Improving the Robustness of Dominance and Selection Inference

Alyssa Lyn Fortier

May 2, 2018

Contents

1 Introduction 2
  1.1 2D Distribution of Fitness Effects 3
  1.2 Dominance 4
  1.3 General Approach 6

2 Results and Discussion 7
  2.1 2D Distribution of Fitness Effects 7
    2.1.1 Methods Development 7
    2.1.1.1 Improvements to fit δaδi 9
    2.1.1.2 Problem 10
    2.1.1.3 Mixture Model Robustness 11
    2.1.2 Application 15
  2.2 Dominance 17
    2.2.1 Methods Development 17
    2.2.1.1 Problem 17
    2.2.1.2 Sum-of-Squares Optimization 22
    2.2.1.3 Divergence Time 23
    2.2.1.4 Selfing Populations 23
    2.2.2 Application 24

3 Conclusion 27
  3.1 2D DFE 27
  3.2 Dominance 27

4 Materials and Methods 28
  4.1 Data 28
  4.2 Software 28

5 Acknowledgements 28
Abstract

Factors such as genetic drift and natural selection shape genetic variation between populations. To understand this variation, population geneticists study the populations’ demographic history, Distribution of Fitness Effects of new mutations (DFE), and dominance coefficient. While current demographic analyses can incorporate two populations, DFE and dominance studies are largely limited to one. The single-population approach assumes independence between populations, though this is not biologically accurate. We hypothesized that two populations would have high, though not perfect, correlation between fitness effects for the same mutations. We developed a Mixture Model for the 2D DFE that is robust to demographic changes, dominance conditions, and sample size. We applied it to human and Drosophila melanogaster genome data, finding that the correlation parameter between Drosophila populations varies widely among functional gene groups. In contrast, human gene groups have extremely high correlation values, indicating that selection acts very similarly in both populations. We also conducted simulation studies to infer the dominance coefficient between two populations, ultimately identifying some situations where the method is robust to errors. Overall, these tools will allow other researchers to better understand genetic variation between two populations of the same species.

1 Introduction

Not all populations of the same species are genetically identical. Genetic drift, demographic changes such as bottlenecks, natural selection, and new mutations all shape genetic variation among populations. Natural selection and demography are well studied, but it is often difficult to characterize new mutations for a given population. The Distribution of Fitness Effects (DFE) describes how many new mutations are fatal, deleterious, neutral, or beneficial, as shown in Figure 1 [9].

Figure 1: The one-dimensional DFE is thought to be a bimodal distribution.
1.1 2D Distribution of Fitness Effects

Unfortunately, the classic one-dimensional DFE ignores population-specific differences like genetic background and environment [16]. As shown in Figure 2, the same mutation could be neutral in one environment but highly deleterious in another. Thus, new methods are necessary to account for these differences and understand how they operate in natural populations.

![Figure 2](image)

Figure 2: Disparate environmental conditions are one way that selection can operate differently on the same mutations in different populations.

How similar is selection on the same mutations in different populations? Answering this question would provide new insight into speciation. Instead of attributing a speciation event to a single mutation, perhaps entire groups of functionally-related mutations are to blame. Resolving this question would also help researchers understand the pitfalls of Genome-Wide Association Studies (GWAS), which use SNP frequencies to correlate specific gene mutations with complex conditions like heart disease. Since these studies typically recruit patients from European populations, they could neglect important variation. If mutations are evolving differently in different human populations, the results of these widely-used studies may not apply to African or Asian populations, for example.

Previous work has shown that fitness effects of mutations can differ among populations [14], but most studies have focused on a few select mutations instead of a whole distribution [16, 22, 18]. Only one group has put forth a model of a two-dimensional distribution of fitness effects, though their theoretical results have not been empirically verified [5]. By creating a two-dimensional DFE model and applying it to whole genomes, we can answer questions regarding selection patterns in different populations.

The 2D DFE takes on some kind of bivariate distribution, and each population’s marginal DFE should be identical to the 1D DFE obtained from that
population alone. Additionally, there is some correlation coefficient that determines how similarly selection acts upon the same mutations in each population. A low-correlation and high-correlation 2D DFE are shown in Figure 3. The axes of these plots are selection coefficients, $\gamma$. Large negative $\gamma$s, located on the lower left, represent extremely deleterious mutations, while $\gamma$s close to 0 represent nearly-neutral selective effects. A high-correlation distribution means that a mutation’s degree of deleteriousness in Population 1 is likely to be similar to its degree of deleteriousness in Population 2. Alternatively, a low-correlation distribution means that knowing a mutation’s degree of deleteriousness in Population 1 provides no information about its degree of deleteriousness in Population 2. Since two populations of the same species are not too far diverged, we might expect a high-correlation 2D DFE, though environmental differences may mean the correlation coefficient is not equal to exactly 1.

