Detecting Selection on Segregating Gene Duplicates in a Population

Tristan L. Stark (1,2, *), Rebecca S. Kaufman (1), Maria A. Maltepes (1), and David A. Liberles (1,*)

[1]Department of Biology and Center for Computational Genetics and Genomics, Temple University, Philadelphia, PA 19122, USA

[2]Current Address: Discipline of Mathematics, University of Tasmania, Hobart, Tasmania 7001, Australia

email contact: TLS: tristan.stark@utas.edu.au; DAL: daliberles@temple.edu

Abstract

Gene duplication is a fundamental process that has the potential to drive phenotypic differences between populations and species. While evolutionarily neutral changes have the potential to affect phenotypes, detecting selection acting on gene duplicates can uncover cases of adaptive diversification. Existing methods to detect selection on duplicates work mostly inter-specifically and are based upon selection on coding sequence changes, here we present a method to detect selection directly on a copy number variant segregating in a population. The method relies upon expected relationships between allele (new duplication) age and frequency in the population dependent upon the effective population size. Using both a haploid and a diploid population with a Moran Model under several population sizes, the neutral baseline for copy number variants is established. The ability of the method to reject neutrality for duplicates with known age (measured in pairwise dS value) and frequency in the population is established through mathematical analysis and through simulations. Power is particularly good in the diploid case and with larger effective population sizes, as expected. With extension of this method to larger population sizes, this is a tool to analyze selection on copy number variants in any natural or experimentally evolving population.

1 Introduction

A major goal in computational genomics is to uncover the intra- and inter-specific changes that affect organismal phenotypes, including those driven by selective forces. An extensive suite of methods exists

to characterize the fixation and divergence of point mutations (Anisimova & Liberles, 2012), but the methods developed to date aimed at studying gene duplicates have been mostly interspecific and based upon on patterns of sequence divergence (Conant & Wagner, 2003) or retention patterns over a phylogeny (Arvestad et al., 2009; Tofigh et al., 2010; Yohe et al., 2019). There is a need for new methods that characterize selection on segregating copy number variants.

Gene duplication affecting a single gene occurs through two predominant processes. Tandem duplication leads to duplicate copies that may encompass the entire length of the gene and are initially found adjacent to one another on a single chromosome (Katju & Lynch, 2006). Transposition, including retrotransposition-mediated processes, is the other common mode of duplication, which leads to unlinked duplicate copies (Innan & Kondrashov, 2010).

Evidence of gene duplication has been found in all three domains of life (Lynch & Force, 2000; J. Zhang, 2003). Within gene families, divergent function has been identified, suggesting gene duplication is an important contributor to genome diversification (Innan & Kondrashov, 2010). A dramatic case involves the convergent expansion of gene families through duplication in the devil worm and in oyster genomes in response to temperature stress (Guerin et al., 2019). Additionally, copy number variation (CNV) is known to be associated with disease (C. Zhang et al., 2013) and the ability to adapt to new or changing environments (Bornholdt et al., 2013; Perry et al., 2007). Despite the apparent importance of duplication in genomic evolution, the mechanisms by which gene duplicates are fixed and maintained in a population are not well understood, including the mutational state at the point of fixation.

Gene duplication occurs in an individual and can be fixed or lost in the population. At the time of duplication, the frequency of the new locus in a haploid population will be 1/N and 1/2N in a diploid population. If the duplicated gene is selectively neutral, its probability of fixation is its frequency in the population. As redundant duplicates segregate, mutations can accumulate. In the absence of non-neutral forces, the probability of a single mutation going to fixation decreases as population size grows. Multiple functions and structures can affect the likelihood of individual genes losing function (pseudogenization), becoming subfunctionalized through partition of ancestral functions, or gaining a new function (neofunctionalization). (Force et al., 1999; Hughes, 1994; Lynch et al., 2001). Mutations that disrupt the function of non-redundant proteins will be selected against at a population genetic level (Ohno, 1970). Gene duplication provides new material for drift or selection to act on, and therefore has been proposed as a major driving force for functional diversification (Hughes, 1994; Ohno, 1970). Identical duplicates with redundant function allows natural selective pressures to be relaxed on both copies while still redundant. Reduced selective pressures on redundant copies of a locus may allow otherwise prohibited mutations to accumulate, potentially leading to novel functions (Hughes, 1994).

Duplicates can be stably maintained in a population when they differ in some aspects of their func-

tion (J. Zhang, 2003). There is a high probability that random genetic drift will cause loss of any given duplicated gene. Most mutations in gene duplicates will be deleterious to function, although potentially selectively neutral when there are redundant copies. Loss of one copy is likely to be through fixation of a null mutation at the duplicated locus, leading to pseudogenization (Lynch & Force, 2000). Persistence of gene duplicates in the population may be driven by fixation of rare, beneficial mutations (neofunctionalization) (Ohno, 1970), subfunctionalization (Force et al., 1999) or dosage balance in cases where genes are duplicated together with interacting partners (Konrad et al., 2011).

The origin of novel function is an important outcome of gene duplication. Positive selection speeds up fixation of nearly neutral substitutions that create a new, but weakly active function (J. Zhang, 2003). In larger populations, positive selection is likely to supercede nearly neutral mutations as a driver for this process. In addition to selection on mutations in the duplicate copy, there can also be selection on the duplicate copy itself. Positive and negative fitness effects have been reported for gene duplicates in many different genes. Copy-number increase in the human salivary amylase gene (AMY1) has enabled adaptation to a high-starch diet (Perry et al., 2007) and segmental duplications of the chemokine gene CCL3L1 gene are associated with decreased susceptibility to HIV infection (Gonzalez et al., 2005). Methods to characterize selection on point mutations within duplicate genes are well established (for example, dN/dS, MacDonald-Kreitman tests, or population genetic outliers), whereas those that detect selection directly on copy number variants do not exist. Probabilistic models for inferring selection on gene duplicates have previously been described in an inter-specific context, but not intra-specifically (Konrad et al., 2011; Stark et al., 2017).

Allele age can be defined as the duration of time a mutant allele has been segregating in a population (De Sanctis et al., 2017). Directional selection, both positive and negative, can lead to a functional allele that is younger than expected given its frequency (Platt et al., 2019). If not lost from the population, an allele under directional selection will reach a given frequency faster than a neutral allele (Maruyama, 1974).

Methods to detect selection on individual SNPs that are segregating in a population (Platt et al., 2019) rely on the complication of examination of tracts of identical descent that have not been interrupted by recombination to establish age. No such method exists for CNVs, but in principle, such methods can be much simpler because the coding sequence of the gene can accumulate synonymous mutations with time anywhere in its sequence. From this, a pairwise dS value is a natural measure of CNV age, with an assumption of the neutrality of synonymous mutations. This assumption is in some cases violated, but is reasonable and dS is commonly used as a molecular clock (Anisimova & Liberles, 2012).

