4-19-2013

# A Stochastic Model for the Development of Complex Bateson- Dobzhanksy-Muller Incompatibilities Incorporating Protein-Protein Interaction Network Structure

Andrius Jonas Dagilis
*Trinity University*, adagilis@trinity.edu

# A Stochastic Model for the Development of Complex Bateson-Dobzhanksy-Muller Incompatibilities Incorporating Protein-Protein Interaction Network Structure

Andrius Jonas Dagilis

A departmental senior thesis submitted to the Department of Biology at Trinity University in partial fulfillment of the requirements for graduation with departmental honors.

April 19, 2013

Thesis Advisor

Department Chair

Associate Vice President
for
Academic Affairs

# A Stochastic Model for the Development of Complex Bateson-Dobzhansky-Muller Incompatibilities Incorporating Protein-Protein Interaction Network Structure

Andrius Jonas Dagilis

April 18, 2013

**ABSTRACT: Bateson-Dobzhansky-Muller Incompatibilities are theorized to be one of the main causes of post-zygotic reproductive isolating barriers. While many quantitative models for the development of such incompatibilities exist, few have dealt with the effect the structure of protein-interaction networks has on the development of incompatibilities, and none deal with incompatibilities between more than two loci. We develop a stochastic model for the development of complex Bateson-Dobzhansky-Muller incompatibilities that incorporates the structure of protein-protein interaction networks, and demonstrate that the structure of the network can drastically change the importance of digenic interactions toward the rate of divergence between populations.**

## Introduction

### Speciation

Under the Biological Species Concept (Mayr, 1942), species are defined by reproductive isolation. While alternative definitions have been proposed, most modern definitions of species generally appeal to our intuition that the populations in question are evolving independently (see De Queiroz (2007) for a review). One of the main ways two populations can be forced to evolve independently, especially if they reproduce sexually, is if they are separated by some kind of reproductive isolating barrier (RIB). Consequently, probably the most active area of speciation research is the development of RIBs. There are multiple models as to how populations may become reproductively isolated, whether by pre or post-zygotic mechanisms. One of the main models explaining how genetics may underlie the development of post-zygotic RIBs is the model formulated in the early to mid 1900s by Bateson (1909), Dobzhansky (1937) and Muller (1942). This model attempts to explain how combinations of genes can give rise to reproductive incompatibility in a hybrid, leading to either hybrid inviability or infertility. It assumes directly that there are two loci with two different alleles that are incompatible by some means, allowing an RIB to develop from their interaction, called a

Bateson-Dobzhansky-Muller Incompatibility (BDMI). This idea has since been extended from digenic, diallelic BDMIs to interactions between multiple loci with multiple possible alleles at each locus. Empirical studies have identified the genes involved in the barriers separating species of many taxa, most prominently in *Drosophila* (some cases in Coyne and Orr (1989), Coyne and Orr (1998), Presgraves (2003)), in *Mimulus* ( Christie and Macnair (1987)) , in *Arabidopsis* (Bikard *et al.* (2009), Burkart-Waco *et al.* (2012)), and in *Sacharomyces* (Kao *et al.* (2010), Anderson *et al.* (2010)). While these studies have been greatly illuminating as to the mechanisms and frequency of BDMIs, examples of BDMIs have largely been limited to highly studied taxa, and few genes interacting to cause BDMIs have been pinpointed. This is largely because it is difficult to perform empirical studies of post-zygotic isolation in highly diverged sister species, or in taxa showing large genetic or phenotypic divergence, as zygotes can be difficult if not impossible to obtain in such cases, leading the majority of the work to be done in close taxa that have speciated relatively recently. While the amount and accessibility of genetic data is constantly increasing, there are still limits to how we can study BDMIs in the lab, especially in regards to the variety of organisms and levels of divergence.

## Theoretical Models

Another fertile avenue for speciation research is theoretical modeling. While studies of natural populations are inflexible when it comes to obtaining data for populations at different stages of differentiation, theoretical models provide us with some insight as to what we would expect to see at the points that data may be hard to obtain for. While the verbal BDMI model had been more or less fully formulated by the late 1940s, the main quantitative model was only put forward by H.A. Orr in the middle of the 1990s (Orr, 1995). This model assumes that two allopatric populations are fixing novel alleles independently, with at most one mutation occurring at each locus. Given that BDMIs can arise when several alleles that have not previously been seen in the wild in combination do occur in the hybrid, the $K^{th}$ fixed substitution is expected to have $K-1$ alleles in its background that it has not been tested with. These can either be novel alleles fixed in the second population, or simply ancestral genes that have not occurred in the background of the $K^{th}$ mutation. Thus the total number of possible BDMIs, or the sum of all the possible BDMIs up to the $K^{th}$ fixed substitution, grows as the binomial coefficient $\binom{K}{2}$. Orr then derives the following probability of speciation:

$$S(K) = 1 - (1-p)^{\binom{K}{2}} \tag{1}$$

Here, $p$ is the probability of a BDMI between any two loci, and $S(K)$ is the probability of speciation after $K$ mutations have been fixed in either population. One of the most important ideas that has stemmed from this model is the "snowball effect" of the probability of speciation: as the populations diverge, the probability of speciation increases exponentially rather than linearly with respect to the number of fixed substitutions in either of the populations. This model of speciation has been highly influential, and many other BDMI models are based off its general structure (see Gavrilets (2003), Kirkpatrick and Ravigné (2002), Turelli *et al.* (2001) for some reviews.) Many such models focus on very specific circumstances for speciation to occur, leading some to claim that theory has led to "balkanized" models (Kirkpatrick and Ravigné (2002)). Some of the more general models do stand out in how they relate theory of BDMIs with other biological concepts. Orr and Orr (1996) investigated the effects of population subdivisions to rate of BDMI acquisition, showing that the rate of BDMI acquisition does not depend on population subdivisions when considering evolution through genetic drift, but the rate is decreased by increasing subdivisions when genes evolve by natural selection. Gavrilets (2004) explored the BDMI model extensively in relation to the possible structure of fitness landscapes, which allowed him to refine the concept of fitness landscapes to adaptive holey landscapes, and produced a wide array of results concerning the timing of speciation events. Kondrashov (2002) incorporated further theory of protein evolution into the BDMI model, leading to closer estimates of the fraction of substitutions in proteins that are compensated pathogenic deviations. Welch (2004) developed a combinatorial model highly similar to Orr's but included the number of introgressed genes and accounted for asymmetry in BDMIs, showing that the Orr model is capable of explaining why BDMIs tend to be asymmetrical and reaffirming some of the results from a slightly different interpretation of the verbal model.

As noted, many models of speciation are suited for very specific conditions only. While this does not generally detract from their findings, ideally a more powerful model would be able to account for different patterns of divergence in vastly different model systems, working for various taxa and conditions of divergence. At the same time, the benefit of the basic BDMI model is that the prevalence of models based on Orr's 1995 work means that the basic model can be extended to account for quite a few factors without too much difficulty as many of its extensions are compatible. In other words, a general model for the acquisition of BDMIs will likely be a result of extending Orr's model to account for the factors that are found to be significant.

## Effects of Protein Interactions

The core assumption of BDMIs is that the result of genes interacting is the cause some kind of incompatibility in the hybrid of two populations. Thus, any model of BDMIs must assume that the causal genes are interacting in order for the BDMI to occur, and empirical studies generally verify this intuition (Presgraves and Stephan (2007)). In Orr's model, the $K^{th}$ fixed mutation is given $K-1$ possible BDMI partners, with equal probability that any of the pairs cause an incompatibility. This assumption implies that all genes interact. However, research into the nature of protein-protein interactions (PPI) has shown that this is not the case (Bansal *et al.* (2009), Hahn and Kern (2005), Zotenko *et al.* (2008), Jordan *et al.* (2003)). If we think of protein interactions as a graphical model, with the protein network represented by an undirected graph with binary edges (see Appendix I for details), Orr's model looks at a complete network, where all proteins interact with all other proteins. Most other models of BDMIs follow similar assumptions to Orr's and either do not account for, or highly simplify, the kind of interactions that can happen. Part of the rationale may be that the probability of a BDMI between any two loci, $p$, would be estimated from real world data and may capture the probability that two genes interact within it. However we do know some details of what PPI-networks look like. Protein interaction networks have shown structure similar to power-law networks, where a few loci have very many interactions, whereas most loci have very few interactions (Hahn and Kern (2005)). It is unclear in such cases whether averages obtained from spotting several BDMIs are meaningful, as the number of possible interactions seems like it would depend highly on which loci fix substitutions, ignoring the remainder of the PPI network. Furthermore, as different species might have highly different PPI network structures, models fit for one set of taxa may not work well for a different set. Thus, a model that accounts for PPI network structure may be much more general than many of the current models, and may provide insights into some of the patterns seen in studies of speciation. While Orr and Turelli (2001) have previously attempted to show that breaking up the network into sub-networks does not have strong effects to initial conclusions gleamed from the original model, their results rely on each of the subnetworks being a complete network itself. Furthermore, it is generally believed that biological networks are connected, so splitting them into subnetworks removes at least some of the possible interactions from the picture. In a simulation based study, Palmer and Feldman (2009) attempt to incorporate genetic network structure somewhat by modeling how sets of genes can be regulated by each other's products, but account for the probability of interaction between the physical products simply by limiting the total number of possibly interacting loci, and do not consider the actual structure of the PPI-network. How PPI network structure affects BDMI development has therefore remained a relatively open and interesting research question.