### 1.2 Dominance

Another important selection parameter is dominance. At the single-site level, dominance describes the degree of recessivity of additivity of a certain allele. When paired with the selection coefficient, $s$, for the particular mutation, the dominance coefficient determines the fitness of the heterozygote, as shown in Table 1. For example, a small selection coefficient means the homozygous dominant phenotype imparts a small advantage over the homozygous recessive, while a large one means this phenotype is hugely advantageous. The heterozygote’s fitness is typically between the two homozygotes’ fitnesses, though examples of each extreme are found in nature.

Table 2 summarizes the fitness of the heterozygote under each possibility for the dominance coefficient. When $h < 0$, the heterozygote is less fit than
either of the heterozygotes. For example, in *Pseudacraea eurytus* butterflies, each homozygote mimics the color pattern of a species of poisonous butterfly. The heterozygote has an intermediate phenotype and cannot adequately mimic either, causing it to have lower fitness than either homozygote [20]. When \( h = 0 \), the heterozygote has the same fitness as the homozygous recessive. In domestic cats, the white coat gene dominates over non-white, although white cats have lower fitness due to their decreased camouflage and increased chance of deafness [3]. When \( 0 < h < 1 \), the phenotypes are additive. Thus, heterozygotes will express a phenotype intermediate of either homozygote, such as the pink flowers shown here. When \( h = 1 \), the heterozygote has the same fitness as the homozygous dominant. Cats express the agouti pattern over other coat types whether they are heterozygous or homozygous dominant, for example [6]. Lastly, \( h > 1 \) signifies heterozygote advantage. The classic case is sickle cell anemia. Whereas the homozygous recessive phenotype is sickle-cell anemia and the homozygous dominant phenotype is susceptibility to malaria, the heterozygote has both normal blood cells and immunity to malaria.

<table>
<thead>
<tr>
<th>Genotype</th>
<th>Homozygous Recessive</th>
<th>Heterozygous</th>
<th>Homozygous Dominant</th>
</tr>
</thead>
<tbody>
<tr>
<td>Relative Fitness</td>
<td>1</td>
<td>1 + ( hs )</td>
<td>1 + ( s )</td>
</tr>
</tbody>
</table>

Table 1: The dominance coefficient \( h \) determines the fitness of the heterozygote for a given selection coefficient, \( s \).

<table>
<thead>
<tr>
<th>( h &lt; 0 )</th>
<th>( h = 0 )</th>
<th>( 0 &lt; h &lt; 1 )</th>
<th>( h = 1 )</th>
<th>( h &gt; 1 )</th>
</tr>
</thead>
<tbody>
<tr>
<td>Fitness</td>
<td>Fitness</td>
<td>Fitness</td>
<td>Fitness</td>
<td>Fitness</td>
</tr>
<tr>
<td>aa Aa AA</td>
<td>aa Aa AA</td>
<td>aa Aa AA</td>
<td>aa Aa AA</td>
<td>aa Aa AA</td>
</tr>
</tbody>
</table>

Table 2: The fitness effects of various dominance coefficients, with examples.

Although each allele has a different dominance effect, the average dominance level across the genome still provides vital information. Dominance influences allele frequencies differently in populations that have undergone different demographic changes, so it is particularly important for studying bottlenecks [2]. Most population genetic models fix the dominance coefficient at \( h = 0.5 \), or completely additive. However, this may not be biologically relevant, as empirical studies suggest that recessive alleles may be more common, resulting in \( h < 0.5 \).
[11, 19]. In any case, the effects of dominance are not well characterized, so developing methods to infer dominance from genomic data is vital to account for this potentially confounding factor.

1.3 General Approach

Once a selection of individuals from each population is sequenced, their single-nucleotide polymorphisms (SNPs) are collected. We can summarize this data in a site frequency spectrum (SFS), as shown in Figure 4.

![Figure 4: The Site Frequency Spectrum (SFS) summarizes SNP frequencies in two populations.](image)

The x-axis ranges from 0 to the total number of chromosomes sampled from the French population, and the y-axis is the same for the Zambian population. The colors represent the number of SNPs that occur in a certain number of French samples and a certain number of Zambian samples. Most of the density is located in the left bottom corner. These singletons are SNPs that occur in just one sample in either of the populations. The rest of the density occurs mostly on the diagonal. These SNPs include beneficial and neutral mutations that have swept to high frequency in both populations, as well as high-frequency variants that were present in the population prior to split.