With this in mind, a continuous time Moran model (Moran, 1958) is proposed, to infer from duplicate age (measured in pairwise dS for the duplicate pair), the expected duplicate frequency in a population

depending upon relevant population genetic and selective parameters. The Moran model is a stochastic model of mutational and selective processes that assumes a fixed population size (N) over generations and can be implemented in either a haploid or a diploid setting. At each instant when the state of the model may change, one gamete is chosen at random to die and is replaced by a new gamete with probabilities assigned to each genotype based on fitness and frequency in the population. Transitions in this Markov chain occur at the death of a single individual. Using this model, we test the null hypothesis of neutral evolution with the aim of building a method that enables detection of non-neutral processes acting on segregating duplicated genes based upon their age (measured in pairwise dS units) and their frequency in the population. This model has been implemented exactly in both haploid and diploid populations and is first described here. Future approximations will need to be introduced to enable extension to realistic population sizes to fit fungal, metazoan, plant, or other datasets.

2 Methods

We consider a locus which has only one allele at the time of duplication, so that the population starts with 1 one individual having two unlinked copies of the gene, and $N-1$ individuals having only a single copy. We present two population genetic models to model the subsequent evolution of such a population, one for the haploid case, and one for the diploid case. Both models are continuous-time Markov chains similar to the classic Moran model (Moran, 1958). We introduce a simple statistical test to detect selection based on this model with the underlying distribution derived from the population genetic model. The test compares the observed proportion of the population carrying the duplicate copy at a particular time to the distribution of proportions that would be attained by the model under neutrality, and can be extended to a time-series test to improve statistical power where such data would be available (for example from experimental evolution studies (Lauer et al., 2018)). The test itself is identical for the haploid and diploid cases, except that the underlying distribution is derived from the corresponding model.

The two models share a common set of assumptions:

- We assume that pseudogenization occurs at Poisson rate μ_p for each duplicate copy, and we model an individual with a pseudogenized copy as equivalent to an individual without the extra copy. E.g. if a haploid individual has two copies of the gene, one of which becomes pseudogenized, we then consider the individual to be equivalent to a 'wild-type' individual with a single copy.
- We assume that neofunctionalization occurs at a Poisson rate μ_n for each duplicate copy, and always confers a fixed fitness benefit. We further assume that only one of the two loci (representing the original and duplicated copies) can become neofunctionalized, and we do not keep track of which is which.

- We assume that individuals in the population are replaced uniformly at random at Poisson rate 1, and that they are replaced by a new individual according to their relative fitness and haploid/diploid breeding regime — the individual to be replaced can participate in breeding their own replacement.

Haploid model

We model the haploid population as a continuous-time Markov chain with state space

$$\mathcal{S} = \{(i, j) : i, j \in 0, \ldots, N \text{ and } i + j \leq N\}, \tag{1}$$

where i tracks the number of individuals carrying an unmodified duplicate copy of the gene in question, and j tracks the number of individuals carrying a duplicate copy which has become neofunctionalized. The number of single-copy 'wild-type' individuals is given by $N - i - j$.

We let f_d and f_n denote the fitness of an individual with an unmodified and a neofunctionalized duplicate copy respectively, relative to the single-copy wild type. We allow f_d to take any value greater than 0, and we can interpret different values of f_d as corresponding to different biological processes.

- $f_d < 1$ represents a situation in which there is a cost to maintaining the duplicate that might be expected based upon the biosynthetic cost of maintaining and synthesizing products from an extra gene (Wagner, 2005).
- $f_d > 1$ represents the situation where dosage effects confer a selective benefit for increasing gene dosage, for example as might occur when the product is limiting in its pathway. The extra copy can lead to increased expression, and in turn improve some physiological function.
- $f = 1$ is the neutral case, where the effects of biosynthetic cost and dosage are negligible.

We will only consider cases here in which $f_n \geq 1$, representing neofunctionalization, although allowing $f_n < 1$ does not break any of our modelling assumptions. We will however consider $1 \leq f_n < f_d$, representing a situation where the neofunctionalized copy is less beneficial than the dosage effect of maintaining an extra copy of the original gene.

The process is characterized by its generator matrix

$$\boldsymbol{Q} = [q_{(i,j)(k,l)}], \tag{2}$$

where the non-zero off-diagonals of $\boldsymbol{Q}$ are given by,

$$q_{(i,j)(k,l)} = \begin{cases} \frac{(N-i-j)}{N} \frac{if_d}{N-i-j)+if_d+jf_n} & \text{for } k = i+1, l = j \\ \frac{j}{N} \frac{if_d}{N-i-j)+if_d+jf_n} & \text{for } k = i+1, l = j-1 \\ \frac{(N-i-j)}{N} \frac{jf_n}{N-i-j)+if_d+jf_n} & \text{for } k = i, l = j+1 \\ \frac{i}{N} \frac{(N-i-j)}{jf_n)+if_d+jf_n} + i\mu_n & \text{for } k = i-1, l = j+1 \\ \frac{j}{N} \frac{(N-i-j)}{N-i-j)+if_d+jf_n} + j\mu_p & \text{for } k = i, l = j-1 \\ \frac{i}{N} \frac{(N-i-j)}{N-i-j)+if_d+jf_n} + i\mu_p & \text{for } k = i-1, l = j \end{cases} \tag{3}$$

In each case the first factor of the first term represents the death of an individual chosen uniformly, while the second factor represents the birth of their replacement, chosen dependent on the relative fitness of the replacement allele to the overall fitness of the population. The second term (where such exists) represents the effect of mutation.

In the neutral case, where $f_d = f_n = 1$ the model can be simplified to a simple birth-and-death process, as in this case 'neofunctionalization' is no longer meaningful, and we need not track the number of neofunctionalized copies. Doing so allows for much more efficient computation for the neutral case, which is of primary concern in our test for selection.

Diploid model

The diploid model is similar to the haploid model, but there are six possible genotypes to consider: $AA--$, $AAA-$, $AAAA$, AAA'-, $AAAA'$, $AAA'A'$, where A represents an unmodified copy, A' represents a neofunctionalized copy, and $-$ represents the absence of a copy. The idea here is that we think of two sets of loci, the original and the duplicated locus, with each of the two chromosomes initially having a single copy of the gene ($AA--$). The copies on both chromosomes are assumed to be duplicated by the initial duplication event ($AAAA$) consistent with an origin through retrotransposition, and the remaining combinations come about through subsequent mating, and neofunctionalization. The state space of this model is the set of 5-vectors with natural valued entries summing to $\leq N$,

$$\mathcal{S} = \{\boldsymbol{s} \in \mathbb{N}^5 : s_1 + s_2 + s_3 + s_4 + s_5 \leq N\}. \tag{4}$$

In the interests of brevity we present the transition rates in a compact form as

$$q_{\boldsymbol{s},\boldsymbol{s}+\boldsymbol{e}_x-\boldsymbol{e}_y} = p_b(x|\boldsymbol{s})p_d(y|\boldsymbol{s}) + \mu_n C_n(\boldsymbol{s},x,y) + \mu_p C_p(\boldsymbol{s},x,y), \tag{5}$$

where $\boldsymbol{s}$ is the state of the process $\boldsymbol{e}_x$ is a unit vector with 1 in the x entry, $p_b(x|\boldsymbol{s})$ is the probability that an individual of type x is born when the process is in state $\boldsymbol{s}$, $p_d(y|\boldsymbol{s})$ is the probability that an individual of type y dies when the process is in state $\boldsymbol{s}$, and $C_n(\cdot)$, $C_p(\cdot)$ are functions counting the number of ways in which $\boldsymbol{s} + \boldsymbol{e}_x - \boldsymbol{e}_y$ can be reached from $\boldsymbol{s}$ by neofunctionalization and pseudogenization respectively. The specifics of each function are easy to determine, but unwieldy when written out, so we omit them here.