We do know, however, that the probability of speciation will at the very least depend on the number of interactions between loci with fixed substitutions. As mutations are fixed at random, let $X(K)$ be a random variable equal to the number of interactions after $K$ fixed mutations. Modifying Equation 1 we obtain the following:

$$S(K) = 1 - (1 - p)^{X(K)}$$

$X(K)$ certainly depends on the PPI network of the common ancestor, but more importantly it depends on where on the network substitutions occur, so we want to formalize the concept of the network of fixed substitutions. Let $\mathcal{P} = (\mathbf{G}, \mathbf{I})$ be the PPI network of a species, where $\mathbf{G}$ is a set of the loci in the network, and $\mathbf{I}$ is a set of the interactions between loci. Now, let $\mathcal{M}(K) = (\mathbf{G}'(K), \mathbf{X}(K))$ be the interaction network after $K$ substitutions have been fixed, where $\mathbf{G}'(K)$ is the set of loci with fixed mutations and $\mathbf{X}(K)$ is a set of the interactions between these loci. These two abstract networks are quite useful as tools when thinking of how PPI-network structure could impact the rate of BDMI development. BDMIs can only occur between loci in the network of fixed mutations, which in turn only have interactions if these exist in the PPI network. Mathematically, we can think of $\mathcal{M}(K)$ as a network with $K$ nodes and randomly distributed edges. As a mutation is fixed on some locus, we choose an $i \in \{\mathbf{G} \setminus \mathbf{G}'(K-1)\}$ by some process and add this node to $\mathbf{G}'(K)$. The interactions between mutated nodes are then defined as the set $\mathbf{X}(K) = \{\{i, j\}$ such that $i, j \in \mathbf{G}'(K)$ and $\{i, j\} \in \mathbf{I}\}$, that is, the set of all the interactions that exist in the PPI network between the substituted loci. We assume for mathematical convenience that no new genes or interactions evolve, and that neither genes nor interactions are lost. We can then attempt to estimate $X(K)$ by modeling how many interactions are in $\mathcal{M}(K)$ with varying structures of $\mathcal{P}$. Under this model, $X(K)$ is equal to the magnitude of the set of edges of $\mathcal{M}(K)$. Since $\mathcal{M}(K)$ is a result of a random process, we model $X(K)$ as a random variable, and since the probability of speciation depends on $X(K)$, we present a stochastic model for the probability of BDMIs that incorporates PPI network structure. Previous research by Livingstone *et al.* focused on digenic BDMIs using such a stochastic model, and showed that the density of the PPI network directly impacts the rate of BDMI development (Livingstone *et al.* (2012)). More importantly, it created a framework for future research on the impacts of PPI network structure to the rate of speciation.

### Complex Interactions

An important topic in the study of speciation is one of incompatibilities between more than two loci. Such interactions have also been identified in several taxa (Cabot *et al.* (1994), Coyne and Orr (1989)). Past models, including Orr's original, predict far higher rates of complex interactions than simple ones (Orr (1995),Gavrilets (2003)). This has been postulated to occur largely due to the combinatorial nature of the number of possible interactions. If there are $\binom{K}{2}$ digenic interactions, then there are $\binom{K}{3}$ interactions between 3 loci, and so on. So, for large $K$, the complex interactions far outnumber digenic ones. Furthermore there are multiple pathways under which the genotypes causing speciation could evolve without hitting the specific combination that is not viable when considering complex incompatibilities, whereas for relatively simple incompatibilities there are fewer such possible paths. Few authors offer examples of the mechanisms by which such complex BDMIs could cause speciation, however several studies have identified either a lack of digenic interactions (Kao *et al.* (2010)) or the presence of complex incompatibilities (Cabot *et al.* (1994),Coyne and Orr (1989)). It is therefore interesting to see whether such interactions are affected by the structure of PPI networks. While our previous model (Livingstone *et al.* (2012)) does not consider complex interactions, it can be extended to do so. To estimate the number of complex interactions, we will need to look at the numbers of $n-$connected components in $\mathcal{M}(K)$ rather than the magnitude of $\mathbf{X}(K)$. The goal of this paper is to investigate how the structure of $\mathcal{P}$ affects the number of $n-$connected components in $\mathcal{M}(K)$, and how this in turn affects the probability of a BDMI developing between the two populations.

## Model and Methods

### General Model

Our general model is based on our earlier work focused on a model of digenic BDMIs in relation to PPI networks (Livingstone *et al.* (2012)). In both models we assume two populations are fixing substitutions independently, and only look at the probability of speciation after $K$ mutations in total have been fixed, making no assumptions either on the fitness of the mutations or their methods of fixation (although as noted in Appendix III.A, our model likely simulates genetic drift). To simplify some aspects of the mathematical assumptions, each locus may only fix one substitution in either population. To define the general case we let the probability of developing a BDMI be $1 - \varphi(K)$, where $\varphi(K)$ is the probability that no BDMIs have occurred after $K$ mutations.

