Testing the Stationary Distribution Model against the Human Population

Priya Pillai and Run Chen

Abstract

The stationary distribution hypothesis, first proposed by Sewall Wright in 1969, argues that the probability of a random allele occurring at a frequency of p is a beta distribution of p, assuming no selection. As we expect $4N_e u$ to be relatively small in the large human population, we are interested in seeing how this theory applies to real world datasets. We hypothesized we would get a beta distribution with alpha and beta parameters of -1. We ran an analysis of SNPs across the entire exome against this theory using the gnomAD dataset, split by population, type of mutation, and chromosome, and found that one, the alpha and beta parameters were fairly far off from -1, and two, the quality of the fit was extremely poor. We then accounted for selection by looking at the log of the probability distribution of synonymous distribution divided by the probability distribution of nonsynonymous distribution. We did find reasonable high statistical significance in the selection coefficient data, but given the lack of evidence for the model matching the data alone, we hesitate to draw many conclusions from this.

Introduction

The stationary distribution model states that the following equation represents the distribution of allele frequencies: $\phi(p) = c(e^{2\int M(x)/V(x)\,dx})/(V(p))$, where M(p) is the change in p after one generation, V(p) is the variance in p, and c is a normalization constant. When mutation is the only force accounted for, $M(p) = -up + (1-p)v$ where u is the forward mutation rate and v is the backward mutation rate, and $V(p) = p(1-p)/(2N_e)$. This leads to the equation $\phi(p) = cp^{4N_e v-1}(1-p)^{4N_e u-1}$, which is a beta distribution with parameters $\alpha = 4N_e v$ and $\beta = 4N_e u$. When incorporating selection into this theory, $M(p) = sp(1-p)/2 - up + (1-p)v$, which changes the distribution to be $\phi_s(p) = cp^{4N_e v-1}(1-p)^{4N_e u-1}e^{4N_e sp}$. Since this only varies from the nonselective model by the addition of the $e^{4N_e sp}$ term, we can in theory take the logarithm of the distributions divided by each other to determine the selection coefficient, as shown here: $\phi_s(p)/\phi(p) = ce^{4N_e sp}$, $log(\phi_s(p)/\phi(p)) = 4N_e sp + c$.

Using the stationary distribution theory, we could possibly determine another metric for measuring the selection coefficient on larger scales. While the stationary distribution theory would provide a sort of weighted average across a large portion of the genome, most current metrics of selection provide finer grained resolution, so it would likely serve as a check against other metrics for validity. Some examples of metrics to determine selection include the long range haplotype test, F_{st}, p_{excess}, Tajima's D, and dN/dS. The long range haplotype test, F_{st}, p_{excess}, and Tajima's D all serve as metrics of selection within a species, whereas dN/dS compares selection between species. Tajima's D works by comparing two estimates of mutation

rate, which will be skewed depending on the amount of and type of selection. This means that Tajima's D has difficulty distinguishing between the mutation rate effects and the selection coefficient, while the stationary distribution could, in theory, separate these two parameters. F_{st} is a commonly used and useful metric of selection, but can be inconsistent-a small F_{st} value may just be indicative of the allele's commonness in the larger population[1]. p_{excess}, while a powerful statistic, relies on knowing the ancestral populations' allele frequencies, of which there are few reliable methods for us to use. The long range haplotype test, a measure of whether linkage disequilibrium decreases at the rate one would expect, also requires knowing the haplotypes which can require an excessive amount of data storage for large datasets, such as the entire genome, and also risk the loss of privacy of the individuals whose DNA contributed to the sample[2]. The biggest weakness of the stationary distribution model to estimate selection involves the amount of data necessary for it to become statistically relevant, pushing its resolution down. However, as a metric for broad estimates of selection, it may become interesting, especially when combined with current methods.

Results

Data Processing

Data was derived from the gnomAD database. We attempted first to process the data using the myVariant.info API to access the data, but found that the process was too complicated and limited the amount of information we could get from the gnomAD database. We therefore did initial testing on single genes using csvs downloaded directly from gnomAD. We finally downloaded the data for the entire exome using gsutil and were able to analyze and process all of the data by directly parsing the text files to access the data. Data acquired and used from each variant included allele count (the number of times that allele was found) and allele number (the number of times that SNP was tested) per population and the type of the variant. Information on the number of variants found per chromosome can be found in the "meta.txt" file within the GitHub repository; population division of number of variants can be found in the "fit_params.txt" files.