In this case, we parameterize selection in terms of the selection coefficients s_d and s_n, and we assume that

$$f_X = 1 + n_d s_d + n_n s_n, \tag{6}$$

where X denotes a genotype, n_d is the total number of duplicate copies, including neofunctionalized, while n_n is the number of neofunctionalized copies. For example, $f_{AAAA'} = 1+2s_d+1s_n$. Effectively, we assume that any selective benefit conferred from a duplicate copy is also conferred to a neofunctionalized copy, ontop of any benefit conferred from the neofunctionalization, and that the selective effects are additive. This parameterization allows us to consider all of the same underlying biology as discussed in the haploid case, but we keep the number of model parameters low by now explicitly assuming additive effects. Figure 1 shows a conceptual diagram of this rationale, where we think of some physiological function increasing with gene dosage towards an asymptote. The relationship between function and dosage can be thought of as having two modes, a 'linear' mode, where function is well approximated as linearly increasing in dosage, and a 'saturation' mode, in which increasing dosage results in little to no increase in function. This then leads to additive selection as parameterized here, where $s_d > 0$ corresponds to the linear mode, and $s_d = 0$ corresponds to the saturation mode. To model the situation in which a neofunctionalized copy is less beneficial than an extra copy of the original gene, we allow for $-s_d < s_n < 0$, representing the difference in benefit conferred from neofunctionalization and dosage.

Similarly to the haploid model, the diploid model can be reduced significantly in the neutral case. Discarding neofunctionalization there are only three genotypes to consider, $AA--$, $AAA-$ and $AAAA$. This leads to a model with state space

$$\mathcal{S} = \{(i,j) : i,j \in 0,\ldots,N \text{ and } i+j \leq N\}, \tag{7}$$

where i tracks the number of double-duplicate ($AAAA$) individuals, and j tracks the number of single-duplicate individuals ($AAA-$). In this case, the non-zero off-diagonals of the generator $\boldsymbol{Q}$ are given

by

$$q_{(i,j)(k,l)} = \begin{cases} \frac{(N-i-j)}{N} p_b(AAAA|(i,j)) & \text{for } k = i+1, l = j \\ \frac{j}{N} p_b(AAAA|(i,j)) & \text{for } k = i+1, l = j-1 \\ \frac{i}{N} p_b(AAA-|(i,j)) + 2i\mu_p & \text{for } k = i-1, l = j+1 \\ \frac{(N-i-j)}{N} p_b(AAA-|(i,j)) & \text{for } k = i, l = j+1 \\ \frac{i}{N} p_b(AA--|(i,j)) & \text{for } k = i-1, l = j \\ \frac{j}{N} p_b(AA--|(i,j)) + j\mu_p & \text{for } k = i, l = j-1 \end{cases} \tag{8}$$

where $p_b(X|(i,j)$ denotes the probability that when a birth occurs it is of an individual with genotype X given that the current state is (i,j). We assume that the population is monoecious (equivalent to a dioecious population without sex biased allele frequencies), that individuals cannot mate with themselves, and that the offspring receives a copy of the gene from each parent at each of the original and duplicated locus, including the possibility of inheriting the absence of any gene at that locus.

$$p_b(X|(i,j) = \begin{cases} \frac{i}{N}\frac{(i-1)}{(N-1)} + \frac{1}{4}\frac{j}{N}\frac{j-1}{N-1} + \frac{j}{N}\frac{i}{N-1} & \text{for } X = AAAA \\ \frac{j}{N}\frac{N-j-i}{N-1} + \frac{j}{N}\frac{i}{N-1} + \frac{1}{2}\frac{j}{N}\frac{j-1}{N-1} + 2\frac{i}{N}\frac{N-j-i}{N-1} & \text{for } X = AAA- \\ \frac{N-j-i}{N}\frac{N-j-i-1}{N-1} + \frac{j}{N}\frac{N-j-i}{N-1} + \frac{1}{4}\frac{j}{N}\frac{j-1}{N-1} & \text{for } X = AA-- \end{cases} \tag{9}$$

Recall that under our model, replacement of individuals occurs at Poisson rate 1 and thus the time to replace N individuals is Erlang distributed, and has expectation N. We therefore say that the 'generation time' under the model is N, but note that this is the time expected to replace N individuals, not the expected time after some time t to replace all individuals who were present at t. The expected time for any particular individual to be replaced is $1/N$.

Note also that the models do not include new duplication events. Rather it is assumed that we start with a duplicate copy and track the subsequent evolution of the population assuming no further duplication events occur. A consequence of this is that the models have an absorbing structure, and permanent fixation of the duplicate is only possible if $\mu_p = 0$. When $\mu_p > 0$ the new duplicate must eventually go extinct since in this case only the state $(0,0)$ is absorbing. However selection acts to increase (or decrease) the timescale over which a duplicate segregates in the population. In the case of strong selection, the time it takes for the duplicate to go extinct could be much larger than the timescales under consideration here.

Testing for selection

To test for selection, we evaluate the 95% prediction band for the proportion of haploids carrying a duplicate copy under the reduced (neutral) model, given that the duplicate copy has not yet gone extinct. If the observed proportion of duplicates in a population falls outside of this region, we can reject the hypothesis of neutrality at the 95% significance level, under the assumptions of our models. Further, we can in principle calculate the prediction bands and expectation for non-neutral models in order to gauge the power of the test to detect selection. We can evaluate the probability of false negatives for given fitness (under the modelling assumptions) by calculating the probability that the proportion of haploids carrying a duplicate copy in the model with selection falls within the prediction bands of the neutral model. However, since the models with selection are less computationally tractable than the neutral population, this is only possible for small population sizes at present. We anticipate that approximations will be forthcoming that allow this to be extended to large populations. However at the time of publication we have not investigated such approximations.