In the digenic case, the limitation on the rate of speciation is the density of the network since the expected number of interactions after $K$ fixed mutations, $E[X(K)]$ is $\alpha_2\binom{K}{2}$, where $\alpha_2 = \frac{E}{\binom{N}{2}}$, $N = |\mathbf{G}|$ is the total number of proteins and $E = |\mathbf{I}|$ is the number of interactions in the network (Livingstone *et al.*, 2012). Letting $\varphi(K) \approx (1-p)^{\alpha_2\binom{K}{2}}$ be the probability that none of the $E[X(K)]$ interactions has caused a BDMI, the probability that one or more BDMIs occur, and the populations develop RBIs is:

$$S(K) \approx 1 - (1-p)^{\alpha_2\binom{K}{2}}$$

**Complex Interactions and Graphs:** Generalizing to complex incompatibilities between $n$ loci, we want to know how many different combinations of $n$ mutated loci are interacting. In terms of our graphical model we are looking at the number of $n-$connected components in $\mathcal{M}(K)$. The jump from digenic interactions to PPI networks is fairly simple - edges only exist if the interaction exist. This is not quite the same for more complex incompatibilities, as complex interactions in PPI-networks are not quite as straightforward. PPI networks are generally not tissue and timing specific, so the fact that edges exist between loci $i, j, k$ does not necessarily mean that these proteins interact as a triple at any point. To the contrary, if $i$ and $j$ are functional duplicates in different tissues, we fully expect to see a connected component in whole-organism with these nodes. However, it seems to be wrong to term this as an interaction between 3 proteins as two of them never actually see each other, unless $k$ is a molecule that moves between the tissues and a single molecule of $k$ may find itself being modified, say, by both $i$ and $j$ at some point. At the same time, limiting ourselves to tissue specific data may also be misleading - $i$ may be a regulator for the expression of $j$ in one tissue, which then travels to its target $k$ in an entirely different tissue. A tissue or timing specific PPI-network may miss such interactions entirely. In short, merely having a connected component in $\mathcal{P}$ does not translate immediately to the existence of a complex interaction, while limiting ourselves to tissue or timing specific data may highly underestimate the number of complex interactions. However, we do know that if a component does NOT exist in $\mathcal{P}$, then a BDMI is impossible as the probability that these proteins can interact in some fashion, in any tissues or at any time, is 0. In effect, we can use the whole-organism PPI network to limit the number of complex interactions, even if it is an overestimate. To counter this, we will make sure that the probability that an $n-$sized interaction causes a BDMI takes into account that an $n-$connected component may not represent an existing interaction (i.e. the probability that an interaction between $n-$loci cause a BDMI can be seen as a product of the probability of causing a BDMI and the probability that the interaction exists). Again, as $\mathcal{M}(K)$ is random, the number of $n-$connected components is a random variable which we will call

$CC_n(K)$.

**Probability of BDMIs:** We assume that an incompatibility between $n$ genes occurs independently of incompatibilities of other complexities and we define $\varphi_n(K)$ as the probability that no $n-$connected components cause a BDMI. Biologically, it may be difficult to verify such BDMIs, and may seem backward to say that a digenic BDMI gives no information about more complex ones. If a digenic incompatibility occurs, then all more complex interactions that include it as an edge should (at least via biological intuition) not work as well. However, we are not looking at the probability that an interaction between $n$ loci works, but rather the probability that such an interaction causes an RIB to form. Under our model knowing that a digenic interaction has occurred does not give any information regarding whether a more complex incompatibility has occurred as well. We assume that somehow it would be possible to verify that it is the interaction between the two genes and not the larger one that is causing speciation or vice versa.

Given this assumption of independence and our previous concerns of overestimating $CC_n$, we also wish to be able to differentiate between the likelihood of BDMIs of different complexities. That is, the probability that $n$ interacting loci cause a BDMI may change with $n$, so we define $p_n$ as the probability that any $n-$connected component causes a BDMI (again, taking into account the probability that the interaction is a false positive in the PPI network). After $K$ mutations, there is at most a $K-$connected component in $\mathcal{M}(K)$. The general probability of speciation can therefore be expressed as:

$$S(K) = 1 - \prod_{n=2}^{K} \varphi_n(K) \approx 1 - \prod_{n=2}^{K} (1 - p_n)^{E[CC_n(K)]} \tag{2}$$

or the probability that at least one BDMI of any complexity occurs. This means that under our model a less complex incompatibility may ”mask” complex incompatibilities involving it or the proteins that make it up. That is, if both an edge and triple containing this edge cause a BDMI according to the model, biologically the two would be indistinguishable from the edge alone causing a BDMI. This may make empirical validation of the model difficult when multiple BDMIs are predicted, but allows for much more straightforward numerical results to be obtained.