Beta Distribution Fitting and Goodness of Fit

The beta distribution was fit to the data using numPy's built in beta distribution fit algorithm, which determines the alpha and beta parameters for the beta distribution. To calculate the goodness of fit of the data to the model, we used the Kolmogorov Smirnov test. The KS test assumes a null hypothesis that the data matches the model, and the p value states the probability that the data does not actually fit the model. Since we were unfamiliar with this statistic prior to using it, we also compared it to a neutral test-if we randomly sampled the same number of values from the corresponding beta distribution, what would the statistic be? We found that the KS test gave p-values of practically 0 ($p<0.0000005$, usually by a large margin) in every chromosome, population, and category (synonymous and nonsynonymous) we tried. By

[1] This is well demonstrated in "Genetic Signatures of Strong Recent Positive Selection at the Lactase Gene" Bersaglieri et. al. *Am J Hum Genet*. 2004 Jun; 74(6): 1111–1120.
[2] Information on selection statistics from "Genomic insights into positive selection" Biswas and Akey. *TRENDS in Genetics*. 2006 Aug; 22(8): 437-446.

comparison, the neutral value for this test ranged from p-values of around 0.001 to .999 in a somewhat normal distribution.

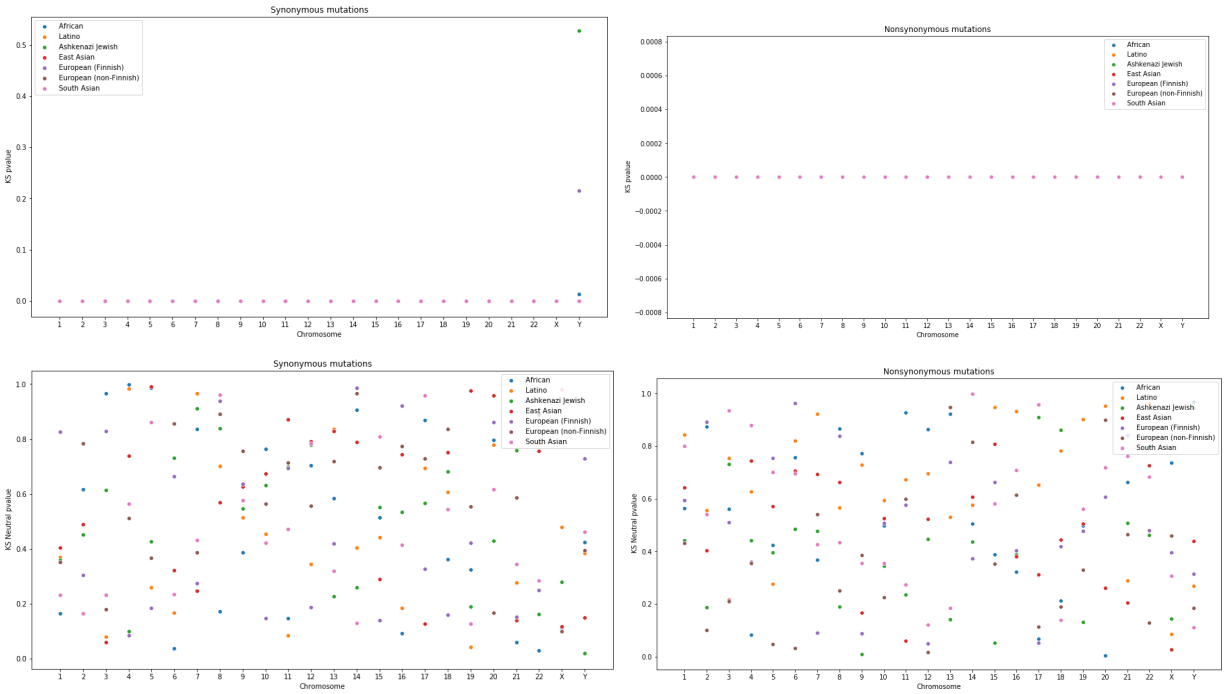

Fig. 1: Top-p-values of the KS test on our variants. Bottom-p-values of the KS test from randomly drawn values from a beta distribution. Left-synonymous. Right-nonsynonymous

These values imply that the stationary distribution under the assumption that there is no selection does not accurately model this data. This could be due to underlying issues with the data, the analysis, or the assumptions made by the model itself. One point to note is that due to time and space constraints, we were only able to test on the exome, not the genome. This highly limited the number of synonymous variants found per chromosome. We expected that the beta distribution would only fit synonymous variants, as we expected there to be little selection on synonymous variants and greater selection on nonsynonymous variants. Even so, we did not see a significantly better fit for the model on synonymous variant data.