One important consideration is the tuning of the parameter u_p. The reduced model for the neutral case has only two parameters, N, and u_p. The parameter N is likely to be known to reasonable accuracy. Treating N as fixed, u_p is solely responsible for tuning the model behaviour, and its value will be dependent on the organism and gene under study. A reasonable estimate of u_p can be obtained by considering the target size for pseudogenizing mutations and becomes a scalar from the background substitution rate dS in the organism and loci under study. Compared to the mutational opportunity for synonymous substitutions per synonymous site (see for example, Nei and Gojobori (1986)), the number of mutations that would introduce an early stop codon leading to a nonfunctional truncated protein, cause a protein to not fold, hit a functional residue, or hit a core region of the promoter sequence that affects all expression domains could be quantified. For our purposes here, this number was set at 35, but a more precise and specific estimate will ultimately be necessary when fitting the model to genomic data. The background substitution rate, dS, could be evaluated for example using PAML for any pair of duplicates (Yang, 2007). It is also, in principle, possible to obtain an empirical rather than a theoretical estimate of the pseudogenization rate when full population (genomic and transcriptomic) sequencing exists by observing and counting the number of pseudogenized copies of the duplicate that are segregating.

The models described here, particularly the diploid model with selection, have very large sparse generator matrices which require special computational consideration. The software package expokit (Sidje, 1998) largely solves these problems for small N (on the order of 1000 in the diploid case), and we used a modified version of the MATLAB implementation of expokit for our analysis. Work is underway to find suitable approximations for large N, but the test for selection itself is tractable in the haploid case for $N = 100,000$ and in the diploid case for $N = 10,000$ thanks to the reduced neutral models discussed

above and the use of expokit.

3 Results

A test for selection on copy number variants in a population has been devised. The neutral expectation and its confidence intervals have been generated for both haploid (N=1000, 100,000) and diploid populations (N=1000). Under a range of selective conditions, both individual realizations and the expectation are presented for both types of population. We also explicitly calculate the probability of correctly rejecting the neutral hypothesis for samples of an example haploid population taken at different times since the duplication event.

Realizations in a Diploid Population

We calculated the conditional (on survival of the duplicate) expected proportion and 95% prediction bands for a diploid population of $N = 1000$ individuals (2000 haploid genomes) with a very low rate of pseudogenization ($\mu_p = 10^{-10}$). Figure 2 shows this example overlaid with 10 simulated sample paths with the same parameters, but with a high rate of highly beneficial neofunctionalization ($u_n = 10^{-6}, s_n = 0.1$). In all 10 simulations the duplicate fixes in the population sooner or later. The shaded region shows the neutral prediction band, outside of which our test for selection rejects the hypothesis of neutrality. After 100 generations ($dS \approx 3 \times 10^{-7}$) half of the simulations have reached duplicate proportions outside of the prediction band, indicating that a sample taken from these populations after this time would reject neutrality. The remaining simulations reach such proportions around by the time of the 200^{th} generation ($dS \approx 6 \times 10^{-7}$).

In this example the rate of neofunctionalization is fast compared to pseudogenization (and hence also compared to our timescale dS, which scales with u_p). The initial neofunctionalization happened quickly in each case, and the neofunctionalized copies quickly spread through the population. The low rate of pseudogenization also results in the prediction band for the neutral case becoming very wide very quickly, but the high rate of neofunctionalization provides a window during which selection is easily detected.

Realizations in a Haploid Population

We repeated the simulation procedure for a haploid population of $N = 10^5$ individuals with a moderately high pseudogenization rate ($u_p = 10^{-7}$). with relatively small selective effects ($f_d = 1.01, f_n = 1.02$). In this case the prediction band stays fairly narrow, however the majority of the 10 sample paths remain inside the prediction band, meaning that we would fail to detect the selection in most of these simulated populations. Figure 3 shows the results of this simulation.

In contrast, when we simulated another 10 sample paths with $f_d = 1.05, f_n = 1.1$, 9 out of 10 sample paths were well outside of the prediction interval by time $ds = 0.025$, as shown in Figure 4. Furthermore, setting $f_d = 0.95$ resulted in an even more pronounced effect, with all 10 sample paths quickly exceeding the upper prediction interval by several orders of magnitude and the largest proportion of duplicates reached by the end of the simulation being ≈ 0.5. The results for this example are shown in Figure 5

Power to detect selection

To better gauge the power of the test, we calculated conditional expectations and prediction bands for the model with selection under some different parameterizations. Figure 6 shows several examples. We can see from Figure 6 that the test is unlikely to detect negative selection acting on copy number alone, as the two different prediction bands almost entirely overlap. However, negative selection on copy number increases the probability of detecting selection in the presence of beneficial neofunctionalization, as can be seen in the middle columns of figure 6. As selection on the copy number becomes positive, we see a significant difference in the two regions, indicating that our test is likely to be able to detect selection. For times after which the two regions no longer overlap, there is a 95% chance that samples from the selected population would fall outside the 95% prediction band of the neutral model, and the test would thus correctly reject the hypothesis of neutrality. Increasing the selective benefit of neofunctionalization results in a similar picture. Where the two prediction bands overlap we can expect the test to be somewhat unreliable, and to correctly reject neutrality in less than 95% of samples.

In terms of the biology underlying the parameterization, $f_d < 1$ reflects the biosynthetic cost of maintaining an extra copy in the genome (Wagner, 2005) whereas $f_d > 1$ reflects a fitness advantage from the linear response region in Figure 1, as was seen with extra copies of amylase in the human population associated with eating a high starch diet (Perry et al., 2007). $f_n > 1$ reflects an advantage to a new function based upon a new mutation arising. The case where $f_d > f_n$ reflects the case where the new function associated with the new mutation provides less of an advantage than extra copies of the original gene, associated presumably with a greater concentration of the encoded protein in relevant cell types.

The power of the test to detect selection of a given magnitude at any point in time can also be calculated explicitly by finding the probability that a sample taken at that time will fall within the prediction band of the neutral model. Figure 7 shows a graph of the probability that a sample taken from a haploid population with parameters $N = 1000, f_d = 0.9, f_n = 1.1, u_p = 10^{-5}, u_n = 10^{-6}$ will result in the rejection of neutrality under our test. The shape of the curve is similar for other parameters, with the strength of selection (and population size) being the main determining factors for how long a duplicate must have been segregating before the test becomes reliable.

4 Discussion

A new approach for characterizing selection on segregating gene duplicates (copy number variants of CNVs) has been established based upon the expected relationship between an allele's age and its frequency in a population when segregating neutrally. This approach characterizes selection on the duplicate copy itself rather than on mutations that occur within the duplicates while segregating. While many models for duplicate gene retention assume that fixation of the duplicate copy occurs before fate determining mutations under selection begin to act (Innan & Kondrashov, 2010), this assumption may be violated frequently for a number of reasons, most particularly when mutation rates and/or effective population sizes are large. These are the scenarios when this method has particular power to reject neutrality.