When two populations become separated, genetic drift and natural selection alter their allele frequencies. Genetic drift is a neutral process by which alleles increase or decrease in frequency as individuals randomly mate or die, and natural selection causes beneficial mutations to increase in frequency and deleterious mutations to decrease in frequency.

Population geneticists typically use synonymous mutations, or those that do not change the corresponding amino acid residue, to infer the populations’ demographic history, as these mutations are not directly affected by selection [10, 4]. The nonsynonymous mutations, which do change the amino acid, are targets for selection. Thus, after we account for the neutral genetic variation using a demographic model, we can incorporate variation influenced by selection [21, 15]. These sources of variation should theoretically add together, resulting in the presently observed allele frequencies between the two populations.
The software package \texttt{dadi}, the current standard in population genetics, allows users to develop and optimize demographic models to fit their own synonymous data [12]. An expansion of the software, \texttt{fitdadi}, allows users to optimize a DFE to their nonsynonymous data using a pre-specified demographic model as a kind of null [15]. We used this general approach to explore both the 2D DFE and dominance, although we build upon \texttt{fitdadi} to handle these more complicated models.

2 Results and Discussion

2.1 2D Distribution of Fitness Effects

2.1.1 Methods Development

In order to work with a 2D DFE, it was necessary to alter \texttt{fitdadi} to accommodate bivariate distributions. For the 1D case, the user provided a range of selection coefficients to \texttt{fitdadi}, then used the software to make a cache of frequency spectra at each value. Then, the user specified a general model for the DFE, and \texttt{fitdadi} determined the optimal parameters for the model by changing the weighting of spectra, summing them, and comparing them to the data, as shown in Figure 5.

![Figure 5: 1D fitdadi mechanism](image)

To work in two dimensions, \texttt{fitdadi} must now cache spectra at each possible pair of selection coefficients within a grid, squaring the necessary computational time. However, once the cache is created, the optimization process proceeds exactly as before, as represented in Figure 6.

After \texttt{fitdadi} was expanded, we needed to determine the best model for the 2D DFE. Because little research has been conducted on the explicit form of the DFE, we chose a model for both convenience and supposed accuracy. The lognormal distribution, shown in Figure 7b, fits the marginal DFEs well, gen-
Figure 6: 2D fit$	heta$adi mechanism

eralizes easily to two dimensions, and has an easily interpretable correlation coefficient, making the bivariate lognormal a good candidate. Other bivariate distributions have multiple definitions, such as the gamma, or too many parameters, such as Fisher’s Geometrical Model [5]. The bivariate lognormal has just three parameters: $\mu$ and $\sigma$, which describe the shape, and $\rho$, which describes the correlation between the marginal distributions.

Figure 7: The two-dimensional lognormal distribution is a well-fitting model for the 2D DFE.

We could now visualize the effect of different $\rho$ values on the SFS. As shown in Figure 8, when $\rho = 0$, SNP density is highly concentrated along the left and bottom edges, meaning the populations have less shared variation. Because SNPs can be beneficial in one population and deleterious in the other, they can be swept to high levels in one population while being purged from the other. When $\rho = 1$, much more of the variation is shared between the two populations, so density is more highly concentrated along the diagonal. In this
case, mutations deleterious in one population are likely to be deleterious in the other, causing them to be purged from both. Similarly, mutations beneficial in one population are likely to be beneficial in the other, resulting in the mutations sweeping to high frequency in both populations.

Figure 8: A low correlation 2D DFE results in much less shared variation between populations than a high correlation 2D DFE.

2.1.1.1 Improvements to fit data

Because caching a 2D grid of spectra is so computationally expensive, it is necessary to restrict the domain of selection coefficients. However, under this approach, it is possible to neglect a significant amount of the probability density, as shown in Figure 9a. SNPs that lie in the bottom left corner in Figure 9b are unlikely to be important, as they are likely lethal in both populations. However, SNPs that are lethal in one population but only slightly deleterious or neutral in the other can greatly affect genetic variation between the populations. By integrating from out to infinity on all sides and incorporating the resulting probability density into the edges of the domain, we can account for the part of the distribution we would otherwise neglect.

This improvement to fit data’s integration method greatly improves the bivariate lognormal’s ability to infer the \( \rho \) parameter. Since the proportion parameter corresponds to a proportion of \( \rho = 0, (1 - \rho) \) from the Mixture Model should theoretically match the \( \rho \) value obtained from the bivariate lognormal. As shown in Figure 10a, the old integration method resulted in a mismatch between the two models’ results, and a large proportion of the groups in the bivariate lognormal model rand to \( \rho = 1 \). However, the new integration method resulted in fewer groups with \( \rho = 1 \), as shown in Figure 10b. This not only improves the bivariate lognormal model, but it also means a better correspondence between the two models, allowing us to be more confident in our results.
Figure 9: Integrating outside of the domain incorporates the remaining probability density into the original domain.

