Estimating Selection Coefficients: Comparative Analysis of GeneBayes and Agarwal Methods

James Ignacio, Yufeng Shen, Tian Lan

Background

My work this summer was centered around the effects of gene variants on human fitness and understanding models that accurately determine the selection fitness metrics of a gene variant. Specifically, we are looking at protein truncating (PTV) and missense variants that could lead to protein loss of function.

The selection coefficient (s) is a key metric in population genetics that quantifies the strength of natural selection acting on a particular genetic variant. It represents the relative fitness difference between individuals carrying the variant and those that do not, which can be represented by 1-s. Variants with high selection coefficients are typically detrimental, reducing reproductive success and thereby lowering the frequency of these variants in a population. Accurately estimating selection coefficients is crucial for understanding the genetic basis of diseases and identifying genes that contribute to disease risk, particularly for conditions like developmental disorders and cancer.

Our overarching goal was to estimate the selection coefficient of gene variants and evaluate and improve the accuracy of these estimates. My work was inspired by the MisFit model (3), developed by the Shen Lab last year. I spent much of my time reading and understanding MisFit, as well as exploring other methods that could enhance it. Eventually, we came across a Bayesian framework called GeneBayes (2).

Previous models have focused on estimating the selection coefficients of individual variants using metrics such as pLI and LOEUF (2, 3) that measure variation constraint, but this approach is challenged by limited data and noise from genetic drift, and also by the fact that these metrics were difficult to interpret because of their loose relation to fitness consequences of loss of function (4). A new method was developed to generate an interpretable metric of fitness reduction (2). However, it was noted that the method lacks sufficient power for genes with fewer than 10 expected unique loss-of-function (LOF) variants, which applies to approximately 25% of genes. Thousands of genes fall into this category, having few expected unique LOF variants under neutrality, often due to their short protein-coding sequences. To improve the accuracy of selection coefficient estimates, a Bayesian framework called GeneBayes was developed. GeneBayes integrates gene-specific features into its prior distribution, potentially enhancing its ability to accurately estimate selection coefficients.

My research this summer was centered on comparing the performance of GeneBayes with another model developed by Ipsita Agarwal (1) from Columbia University's Department of Biological Sciences. Agarwal's model, which estimates a similar metric called hs, does not incorporate gene features into its priors, making it an ideal baseline for comparison. I hoped to determine whether GeneBayes is worth further exploration as the Shen Lab works to improve its MisFit estimation model.

Goals of the Project

The primary goal of this project was to evaluate the accuracy of the GeneBayes model in estimating selection coefficients for gene variants, particularly in comparison to the Agarwal model. The objectives were:

1. To identify genes with significant differences between the selection coefficients estimated by GeneBayes and Agarwal's model.
2. To compare both model estimates with those produced by the MisFit model to determine accuracy
3. To assess the correlation between GeneBayes estimates and other genetic features, such as missense variant constraint and gene length.

Methods

GeneBayes employs a Bayesian framework that combines a population genetics model with machine learning techniques to estimate an average selection coefficient (S_het) for nearly all PTVs within a gene. The model makes the assumption that all PTVs in the same gene have similar selection coefficients due to their shared impact on protein function. GeneBayes further incorporates gene-specific features into the prior distribution of S_het, using natural gradient boosting to learn these parameters. The model integrates a discrete-time Wright-Fisher population genetics model to account for mutation and genetic drift.

In contrast, Agarwal's model estimates selection coefficients without incorporating gene-specific features into its priors. It uses a Monte Carlo Approximate Bayesian Computation method in conjunction with a Wright-Fisher model, focusing on mutation rates, genetic drift, and other demographic features.This is the perfect model to compare GeneBayes to because it works similarly to GeneBayes, but doesn't take into account gene properties as parameters for learning.

For analysis, I used Python along with various libraries for statistical and computational tasks. GeneBayes simulations were conducted using Python and R, while Agarwal's model was implemented in C++. Data for analysis was sourced from GitHub repositories, as well as the UCSC Genome Browser. The analysis covered 16,229 genes, focusing on the comparison of selection coefficient metrics and the identification of significant discrepancies between the two models. Descriptive statistics, correlation analyses (using Pearson's r), and comparisons with the MisFit model were performed to assess the accuracy and reliability of the estimates.