$4N_e s$ Coefficient Estimation

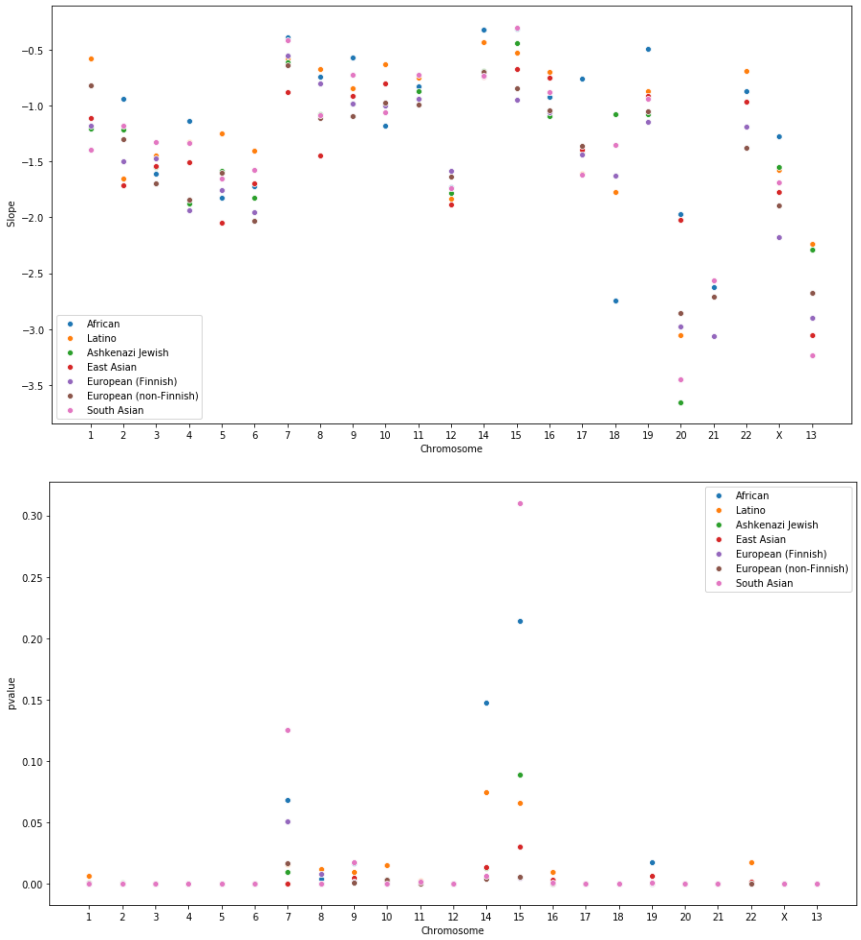

Fig. 2: Top-Slopes of linear regression of $log(\phi_s(p)/\phi(p))$; Bottom-p-values of the linear regression

To estimate $4N_e s$, we first binned the allele frequencies into bins of size 0.01, then ran a linear regression to estimate the slope and the intercept of the model.

Assumptions

There were a lot of assumptions made while analyzing this data. One, we assume gnomAD is a random sampling of alleles given subpopulations. While we do not have clear evidence for this, we would need to make this assumption with practically any dataset. We also assume that the stationary distribution hypothesis can be applied to large classes of mutations, rather than being a theoretical distribution of alleles over a specific site. The stationary distribution is generally discussed as being derived for a random mutation at a given locus rather than a genome-wide distribution. It also assumes, naturally, that the population is at a stable equilibrium-assuming there are only consistent and stable forces of selection, migration,

and drift. Given that we know that the human population has been rising rapidly (changing the force of drift) as well as dramatic changes to human lifestyles and environments that likely have had an effect on both selection and migration, these assumptions seem extremely shaky. [3]

Discussion

The data processing aspect of this work should be noted particularly for the fact that we focused exclusively on exome data rather than genomic data. While this was necessary due to space constraints-the uncompressed version of the exomic data was around 400 GB-this significantly reduced the number of synonymous variants we were able to analyze. The difference between the number of variants tested for nonsynonymous and synonymous conditions could have an effect on the quality of our selection coefficient results in particular.

The metric we used to estimate the slope for the selection coefficient, as well as its p-value, is reliant on the bin size used. Using too large a bin size could erroneously give a good fit to the line; too small with not enough data may erroneously give a poorly fitting line. That being said, there are some very strong trends in the data with highly significant p-values, which indicates that there may be some validity to the selective model for the stationary distribution. More conservatively, it does imply that there is a consistent distribution underlying the spread of allele frequencies that partially relies on the distinction between synonymous and nonsynonymous variants. Whether this distribution can be described by the stationary model still remains unclear-the slope may be representative of some other metric.

This work unfortunately did not lead to particularly conclusive results on whether the stationary distribution model can be used at all for the human population, but the preliminary results we have so far suggest against it. That being said, we could continue to test this theory on the human population by analyzing more specific sets of variants or including classes of functionality to separate out the data.

Supplemental Information

The GitHub repository for our code for this project can be accessed here: https://github.com/priyappillai/rare_muts

A supplemental zip file has been attached containing raw data.

[3] Information on the assumptions for the stationary distribution from “Stationary Allele Frequency Distributions.” Rannala. *eLS*. 2013 Apr