Similar approaches have recently been applied to characterize selection on SNPs segregating in a population (Platt et al., 2019). In those scenarios, characterizing the age of an allele depends upon characterizing tracts of identity by descent. Here, characterization of allele (duplicate) age is much simpler, relying only upon the pairwise dS value between the copies. More complex schemes to examine selection on CNVs have been presented (Itsara et al., 2010), but use orthogonal information to that used in this method.

In this paper, in addition to presenting the neutral baseline and prediction bands about it, we have analyzed the statistical power of the test under a number of simple but realistic selective schemes and presented cases where one would expect to have the power to reject neutrality. The cases where one expects this method to have sufficient power include population sizes much smaller than would be expected for most species of interest to the population ecology/ecological genomics community.

While the approach presented here is an exact solution that has not yet reached population sizes that are reflective of those for many eukaryotic populations of interest, approximations to the neutral baseline are currently under development that will enable generation of the test statistic to compare to population ecological genomic data for any species of interest. The work here lays the basic science foundations for such a future application.

5 Acknowledgments

We would like to thank Ryan Houser for careful reading of an early version of the manuscript and for helpful discussions. We thank Gene Maltepes for computational support. We also thank Catherine Browne for technical assistance in the preparation of the manuscript.

6 Data Accessibility

No original research data was presented in this paper. Code used to perform the analysis will be made available at https://github.com/TristanLStark/DetectingSelection.

7 Author Contributions

This study was conceived by DAL and TLS. Modeling and theoretical results were generated by TLS and RSK. Computer code for simulations was written and run by TLS, RSK, and MAM. The manuscript was written by DAL, TLS, RSK, and MAM.

References

Anisimova, M., & Liberles, D. (2012). Detecting and understanding natural selection. In G. M. Cannarozzi & A. Schneider (Eds.), *Codon evolution: Mechanisms and models* (pp. 73–96). Oxford University Press.

Arvestad, L., Lagergren, J., & Sennblad, B. (2009). The gene evolution model and computing its associated probabilities. *J. ACM*, *56*(2). https://doi.org/10.1145/1502793.1502796

Bornholdt, D., Atkinson, T. P., Bouadjar, B., Catteau, B., Cox, H., De Silva, D., ... Grzeschik, K.-H. (2013). Genotype–phenotype correlations emerging from the identification of missense mutations in MBTPS2. *Human Mutation*, *34*(4), 587–594. https://doi.org/10.1002/humu.22275

Conant, G. C., & Wagner, A. (2003). Asymmetric sequence divergence of duplicate genes. *Genome Research*, *13*(9), 2052–2058. https://doi.org/10.1101/gr.1252603

De Sanctis, B., Krukov, I., & de Koning, A. J. (2017). Allele age under non-classical assumptions is clarified by an exact computational markov chain approach. *Scientific Reports*, *7*(1), 1–11. https://doi.org/10.1038/s41598-017-12239-0

Force, A., Lynch, M., Pickett, F. B., Amores, A., Yan, Y.-L., & Postlethwait, J. (1999). Preservation of duplicate genes by complementary, degenerative mutations. *Genetics*, *151*(4), 1531–1545.

Gonzalez, E., Kulkarni, H., Bolivar, H., Mangano, A., Sanchez, R., Catano, G., ... Ahuja, S. K. (2005). The influence of CCL3L1 gene-containing segmental duplications on HIV-1/AIDS susceptibility. *Science*, *307*(5714), 1434–1440. https://doi.org/10.1126/science.1101160

Guerin, M. N., Weinstein, D. J., & Bracht, J. R. (2019). Stress adapted mollusca and nematoda exhibit convergently expanded hsp70 and AIG1 gene families. *Journal of Molecular Evolution*, *87*(9-10), 289–297. https://doi.org/10.1007/s00239-019-09900-9

Hughes, A. L. (1994). The evolution of functionally novel proteins after gene duplication. *Proceedings of the Royal Society of London B: Biological Sciences*, *256*(1346), 119–124. https://doi.org/10.1098/rspb.1994.0058

Innan, H., & Kondrashov, F. (2010). The evolution of gene duplications: Classifying and distinguishing between models. *Nature Reviews Genetics*, *11*(2), 97–108. https://doi.org/10.1038/nrg2689

Itsara, A., Wu, H., Smith, J. D., Nickerson, D. A., Romieu, I., London, S. J., & Eichler, E. E. (2010). De novo rates and selection of large copy number variation. *Genome Research*, *20*(11), 1469–1481. https://doi.org/10.1101/gr.107680.110

Katju, V., & Lynch, M. (2006). On the formation of novel genes by duplication in the caenorhabditis elegans genome. *Molecular Biology and Evolution*, *23*(5), 1056–1067. https://doi.org/10.1093/molbev/msj114

Konrad, A., Teufel, A. I., Grahnen, J. A., & Liberles, D. A. (2011). Toward a general model for the evolutionary dynamics of gene duplicates. *Genome Biology and Evolution*, *3*, 1197–1209. https://doi.org/10.1093/gbe/evr093

Lauer, S., Avecilla, G., Spealman, P., Sethia, G., Brandt, N., Levy, S. F., & Gresham, D. (2018). Single-cell copy number variant detection reveals the dynamics and diversity of adaptation. *PLoS Biology*, *16*(12), e3000069. https://doi.org/10.1371/journal.pbio.3000069

Lynch, M., & Force, A. (2000). The probability of duplicate gene preservation by subfunctionalization. *Genetics*, *154*(1), 459–473.

Lynch, M., & Force, A. G. (2000). The origin of interspecific genomic incompatibility via gene duplication. *The American Naturalist*, *156*(6), 590–605. https://doi.org/10.1086/316992

Lynch, M., O'Hely, M., Walsh, B., & Force, A. (2001). The probability of preservation of a newly arisen gene duplicate. *Genetics*, *159*(4), 1789–1804.

Maruyama, T. (1974). The age of an allele in a finite population. *Genetics Research*, *23*(2), 137–143. https://doi.org/10.1017/S0016672300014750

Moran, P. A. P. (1958). Random processes in genetics. *Mathematical Proceedings of the Cambridge Philosophical Society*, *54*(1), 60–71. https://doi.org/10.1017/S0305004100033193

Nei, M., & Gojobori, T. (1986). Simple methods for estimating the numbers of synonymous and nonsynonymous nucleotide substitutions. *Molecular Biology and Evolution*, *3*(5), 418–426. https://doi.org/10.1093/oxfordjournals.molbev.a040410

Ohno, S. (1970). The enormous diversity in genome sizes of fish as a reflection of nature's extensive experiments with gene duplication. *Transactions of the American Fisheries Society*, *99*(1), 120–130. https://doi.org/10.1577/1548-8659(1970)99⟨120:TEDIGS⟩2.0.CO;2

Perry, G. H., Dominy, N. J., Claw, K. G., Lee, A. S., Fiegler, H., Redon, R., ... Stone, A. C. (2007). Diet and the evolution of human amylase gene copy number variation. *Nature Genetics*, *39*(10), 1256–1260. https://doi.org/10.1038/ng2123