**Other Measures:** While knowing the probability of speciation is important in and of itself, for our analysis we also wish to know the relative importance of different complexities of interactions. There are several ways to do this. First, we can look at the probability that at least one $n-$sized BDMI occurs, $Q_n(K) = 1 - \varphi_n(K)$. Alternatively, we can also look at the ratio of the probability that *only* an $n-$sized interaction causes a BDMI over the total probability of speciation. Calling this quantity $R_n(K)$, we can define it as:

$$R_n(K) = \frac{Q_n(K)\prod_{i=2}^{K}\varphi_i(K)}{S(K)} \text{ where } i \neq n$$

With some simple algebra, we can show that:

$$R_n(K) = \frac{Q_n(K)}{S(K)}\frac{1 - S(K)}{1 - Q_n(K)}$$

In terms of analysis, $R_n(K)$ shows how much of the probability of speciation comes from $n-$sized incompatibilities alone. In other words, by looking at $R_n(K)$ we can see when and if incompatibilities of certain complexities make up more of the probability of speciation than others. A third way of interpreting $R_n(K)$ is as a conditional probability. Given that two populations have at least one BDMI isolating them reproductively after $K$ mutations, $R_n(K)$ is the probability that all existent BDMIs are of complexity $n$.

Much of the probability of speciation, at least later in divergence, would likely be dominated by the probabilities that incompatibilities of several different sizes have occurred. If we wish to exclude such probabilities and claim that after $K$ mutations we expect at most a single type of BDMI to have occurred, we can look at this probability in isolation. Calling it $T_n(K)$, we know that:

$$T_n(K) = Q_n(K)\prod_{i=2}^{K}\varphi_i(K) \text{ where } i \neq n$$

Verbally, $T_n(K)$ is the probability that at least one BDMI of size $n$ occurs, and no other BDMIs have happened. Letting $T(K) = \sum_{n=2}^{K} T_n(K)$, we can treat $T(K)$ as a modified probability of speciation excluding the possibility of multiple BDMIs of different sizes, and look at the ratio of $T_n(K)$ to $T(K)$ as another measure of the relative importance of different complexities of BDMIs.

Finally, we may also want to know the expected number of incompatibilities, both cumulatively and for BDMIs of different complexities. Again, as we assume independence, this quantity does not represent the number of interactions between $n$ proteins that are not functional, but the number of such interactions that are causing reproductive isolation independently of the interactions that make them up. Let $I(K)$ be the total number of BDMIs after $K$ substitutions, and let $I_n(K)$ be the number of BDMIs consisting of $n$ loci.

Since each BDMI is independent of others, $I_n(K)$ follows the binomial distribution and so:

$$E[I_n(K)] \approx E[CC_n]p_n \tag{3}$$

$$E[I(K)] \approx \sum_{n=2}^{K} E[CC_n]p_n$$

Unfortunately, no general solution for the average number of substitutions until reproductive isolation occurs could be found, however, some numerical results could be obtained and are presented.

## Numerical Methods

The model as structured above does not lend itself well to pure analysis, as $CC_n(K)$ is difficult to treat analytically and perhaps even more difficult to obtain data for, while $\alpha_2$ and $p_n$ are biological variables that are similarly hard to estimate, so we turn to numeric approximation over several ranges of variables.

Ideally, we would wish to find the $E[CC_n(K)]$ for biological networks given their precise structure, however, looking at complex interactions mathematically is particularly hard when estimating $E[CC_n(K)]$. This is difficult because even though we may know the structure of a PPI network, this does not easily transfer to information about the network of substitutions. As an example, think of a small-world network such as a network of all airports in the US. If we pick airports at random from a list of all the airports in the United States, we are quite unlikely to pick airports with connecting flights, as the number of airports with many connecting flights (hubs) is very small compared to the total number of airports. However if we keep repicking $K$ airports we will once in a while pick a hub as well. Thus the average network picked at random from $\mathcal{P}$ may be quite different in its structure. In other words, even if we know a species' PPI-network accurately, we would have to know the expected degree sequence of the random graph $\mathcal{M}(K)$ to obtain biologically relevant results, and such data are quite difficult to obtain. Furthermore, even if we assume that $\mathcal{M}(K)$ replicates the degree distribution of $\mathcal{P}$, there are no simple ways to approximate the number of connected components for scale-free networks or other types of networks thought to occur most frequently in biology. In fact, a similar problem of finding the largest connected component in a random graph is NP-complete. Lastly, the methods that do exist for estimating the size of the largest connected component in power-law graphs look at very large networks (Aiello *et al.* (2000), Chung and Lu (2002)) whereas we are interested in connected components ranging in size from 2 to $K$, where $K$ is not necessarily very large.

To get around these computational difficulties we return to our previous work (Livingstone *et al.* (2012)), where we showed that on average, the $K^{th}$ mutation will add $\alpha_2(K-1)$ new interactions. This leads us to estimate that there are on average $\alpha_2\binom{K}{2}$ interactions in $\mathcal{M}(K)$, so we know that at least the density of the networks remains the same on average. However, in our previous paper we do not discuss how these interactions are distributed and so predicting $CC_n(K)$ remains somewhat difficult. Simulation results demonstrated that on average, the degree of a node in $\mathcal{M}(K)$ followed a Beta-Binomial distribution (results not shown) which is unfortunately very cumbersome to work with. Approaching the problem directly does not seem possible.