Results

The analysis revealed significant differences in the selection coefficient estimates between GeneBayes and Agarwal's model for 1,403 out of 16,229 genes. Of these, 128 genes were flagged as pathogenic by GeneBayes but not by Agarwal's model. GeneBayes estimates also showed stronger correlation with missense Z-scores, which quantify the significance and tolerance of missense variants, with a Pearson correlation coefficient (r) of 0.564 compared to 0.517 for Agarwal's model. Both scores are significant, suggesting that missense constraint was a very influential feature, especially in GeneBayes estimation.

Correlation with gene length was weak, with Pearson r of 0.099 for GeneBayes and 0.066 for Agarwal.

The comparison with the MisFit model showed that GeneBayes produced more accurate estimates, with a Pearson r of 0.608 for protein-truncating variants and 0.516 for missense variants, compared to 0.59 and 0.393 for Agarwal's model, respectively.

GeneBayes estimates were particularly aligned with MisFit estimates for both protein-truncating and missense variants.

Discussion

The results indicate that GeneBayes outperforms Agarwal's model in estimating selection coefficients. The integration of gene-specific features into GeneBayes' prior distribution appears to enhance its accuracy, especially for identifying pathogenic variants. However, there are limitations to this approach. For example, in genes like the DNA methyltransferase 3 alpha (DNMT3A) gene, where functionally important missense variants under strong selection and protein-truncating variants not under strong selection have different modes of action, GeneBayes may overestimate the selection coefficient due to its reliance on missense constraint.

This raises a critical question about the generalizability of the GeneBayes model across different types of genes. While it performs well for genes with uniform functional constraints, it may require adjustments for genes with heterogeneous effects. Future work should focus on extending the model to account for such variations and exploring other gene-specific features that could improve the accuracy of selection coefficient estimates.

Limitations

One limitation of the GeneBayes model is its assumption that all PTVs within a gene have similar selection coefficients. This assumption may not hold true for all genes, particularly those with complex functional dynamics, such as DNMT3A. Additionally, the model's reliance on missense constraint as a key feature may lead to overestimation of

selection coefficients in genes where missense and protein-truncating variants have different effects. The current analysis also did not explore other gene features that might contribute to discrepancies in selection coefficient estimates. These features include gene structure, gene expression, biological pathways, connectedness in protein-protein interaction (PPI) networks, co-expression, gene regulatory landscape, conservation, protein embedding features and subcellular localization (5).

Conclusion

GeneBayes effectively combines gene-specific features with population genetics models to estimate selection coefficients for gene variants. Its incorporation of missense variant constraint and other features enhances its accuracy compared to models that do not use such features, such as the Agarwal model. I believe GeneBayes' methods are worth further exploration as the Shen Lab continues to improve the MisFit model.

Future Studies

Future research should focus on expanding the analysis of GeneBayes estimates across a broader range of genes, particularly those with complex functional dynamics like DNMT3A. Additional gene-specific features should be explored to improve the model's accuracy and generalizability. Comparative analysis with other models, including further evaluation against the MisFit model, should be conducted to identify areas for improvement.

Bibliography

1. Agarwal I, Fuller ZL, Myers SR, Przeworski M. Relating pathogenic loss-of-function mutations in humans to their evolutionary fitness costs. eLife. 2023;12:e83172.
2. Cassa CA, Weghorn D, Balick DJ, Jordan DM, Nusinow D, Samocha KE, et al. Estimating the selective effects of heterozygous protein-truncating variants from human exome data. Nature Genetics. 2017;49(5):806-10
3. Fuller ZL, Berg JJ, Mostafavi H, Sella G, Przeworski M. Measuring intolerance to mutation in human genetics. Nature Genetics. 2019;51(5):772-6.
4. Karczewski KJ, Francioli LC, Tiao G, Cummings BB, Alföldi J, Wang Q, et al. The mutational constraint spectrum quantified from variation in 141,456 humans. Nature. 2020;581(7809):434-43
5. Zeng, T., Spence, J.P., Mostafavi, H. et al. Bayesian estimation of gene constraint from an evolutionary model with gene features. Nat Genet 56, 1632–1643 (2024). https://doi.org/10.1038/s41588-024-01820-9
6. Zhao Y, Zhong G, Hagen J, Pan H, Chung WK, Shen Y. A probabilistic graphical model for estimating selection coefficient of missense variants from human population sequence data. medRxiv [Preprint]. 2023 Dec 22:2023.12.11.23299809. doi: 10.1101/2023.12.11.23299809. PMID: 38168397; PMCID: PMC10760286.