Platt, A., Pivirotto, A., Knoblauch, J., & Hey, J. (2019). An estimator of first coalescent time reveals selection on young variants and large heterogeneity in rare allele ages among human populations. *PLoS Genetics*, *15*(8), e1008340. https://doi.org/10.1371/journal.pgen.1008340

Sidje, R. B. (1998). Expokit: A software package for computing matrix exponentials. *ACM Transactions on Mathematical Software (TOMS)*, *24*(1), 130–156. https://doi.org/10.1145/285861.285868

Stark, T. L., Liberles, D. A., Holland, B. R., & O'Reilly, M. M. (2017). Analysis of a mechanistic markov model for gene duplicates evolving under subfunctionalization. *BMC Evolutionary Biology*, *17*(1), 1–16. https://doi.org/10.1186/s12862-016-0848-0

Tofigh, A., Hallett, M., & Lagergren, J. (2010). Simultaneous identification of duplications and lateral gene transfers. *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, *8*(2), 517–535. https://doi.org/10.1109/TCBB.2010.14

Wagner, A. (2005). Energy constraints on the evolution of gene expression. *Molecular Biology and Evolution*, *22*(6), 1365–1374. https://doi.org/10.1093/molbev/msi126

Yang, Z. (2007). PAML 4: Phylogenetic analysis by maximum likelihood. *Molecular Biology and Evolution*, *24*(8), 1586–1591. https://doi.org/10.1093/molbev/msm088

Yohe, L. R., Liu, L., Dávalos, L. M., & Liberles, D. A. (2019). Protocols for the molecular evolutionary analysis of membrane protein gene duplicates. In T. Sikosek (Ed.), *Computational methods in protein evolution* (pp. 49–62). Humana Press. https://doi.org/10.1007/978-1-4939-8736-8_3

Zhang, C., Zhang, C., Chen, S., Yin, X., Pan, X., Lin, G., ... Wang, W. (2013). A single cell level based method for copy number variation analysis by low coverage massively parallel sequencing. *PloS ONE*, *8*(1), e54236. https://doi.org/10.1371/journal.pone.0054236

Zhang, J. (2003). Evolution by gene duplication: An update. *Trends in Ecology & Evolution*, *18*(6), 292–298. https://doi.org/10.1016/S0169-5347(03)00033-8

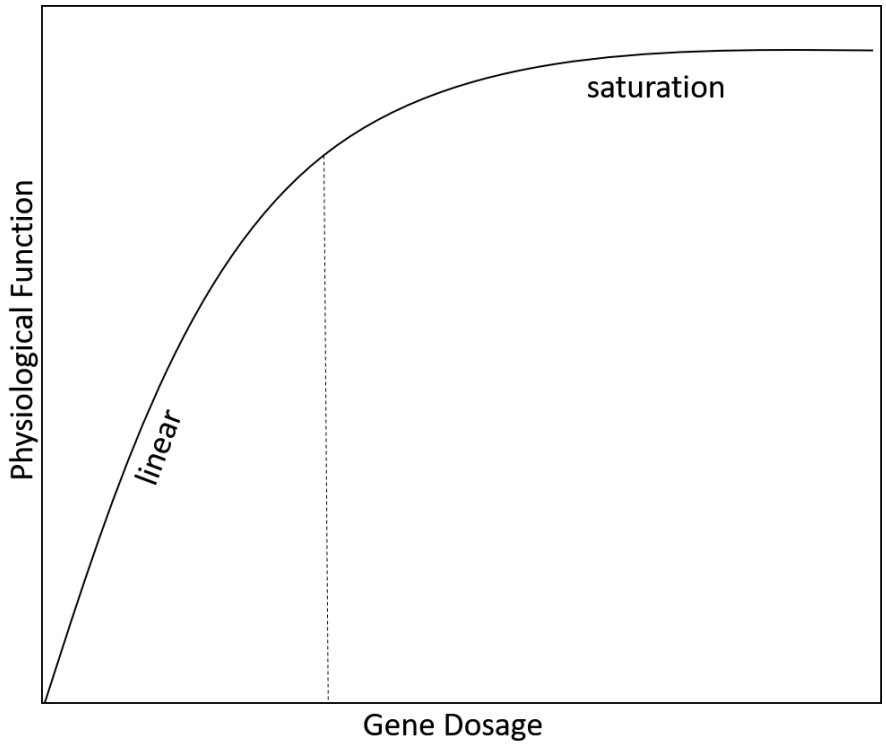

Figure 1: A conceptual diagram of the relationship between physiological function (where fitness and selection relate directly to gene product concentration) and gene dosage (hence, copy number).

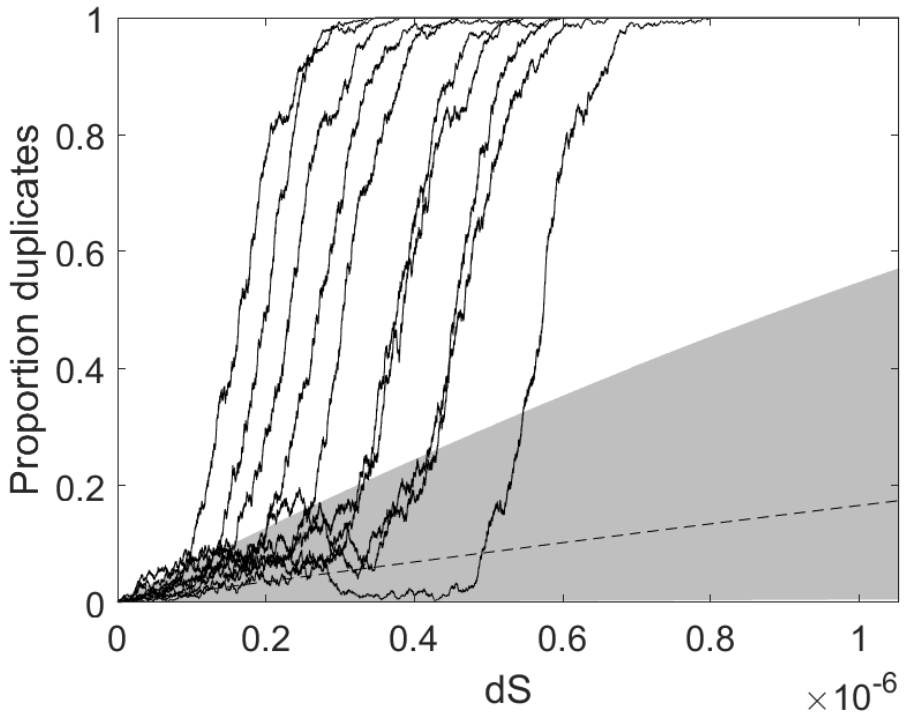

Figure 2: Conditional expected proportion (dashed line) of haploid genomes in a diploid population of $N = 1000$ individuals with a duplicated copy under neutrality ($u_p = 10^{-10}$), overlaid with 10 simulated populations experiencing positive selection ($u_n = 10^{-6}, s_n = 0.1$). The neutral prediction band is shaded.