Thankfully we do know that in the special case when $\mathcal{P}$ is a random network fitting the Erdős-Rényi model, such that all edges have a uniform probability of existing, the average $\mathcal{M}(K)$ is a network of the same type as well. A random network $G(N,p)$ following the Erdős-Rényimodel, only requires two parameters - the number of nodes $(N)$ and the density of edges $(p)$. As we already noted, the network of substitutions will have the same density as $\mathcal{P}$, and we also know that it will have $K$ nodes - that is the number of fixed substitutions. Therefore, we estimate $\mathcal{M}(K)$ as an Erdős-Rényi network with $K$ nodes and density $\alpha_2$. Additionally, such graphs have a binomial distribution of degrees, so while it is not a perfect fit for our biological estimate of $\mathcal{M}(K)$ which showed a Beta-Binomial distribution of edges, the two distributions are similar enough that it does present a good estimate at the very least. Finally, we also have exact expressions for the probability of an $n-$connected component in a typical random graph (Gilbert (1959)), given by:

$$P_n = 1 - \sum_{k=1}^{n-1} \binom{n-1}{k-1} P_k (1-\alpha_2)^{k(n-k)} \tag{4}$$

Here, $P_n$ is the probability of $n$ randomly selected nodes being connected and $\alpha_2$ is the density of the network as defined previously. This equation is quite useful to our model, as it allows us to estimate that the expected number of interactions after $K$ substitutions is $P_n\binom{K}{n}$. Furthermore, for large networks it is almost surely the case that $P_n\binom{N}{n} = E_n$, where $E_n$ is the number of $n-$connected components in $\mathcal{P}$, so $P_n = \frac{E_n}{\binom{N}{n}}$. In effect the probability of an n-connected component occurring in $\mathcal{M}(K)$, for large $K$, is the same as the density of such components in $\mathcal{P}$. Our previous research showed that $K$ does not need to be very large for the density of $\mathcal{M}(K)$ to approach $\alpha$, and so we assume it is fixed for all $K$. Therefore, given only the density of the PPI network, we can use Equation 4 to approximate the probability of a BDMI of any complexity occurring.

$$E[CC_n(K)] = \alpha_n \binom{K}{n} \tag{5}$$

Similarly to the overall density, $\alpha_n$ is both the density of $n-$connected components on the graph and the probability that $n$ random nodes are connected, so long as $K$ is large. We can now estimate the probability that an $n-$connected component does not cause speciation after $K$ fixed mutations in a random network, by plugging in the expected value of $CC_n(K)$, into the expression for $\varphi_n(K)$ giving:

$$\varphi_n(K) \approx (1-p_n)^{\alpha_n \binom{K}{n}}$$

Finally, we can express the probability of speciation as:

$$S(K) = 1 - \prod_{n=2}^{K} \varphi_n(K) \approx 1 - \prod_{n=2}^{K} (1-p_n)^{E[CC_n(K)]} = 1 - \prod_{n=2}^{K} (1-p_n)^{\alpha_n \binom{K}{n}} \tag{6}$$

By assuming that $\mathcal{M}(K)$ is a random network we now only have to worry about our choices for $\alpha_2$ and $p_n$, reducing the parameter space considerably. It can be quite difficult to find $CC_n$ in general, but by assuming random-network structure we only need to know $CC_2$ to estimate the rest, which is much easier.

### Several Issues

It is important to recall that random networks are quite distinct from the power-law networks that have generally been seen in biological organisms. Intuitively, if mutations occur uniformly throughout a scale-free or small-world network, as soon as a ”hub” mutates complex interactions will accumulate much faster than on a random graph, where we do not expect to see any such hubs. That is, for large $n$ and $K$, we might expect $CC_n(K) \gg \alpha_n \binom{K}{n}$ to hold. However, both for very large $n$, and early in differentiation this should not be the case. First of all, so long as $\alpha_2 > 0$, $\alpha_n \to 1$ as $n \to \infty$. This means that for very large $n$, all $n-$connected components occur even in very sparse networks. Second, hubs are less likely to have fixed substitutions both because there are less of them and because they may also be less likely to fix mutations due to heavier genetic constraints (Zotenko *et al.* (2008),Hahn and Kern (2005),Wuchty and Almaas (2005),Ramsay *et al.* (2009)). In effect, by the time that a hub has likely fixed a substitution, and therefore more large interactions may be seen in a small-world network than a random network, the number of $n-$connected components in the random network will be equal to or greater than the small-world network, for most $n$. In short, while error most certainly exists in our model, it is most likely a slight underestimate of the overall probability of speciation, and should not impact the relative contributions of BDMIs of different complexities.