Figure 10: The new integration method results in better correspondence between the mixture model and bivariate lognormal.

2.1.1.2 Problem Ideally, inference of the 2D DFE would be robust to small changes in the demographic model. Since we can never model nature exactly, we need to make sure that our approximations do not significantly affect our inference later on. In this case, we want slightly less-likely demographic models to lead to equally slight errors in the inference of the selection parameters. Unfortunately, the bivariate lognormal distribution was sensitive to demography. The best-fit demographic model on the y-axis and a slightly less likely model on the x-axis. If the DFE model worked perfectly, each of the points, representing groups of SNPs, would fall along the line. Because the bivariate lognormal results in erroneous values for many of the groups, it must be improved.

We hypothesized that the bivariate lognormal might be faring poorly because \( \rho \) values are typically very high. The optimization methods we employ may be unable to discern between 1 and values close to 1. Also, when \( \rho \) is very close to 1, the bivariate lognormal distribution exists as an extremely thin line. It is possible that the method has trouble integrating on the close-to-one-dimensional distribution, resulting in values that run to 1. We hypothesized that we could
avoid these problems by not allowing $\rho$ to vary continuously. We developed the Mixture Model, which is composed of a proportion of a bivariate lognormal with $\rho = 0$ and a proportion of a bivariate lognormal with $\rho = 1$, as shown in Figure 12. Though $\rho$ is now fixed, we need to optimize a proportion parameter, $p$, that represents the proportion of the $\rho = 0$ distribution, with $(1 - p)$ being the proportion of the $\rho = 1$ distribution. Although we logically expect correlation between the two populations to be $\approx 1$, we chose to make $p$ the proportion of $\rho = 0$. This is because the optimizer explores the range $[0, 1]$ in log space, so it will be able to explore values near 0 more finely than values near 1, leading to more precise results.

2.1.1.3 Mixture Model Robustness  We verified the mixture model’s effectiveness using simulated data. We generated fake synonymous and nonsynonymous data using realistic parameters, as described in Table 3.

We first explored the mixture model’s ability to accurately infer the correct parameters at various sample sizes. We projected the fake data downward to sample sizes of 2, 4, 6, 20, 50, and 100, then created 10 random samples of each. As shown in Figure 13, the mixture model is able to return the parameters.
Figure 12: The mixture model consists of a proportion of a distribution with no correlation as well as a proportion with full correlation.

Table 3: Fake data parameters for testing mixture model

<table>
<thead>
<tr>
<th>Demographic Parameters</th>
<th>Selection Parameters</th>
</tr>
</thead>
<tbody>
<tr>
<td>s</td>
<td>ν₁</td>
</tr>
<tr>
<td>0.93</td>
<td>2.9</td>
</tr>
</tbody>
</table>

well even for low sample sizes, although it does better as sample size increases. This suggests that the mixture model can work well even for noisy real data using small samples.

Dominance is a potential confounding factor when inferring 2D DFEs, since it influences allele frequencies differently in populations that have and have not undergone a bottleneck. The classic assumption is that the dominance coefficient \( h = 0.5 \), meaning alleles’ selective effects are completely additive, on average. We simulated frequency spectra under dominance coefficients \( h = 0.25 \), \( h = 0.5 \), and \( h = 0.75 \), then optimized the DFE parameters under the assumption \( h = 0.5 \). As shown in Figure 14, the different dominance conditions slightly alter \( \mu \) and \( \sigma \), changing the shape of the distribution, but do not substantially affect \( p \), the correlation between the marginal distributions. Robustness in the proportion parameter is the most important conclusion, since this parameter allows us to answer our question and is the most biologically interpretable. While \( \mu \) and \( \sigma \) are important mathematically, the correlation parameter is most essential to characterizing differences between populations.
(a) True $\mu = 3.6$  
(b) True $\sigma = 5.1$

(c) True $p = 0.1$  
(d) Log likelihood

Figure 13: The mixture model fares well even for small sample sizes.
Figure 14: The mixture model fares well even for alternate dominance assumptions.
2.1.2 Application

We chose to apply the model to two species, *Drosophila melanogaster* and humans, because both contain populations that have undergone bottlenecks. In humans, a portion of the ancient African population underwent a bottleneck to form the European population, while in *Drosophila*, the French population arose from a bottlenecked Zambian population. *Drosophila melanogaster* is an interesting species to study because it is a widely-used model organism, has well-studied genetics, and has a very large effective population size. Humans are also a good species for this analysis because we also have well-studied genetics, have a much lower effective population size, and are intrinsically interesting.