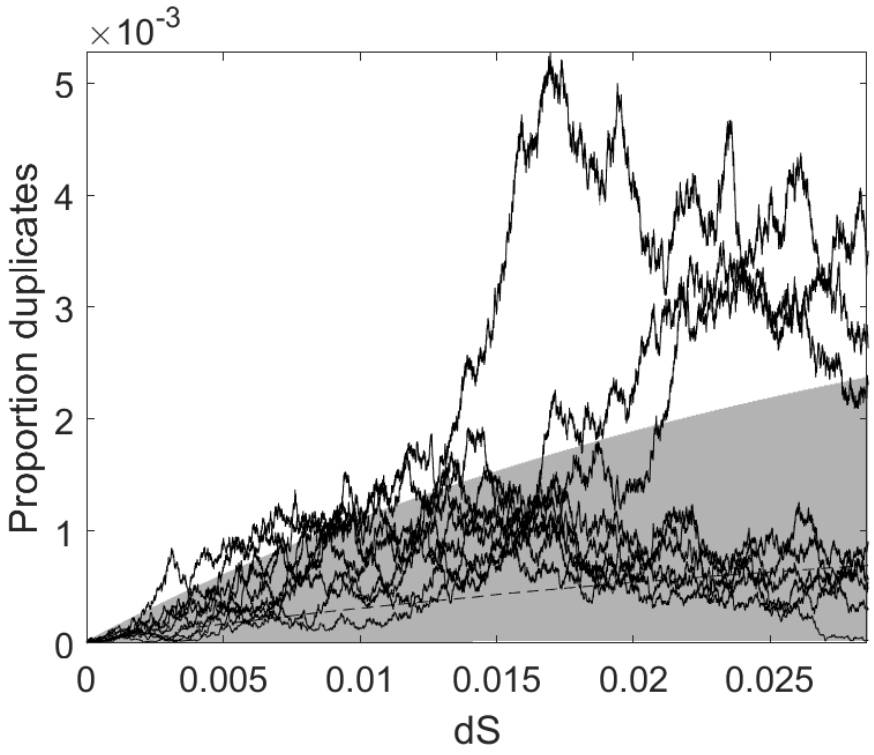

Figure 3: Conditional expected proportion of haploid genomes (dashed line) in a haploid population of $N = 10^5$ individuals with a duplicated copy under neutrality ($u_p = 10^{-7}$). Overlaid are 10 simulated populations experiencing mild dosage selection and with a moderate rate of mildly beneficial neofunctionalization ($u_n = 10^{-8}, f_d = 1.01, f_n = 1.02$). The neutral prediction band is shaded.

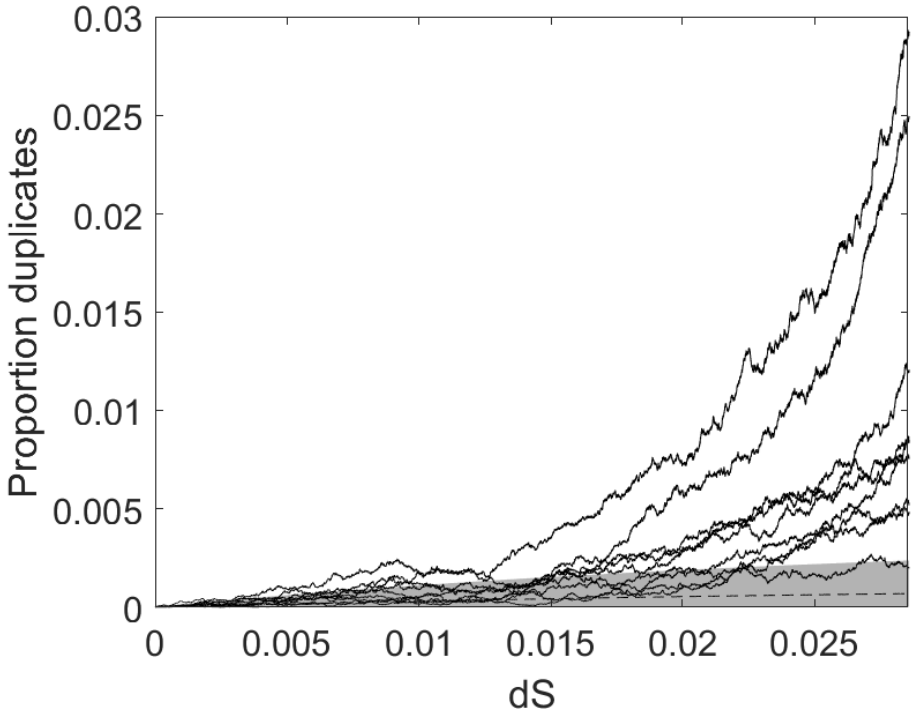

Figure 4: Conditional expected proportion of haploid genomes (dashed line) in a haploid population of $N = 10^5$ individuals with a duplicated copy under neutrality ($u_p = 10^{-7}$). Overlaid are 10 simulated populations experiencing moderate dosage selection and with a moderate rate of highly beneficial neofunctionalization ($u_n = 10^{-8}, f_d = 1.05, f_n = 1.1$). The neutral prediction band is shaded.