A second problem with using Equation 4 to estimate $E[CC_n(K)]$ is that for small $\alpha_2$ on the order of the density seen in previous research ($10^{-3}$, Livingstone *et al.* (2012)), $P_n$ can be extremely small for large $n$, which makes numerical approximations more difficult. While we know that $\alpha_n = 1$ for very large $n$, since equation 4 is recursive, if $\alpha_n$ overflows the numeric type of whatever we are using to perform the recursive calculations, we cannot trust the predicted density of all components larger than $n$, as they are based at least partially on a false probability. To remove possible concerns of accuracy from the analysis of our results, we limit ourselves to the relatively small maximum complexity of $n = 10$. While larger values were evaluated assuming networks with higher density (results not shown), for this study we limit ourselves to $n = 10$ as it works for all three of the network densities on which we focused (Figure 1).

### Network Density Estimates

As it is not practical to evaluate all possible densities of networks, we focus on a range that seems both interesting theoretically and feasible biologically. As our lower limit, and what we will call the "scarce" network, we take the sharp threshold of the connectedness of an Erdős - Rényi graph $G(N, \alpha_2)$ which is defined as $\alpha_2 = \frac{ln(N)}{N}$. This is the density above which $G(N, \alpha_2)$ is almost surely connected and below which there are almost surely isolated vertices (Erdos and Renyi (1960)) In our previous research (Livingstone *et al.* (2012)), we used the *S. cerevisiae* PPI network data (Stark *et al.* (2006)) which has a density of .00872, and $N = 6018$. Keeping this $N$ and $\alpha_2 = .00872$ as biologically feasible quantities, we let the scarce network density be $\alpha_2 = \frac{\ln(6018)}{6018} \approx .00144$, and use $\alpha_2 = 2(.00872) = .01744$ as an upper bound, or a "dense" network. Thus, in our toy model of $\mathcal{P}$, each protein has on average 100 interactions in the dense, 52 in the biological and 8 interactions per protein in the scarce networks. Note that given that $\mathcal{M}(K)$ is a random network we know that $\mathcal{M}(K)$ is almost surely not connected until $K\alpha_2 = ln(K)$, and so the graph is almost surely disconnected until $K = 6018, K = 760, K = 333$ for the scarce, biological and dense networks respectively.

### Estimating Probabilities of BDMIs

To evaluate the probability of speciation, we must also make certain assumptions concerning the probability that a BDMI develops in an $n-$connected component, $p_n$. One assumption could be made that $p_n$ is fixed for all $n$. However, it does not seem that the average complex between 10 proteins, for example, would be as likely to interact in such a way as to cause a DBMI as the average digenic interaction. Furthermore, as we noted earlier, not all the complex interactions present on a PPI network would likely occur *in vivo*. Thus, a flat value of $p_n$ may be seen either as a claim that complex incompatibilities are inherently more likely or as a negation of the claim that the PPI network over-represents the number of complex interactions. We feel confident, however, that such scaling overestimates the probability of complex BDMIs.

We therefore consider several other ways that $p_n$ may scale, specifically as a decreasing function of $n$. A linear relationship would not be too interesting, as it would be equivalent to the fixed case aside from a relatively small scaling factor. Additionally, we would like $\varphi_2$ to remain fixed for all parameters so that we investigate the changes in the effects of complex BDMIs only. Numerical results were therefore obtained for the following cases:

- $p_n \propto 1$ for $n = 2, 3, \ldots$ as the fixed case.
- $p_n \propto \frac{1}{10^{n-2}}$ as one case where $p_n$ decreases quickly.
- $p_n \propto \frac{1}{(n-1)^{n-2}}$ as another case where $p_n$ decreases increasingly quickly.

For the biological network ($\alpha_2 = .00872$), we let the flat $p_2 = 10^{-4}$, while the scaling factors are multiplied by the flat probability. Solutions of $p_2$ such that $\varphi_2(K)$ remains roughly fixed were obtained, giving $p_2 = \frac{10^{-4} 0.00872}{\alpha_2}$ as the $p_2$ for any network with density $\alpha_2$. For the scarce and dense networks this results in $p_2 = 6.05510^{-4}$ and $p_2 = 5.01110^{-5}$ respectively. We used these parameters to verify that $\varphi_2(K)$ remained roughly invariable based on densities. The choice of $p_2 = 10^{-4}$ is relatively arbitrary. Higher $p_2$ caused speciation too rapidly to be biologically realistic, and made it difficult to observe the effects of complex interactions, as speciation was near certain before the number of such interactions became appreciable. Varying this parameter by several magnitudes lower resulted in results very similar in shape of $S(K)$, $R_n(K)$ and $Q_n(K)$ but stretched over longer time scales (results not shown). Since our primary goal is to see the influence of introducing PPI-network structure to the importance of complex interactions, and estimating the time until speciation is a secondary goal at most, this was sufficient reason to not investigate the effects of scaling $p_2$ further, at least for this study.