First, we optimized demographic models to the synonymous data for each species. A model of isolation, migration, and exponential growth was suitable for the Zambian and French *Drosophila* populations, while a slightly different model with a period of growth prior to split was optimal for the African and European human populations. These models are visualized in Figure 15, and their parameters are listed in Table 4 and Table 5.

![Image of (a) The Isolation with Migration (IM) demographic model and (b) The Isolation with Migration and instantaneous growth prior to split (IM pre) demographic model.](image)

Figure 15: The IM model is appropriate for *Drosophila* while the IM pre model is appropriate for humans.

Table 4: Demographic model parameters for Zambian and French *Drosophila melanogaster* populations.

<table>
<thead>
<tr>
<th>s</th>
<th>(v_{Zm})</th>
<th>(v_{Fr})</th>
<th>T</th>
<th>(m_{Zm-Fr})</th>
<th>(m_{Fr-Zm})</th>
<th>misID</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.959</td>
<td>3.45</td>
<td>0.207</td>
<td>0.833</td>
<td>0.802</td>
<td>3.58</td>
<td>0.0498</td>
</tr>
</tbody>
</table>

Table 5: Demographic model parameters for African and European human populations.

<table>
<thead>
<tr>
<th>(v_{pre})</th>
<th>(T_{pre})</th>
<th>s</th>
<th>(v_{Af})</th>
<th>(v_{Eu})</th>
<th>T</th>
<th>(m_{1,2})</th>
<th>(m_{2,1})</th>
<th>misID</th>
</tr>
</thead>
<tbody>
<tr>
<td>1.85</td>
<td>0.52</td>
<td>0.96</td>
<td>3.34</td>
<td>4.28</td>
<td>0.099</td>
<td>0.29</td>
<td>0.26</td>
<td>0.015</td>
</tr>
</tbody>
</table>
Next, we fit the mixture model to the entire human and \textit{Drosophila} nonsynonymous spectra. The model comparisons are shown in Figure 16. The upper left corner is the nonsynonymous data SFS and the upper right is the best-fit model SFS. The lower left is a plot of the residuals, determined by subtracting the model SFS from the data SFS. We expect a well-fitting model to result in randomly distributed residuals, indicative of noise, while a poorly-fitting model would result in systematic overestimates or underestimates across the spectrum. The lower right is a histogram of the absolute values of the residuals, which we would expect to be normally distributed about 0 for a well-fitting model. The optimal parameters are listed in Figure 6. Overall, humans have a much lower proportion value, or proportion of the distribution with $\rho = 0$. Thus, across the whole genome, selection is acting more similarly across the two human populations than across the two \textit{Drosophila} populations.

![Figure 16: The two-dimensional lognormal distribution is a well-fitting model for the 2D DFE.](image)

**Table 6:** Mixture model parameters for whole human and \textit{Drosophila} datasets.

<table>
<thead>
<tr>
<th></th>
<th>$\mu$</th>
<th>$\sigma$</th>
<th>$\rho$</th>
<th>misID</th>
</tr>
</thead>
<tbody>
<tr>
<td>Human</td>
<td>2.5</td>
<td>4.5</td>
<td>0.0071</td>
<td>0.0087</td>
</tr>
<tr>
<td>\textit{Drosophila}</td>
<td>5.5</td>
<td>3.7</td>
<td>0.038</td>
<td>0.027</td>
</tr>
</tbody>
</table>

However, applying the mixture model to the whole genome does not provide too much information. We hypothesize that different functional groups of genes may have different correlation coefficients. For example, SNPs in genes with critical functions, like reproduction, may have higher correlations between populations than genes that are perhaps less critical, such as stress response.

We used the \textit{Gene Ontology} (GO), which is a manually-curated hierarchy of gene annotations useful for sorting genes into functional categories [1, 8]. We chose to use the GO Generic Slim, which is a subset of the GO that contains
core functions that are not species-specific. Since the GO Slim is better curated and contains terms seen across a wide range of species, we figured that it would be the most biologically interesting way to divide the SNPs.

In order to sort SNPs into their functional groups, we used ANNOVAR [23] to annotate SNPs to their corresponding genes. We then used Ensembl Biomart [24] to annotate genes to the most relevant Gene Ontology Slim annotation. The direct children of GO:0008510, biological process, became the targets of our analysis. Any other GO Slim descendants of these children were combined with their parent term, and could be included in multiple parent groups. Thus, each resulting SNP group represented the SNPs directly associated with that annotation as well as the SNPs in all descendant terms.