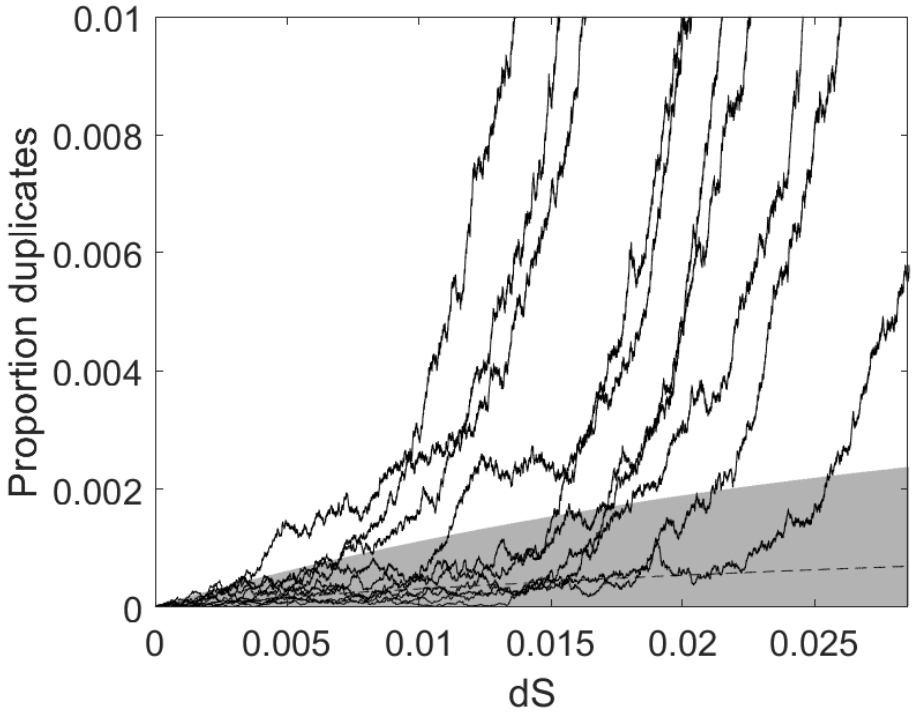

Figure 5: Conditional expected proportion of haploid genomes (dashed line) in a haploid population of $N = 10^5$ individuals with a duplicated copy under neutrality ($u_p = 10^{-7}$). Overlaid are 10 simulated populations experiencing moderate dosage selection and with a moderate rate of highly beneficial neofunctionalization ($u_n = 10^{-8}, f_d = 0.955, f_n = 1.1$). The neutral prediction band is shaded.

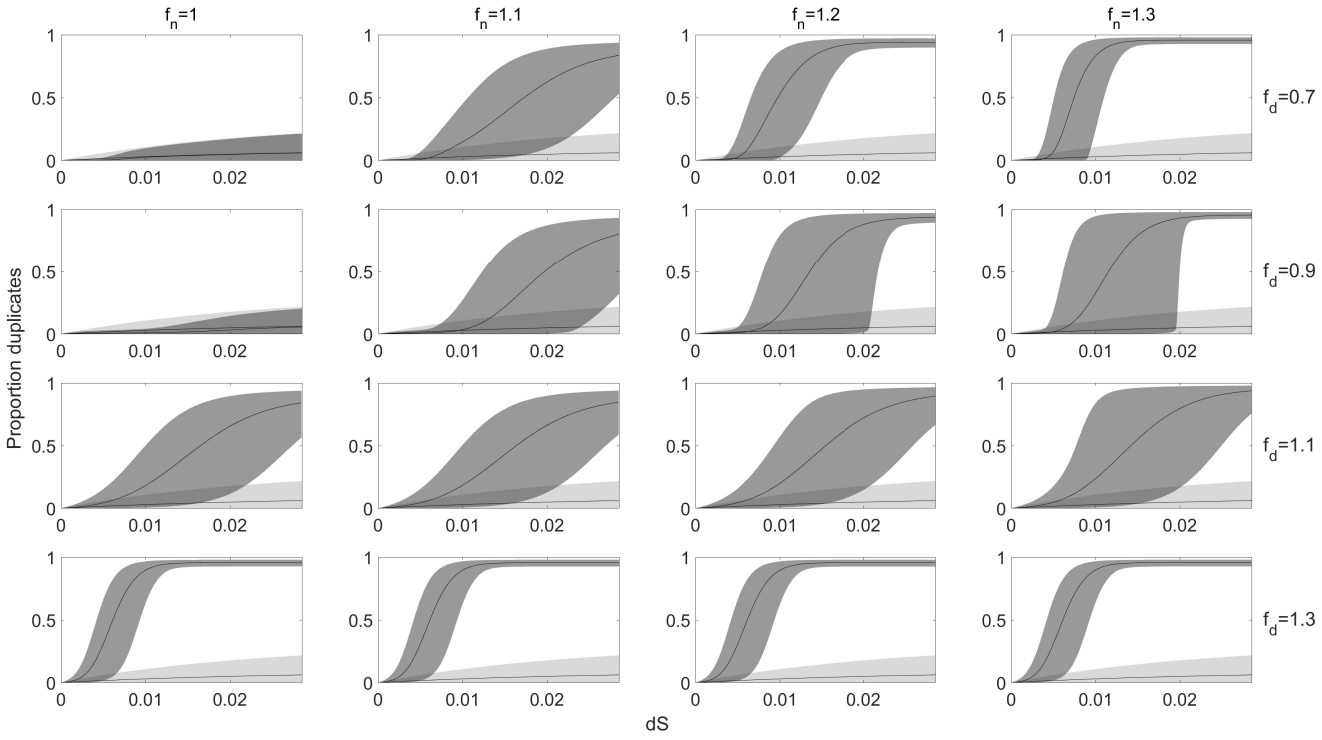

Figure 6: Comparison of conditional expectations and prediction bands between the neutral haploid model and haploid model with selection for several fitness parameters ($N = 1000, \mu_p = 10^{-5}, \mu_n = 10^{-6}$). f_n is increasing in the columns, while f_d is increasing in the rows.

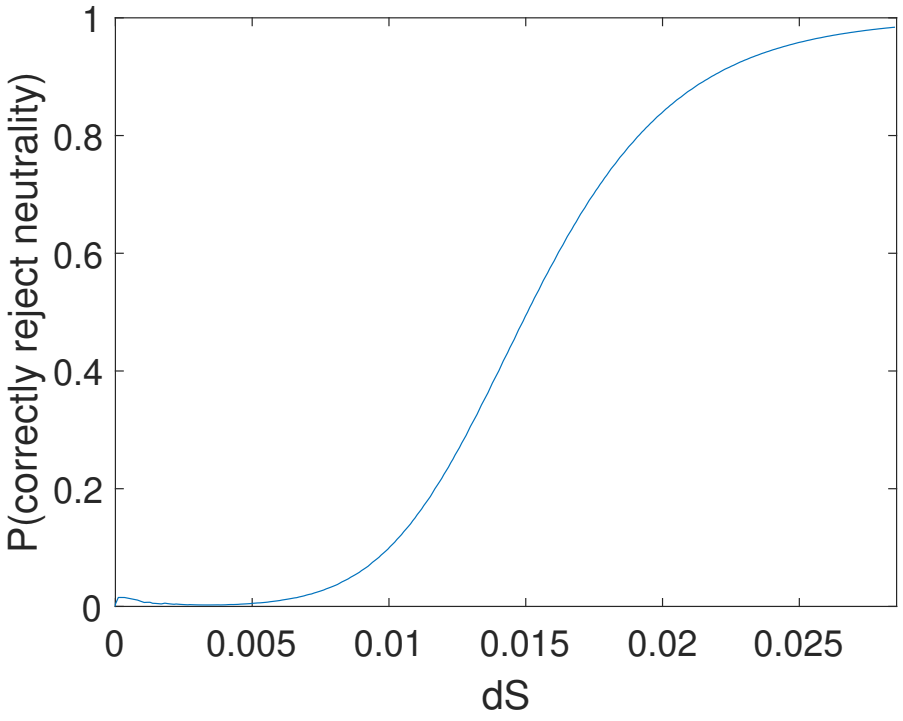

Figure 7: Probability of correctly rejecting neutrality under the haploid model with $N = 1000, f_d = 0.9, f_n = 1.1, u_p = 10^{-5}, u_n = 10^{-6}$.