Using the synonymous SNPs for each GO Slim group, we determined the $\theta$ value, or data scaling factor, for each. Larger $\theta$s mean that the Slim group contains more data, while smaller $\theta$s mean that the Slim group contains less data, and uncertainties will therefore be higher. Figure 17 shows the Mixture Model results against the bivariate lognormal results for each Slim group in *Drosophila*, while Figure 18 shows the comparison in humans.

*Drosophila* have a much wider range of proportion values than humans, indicating that selection is acting more differently on the same mutations in *Drosophila* populations than in human populations. The human proportion values are all very close to 1, indicating that human populations are not all that different from each other.

Table 7 shows specific values for GO Slim groups in both humans and *Drosophila* for all human groups with $\theta > 100$ and all *Drosophila* groups with $\theta > 1000$. The table is sorted by the *Drosophila* proportion values. There is no correlation between the *Drosophila* and human values for each GO Slim group, although it could simply be difficult to discern because the human proportion values are universally high.

When using the bivariate lognormal model of the DFE, the parameters were sensitive to changes in the demographic model, demonstrated in Figure 11. In contrast, the Mixture Model is more robust, as shown in Figure 19.

2.2 Dominance

2.2.1 Methods Development

For this analysis, we assume a one-dimensional DFE, meaning that selection is acting exactly the same on the same mutations in both populations. Though we know from our study of the 2D DFE that this is not entirely accurate, $\rho$ values are high in reality, meaning the assumption is good enough to allow us to develop methodology.

First, we create fake synonymous and nonsynonymous data using realistic parameters, as described in Table 8.

2.2.1.1 Problem To test how well we can infer $h$, we used $\partial a \partial h$ to optimize 22 incorrect demographic models against our simulated synonymous data. We
Figure 17: The mixture model results for different GO slim groups in *Drosophila*.
Figure 18: The mixture model results for different GO slim groups in humans.
<table>
<thead>
<tr>
<th>GO Slim Group</th>
<th>Definition</th>
<th>Drosophila $(1 - p)$</th>
<th>Human $(1 - p)$</th>
</tr>
</thead>
<tbody>
<tr>
<td>GO:0030234</td>
<td>enzyme regulator activity</td>
<td>0.537</td>
<td>0.998</td>
</tr>
<tr>
<td>GO:0008233</td>
<td>peptidase activity</td>
<td>0.548</td>
<td>—</td>
</tr>
<tr>
<td>GO:0007059</td>
<td>chromosome segregation</td>
<td>0.612</td>
<td>—</td>
</tr>
<tr>
<td>GO:0003700</td>
<td>DNA binding transcription factor activity</td>
<td>0.672</td>
<td>0.997</td>
</tr>
<tr>
<td>GO:0002376</td>
<td>immune system process</td>
<td>0.691</td>
<td>0.940</td>
</tr>
<tr>
<td>GO:0051276</td>
<td>chromosome organization</td>
<td>0.695</td>
<td>0.998</td>
</tr>
<tr>
<td>GO:0004641</td>
<td>cellular nitrogen compound metabolic process</td>
<td>0.750</td>
<td>0.993</td>
</tr>
<tr>
<td>GO:0009056</td>
<td>catabolic process</td>
<td>0.751</td>
<td>1.000</td>
</tr>
<tr>
<td>GO:0008283</td>
<td>cell proliferation</td>
<td>0.784</td>
<td>0.987</td>
</tr>
<tr>
<td>GO:0006950</td>
<td>response to stress</td>
<td>0.784</td>
<td>0.992</td>
</tr>
<tr>
<td>GO:0007049</td>
<td>cell cycle</td>
<td>0.801</td>
<td>0.999</td>
</tr>
<tr>
<td>GO:0007155</td>
<td>cell adhesion</td>
<td>0.803</td>
<td>0.992</td>
</tr>
<tr>
<td>GO:0048856</td>
<td>anatomical structure development</td>
<td>0.810</td>
<td>0.998</td>
</tr>
<tr>
<td>GO:0030154</td>
<td>cell differentiation</td>
<td>0.821</td>
<td>0.999</td>
</tr>
<tr>
<td>GO:0061024</td>
<td>membrane organization</td>
<td>0.836</td>
<td>0.970</td>
</tr>
<tr>
<td>GO:0040011</td>
<td>locomotion</td>
<td>0.839</td>
<td>0.965</td>
</tr>
<tr>
<td>GO:0007165</td>
<td>signal transduction</td>
<td>0.840</td>
<td>1.000</td>
</tr>
<tr>
<td>GO:0006464</td>
<td>cellular protein modification process</td>
<td>0.843</td>
<td>0.998</td>
</tr>
<tr>
<td>GO:0022607</td>
<td>cellular component assembly</td>
<td>0.849</td>
<td>0.979</td>
</tr>
<tr>
<td>GO:0000003</td>
<td>reproduction</td>
<td>0.861</td>
<td>0.999</td>
</tr>
<tr>
<td>GO:0009058</td>
<td>biosynthetic process</td>
<td>0.861</td>
<td>0.979</td>
</tr>
<tr>
<td>GO:0050877</td>
<td>nervous system process</td>
<td>0.864</td>
<td>0.972</td>
</tr>
<tr>
<td>GO:0007010</td>
<td>cytoskeleton organization</td>
<td>0.872</td>
<td>0.988</td>
</tr>
<tr>
<td>GO:0051301</td>
<td>cell division</td>
<td>0.880</td>
<td>—</td>
</tr>
<tr>
<td>GO:0044281</td>
<td>small molecule metabolic process</td>
<td>0.881</td>
<td>0.982</td>
</tr>
<tr>
<td>GO:0008219</td>
<td>cell death</td>
<td>0.889</td>
<td>0.988</td>
</tr>
<tr>
<td>GO:0006629</td>
<td>lipid metabolic process</td>
<td>0.892</td>
<td>1.000</td>
</tr>
<tr>
<td>GO:0016491</td>
<td>oxidoreductase activity</td>
<td>0.894</td>
<td>—</td>
</tr>
<tr>
<td>GO:0006810</td>
<td>transport</td>
<td>0.907</td>
<td>0.998</td>
</tr>
<tr>
<td>GO:0016301</td>
<td>kinase activity</td>
<td>0.908</td>
<td>0.972</td>
</tr>
<tr>
<td>GO:0007267</td>
<td>cell-cell signaling</td>
<td>0.921</td>
<td>1.000</td>
</tr>
<tr>
<td>GO:0042592</td>
<td>homeostatic process</td>
<td>0.962</td>
<td>0.995</td>
</tr>
<tr>
<td>GO:0040007</td>
<td>growth</td>
<td>0.964</td>
<td>0.994</td>
</tr>
<tr>
<td>GO:0005975</td>
<td>carbohydrate metabolic process</td>
<td>0.992</td>
<td>0.999</td>
</tr>
<tr>
<td>GO:0021700</td>
<td>developmental maturation</td>
<td>1.000</td>
<td>—</td>
</tr>
</tbody>
</table>

Table 7: Proportion values for different GO slim groups in both Drosophila and humans.

Table 8: Fake data parameters for testing dominance inference

<table>
<thead>
<tr>
<th>Demographic Parameters</th>
<th>Selection Parameters</th>
</tr>
</thead>
<tbody>
<tr>
<td>$s$</td>
<td>$\nu_1$</td>
</tr>
<tr>
<td>0.95</td>
<td>3.4</td>
</tr>
</tbody>
</table>
Figure 19: The mixture model is robust to demographic assumptions in humans.
then cached spectra based on these incorrect models, then optimized the DFE parameters and h. Unfortunately, the inference of h was extremely sensitive to our choice of underlying demographic model, typically resulting in huge over-estimates of h, as shown in Figure 20. Each of the colored lines represents a different wrong demographic model. For example, IM_sym_mig is an IM demographic model with symmetric migration, which is slightly less realistic than the best-fitting IM model that allowed migration rates to differ in each direction. Only the perfect IM model was able to return the correct h value, while even the slightly-incorrect other models failed to do so.

Figure 20: h is overestimated when the demographic model is wrong.

2.2.1.2 Sum-of-Squares Optimization In an attempt to solve this problem, we implemented a different optimization. Before, we optimized models based on log-likelihood comparisons of model to data. Because we noticed that the best-fit demographic models would often have large residuals when compared to the synonymous data, we hypothesized that we could infer h more robustly by optimizing the synonymous residuals against the nonsynonymous. We implemented our sum-of-squares optimization, which involved minimizing the sum of the squares of entries of the metaresidual, which the difference between the synonymous and nonsynonymous residuals. An example of this comparison is shown in Figure 21. In this plot, the upper left corner is the residual from the optimal demographic model, and the upper right corner is the residual from the optimal selection model. The metaresidual, which compares the two regular residuals, is shown in the bottom left corner, while a histogram of its values is located on the bottom right.

The sum-of-squares optimization fared better than the log-likelihood approach for some of the incorrect demographic models, such as the ones that were only slightly wrong. However, this approach did equally as terribly as
log-likelihood for the worse models, though it tended to underestimate instead of overestimate $h$, as shown in Figure 22.

Next, we hypothesized that we could improve the inference of $h$ by projecting the data down to smaller sample sizes, thereby reducing noise. We randomly sampled our fake data ten times, then projected each spectrum down to sample sizes 17, 37, and 47. As shown in Figure 23, we could infer $h$ better in smaller projections for the best-fit model. $h$ inference was still sensitive to changes in the demographic model, and even the slightly-wrong demographic models fared poorly.

### 2.2.1.3 Divergence Time
Since we did not discover a way to systematically improve the inference of $h$, we hypothesized that we can use the method only in limited cases. We suspected that inferring $h$ becomes more difficult as the time post-bottleneck increases. Thus, we tested our model under fake data with different divergence times, with the rest of the parameters remaining the same. As shown in Figure 24, $h$ inference is most robust for small divergence times.

### 2.2.1.4 Selfing Populations
One recent manuscript [13] suggested that dominance inference was more robust when one population was selfing, or highly inbreeding. To test this, we created fake data where $h$ for one population was fixed at 1 and the other was free to vary. As shown in Figure 25, dominance was still underestimated, although inference was much more robust, even for some of the worst demographic models.
2.2.2 Application

Applying our dominance model to humans and *Drosophila*, we obtained unrealistically high $h$ values near 1. Because our simulation studies reveal that even slightly incorrect demographic models tend to cause $h$ to go to 1, perhaps our demographic model is imperfect, the data is too noisy, or the model is simply not applicable to these particular populations. Further simulation studies are necessary before making any conclusions from real data.
Figure 23: The sum-of-squares optimization fares better when spectra are projected down.
Figure 24: Inference of $h$ works best when divergence time is low.

Figure 25: Inference of $h$ works best when one of the populations is selfing.
3 Conclusion

3.1 2D DFE

Through simulation studies, we show that the Mixture model is robust to changes in underlying demography, sample size, and dominance. Applying the model to real data in Drosophila and humans reveals that humans have much higher correlation coefficients than Drosophila. This could be due to the fact that humans have a longer generation time, so natural selection has been less effective in altering allele frequencies between the two populations. Additionally, since humans are better able to alter their own environments, environmental differences may impact selection less than in Drosophila.

We ascertained a wide range of proportion values, especially in flies. One interesting result was immune system process, which had the lowest believable proportion value in humans, at 0.948, and an even lower value in flies, 0.691. It makes sense that this value is low for both species, as the strain on a population’s immune system is conceivably linked to the environment - particularly the types of pathogens and other organisms present in each environment. Groups like carbohydrate metabolic process, growth, and homeostatic process had sufficient amounts of data as well as correlation coefficients above 0.96 in both humans and Drosophila. This suggests that these processes are essential to both species, which also makes sense logically. Once we obtain uncertainty values for these groups, we will be able to say more conclusively which groups are significantly different from others.

Moving forward, I am to quantify uncertainties in order to make statistically significant conclusions about the different gene groups. I also aim to infer a proportion of positive selection along with the Mixture Model in Drosophila, as we know that positive selection is realistic for this species.

3.2 Dominance

The traditional method of inferring dominance was sensitive to the underlying demographic model. Unfortunately, altering the optimization method and reducing sample sizes did not make dominance inference robust to demography. However, we verified through simulations that it is possible to infer dominance accurately for populations with low divergence time or situations in which one population is selfing. Thus, the method may be able to help determine dominance for recently-diverged species, such as dogs and wolves, or for selfing plants, such as Arabidopsis thaliana.

For the moment, the 2D DFE model appears more promising to continue than dominance. However, the new integration in fitda@ was implemented after doing these analyses. Perhaps revisiting this model using the new integration will improve the results.
4 Materials and Methods

4.1 Data
We obtained human data from the 1000 genomes project [7] and Drosophila data from the Drosophila Genome Nexus [17].

4.2 Software
We used the software packages ∂a∂h for demogrpahic inference [12] and fit∂a∂h [15] for selection inference. We used ANNOVAR [23] for annotating SNPs, Biomart Ensembl [24] and Gene Ontology [1, 8] for sorting genes into functional groups.

5 Acknowledgements
This research was sponsored by the Senior Vice President for Research, Private Donors, and NSF grant #DEB-1146074. Thank you to Dr. Gutenkunst for his mentorship, Travis Struck for his guidance, and UBRP for supporting aspiring researchers.

References