# Results and Discussion

Numerical results were obtained by calculating functions of interest over the chosen densities and probability scaling methods. Results were generally evaluated for $K = 2 \ldots 1000$ and $n \leq 10$ unless otherwise noted. Runs were performed in Scala, while figures were generated in R. Code for both is available upon request.

## The Total Number of Interactions of Different Complexities

The total number of incompatibilities increases exponentially for all complexities. As in other studies, the number of possible complex interactions quickly outnumbers the number of digenic interactions significantly (Figure 2). We see a clear snowball effect when looking at both the number of interactions, and the expected number of incompatibilities. However, the expected number of interactions of larger sizes is far smaller than would be expected under a complete network model. For instance, there would be $\binom{50}{10} \approx 10^{10}$ interactions in a complete network after 50 fixed substitutions, whereas even in our higher density network there are only a predicted $10^4$ such interactions under the same circumstances. The difference is not nearly as great for digenic interactions, being only several orders of magnitude apart. This of course ties back to the probability of such components under a random network model (Figure 1).

Our previous model (Livingstone *et al.* (2012)) presented the estimate that $K_S = \sqrt{\frac{\pi}{2\alpha p}}$ substitutions on average would lead to speciation. While we could not find a general solution when considering complex interactions, it is possible to find a numerical estimate. Following similar methodology, we let $K_S$ be the smallest $K$ such that $E[I(K)] = 1$, and try to find a numerical solution. While this is not exactly the same kind of estimate to the time until speciation as in our earlier paper, it makes sense to say that we expect least one BDMI should have occurred when the expected number of BDMIs is greater than or equal to 1.

Solving for $K_S$ we find:

$$\sum_{n=2}^{K} \alpha_n p_n \binom{K_S}{n} = 1$$

Evaluating numerically, we obtain the following results:

| $\alpha_2$ | .0174 | | | 0.00872 | | | .00144 | | |
|---|---|---|---|---|---|---|---|---|---|
| $p_n$ | flat | $\propto \frac{1}{n^{n-2}}$ | $\propto \frac{1}{10^{n-2}}$ | flat | $\propto \frac{1}{n^{n-2}}$ | $\propto \frac{1}{10^{n-2}}$ | flat | $\propto \frac{1}{n^{n-2}}$ | $\propto \frac{1}{10^{n-2}}$ |
| $K_s$ | 78 | 350 | 430 | 128 | 528 | 681 | 478 | 1118 | 1347 |

Note that this is after $p_n$ is taken into account, but that even before this is done (Figure 2) the structure of the PPI network alone has a strong impact on the number of potential BDMIs, while the combination of varying densities and $p_n$ produces diverse estimates.

## Probability of Speciation

Probability of speciation (Figure 3) is overall far higher if we include the possibility of complex interactions. This is unsurprising, as allowing for more possible cases of developing incompatibilities should give a higher probability of developing a BDMI. However, for sparse networks, and scaling $p_n$, the difference becomes relatively small. Surprisingly, the difference between networks of relatively similar density (.0174 and .0872) resulted in large differences when scaling of $p_n$ was considered (Figure 3). In fact, under 5 of the 9 parameter combinations, the probability of speciation is less than .5 even after 500 substitutions have been fixed. In contrast, if we assumed $\alpha_2 = 1$, $S(K) = 1$ after 40 fixed substitutions. Note that $Q_2$ is fixed for all the tested parameters (Figure 4), so these changes are entirely dependent on more complex interactions. Unfortunately, as there is little data to estimate just how $p_n$ should scale to accurately reflect biological processes and there is quite little data on BDMIs in nature, it is difficult to ascertain which rate is closest to nature. However, the case of relatively high network density and lack of scaling probability is clearly the least feasible - a BDMI is certain in less than 150 fixed substitutions. Likewise, flat scaling in general causes speciation far too rapidly to be biologically feasible, unless $p_n$ is presumed to be much smaller.

## Relative Importance of Digenic BDMIs

Unlike previous models of BDMIs, we find that throughout early divergence digenic BDMIs are the major proportion of the probability of speciation in scarce networks or when $p_n$ scales exponentially (Figures 5, 6 and 7). Furthermore, when networks are scarce and $p_n$ leads to unfavorable probability of complex BDMIs, the probability that only a digenic BDMI occurs remains high even as the probability of speciation becomes large (Figure 8). Thus, under certain circumstances, the combinatorial nature of complex interactions may not be sufficient to make them the major contributor to the probability of reproductive isolation. Of course, the results discussed here only take into account relatively small maximum sizes of BDMIs, however, given similar scaling for $p_n$ there is no reason to believe these results would not hold for higher values as well.

It is important to stress that digenic BDMIs are especially important early in the speciation process, no matter the parameters tested. Previous models agree that initially the number of possible digenic interactions is higher than that of more complex ones. However, our model extends the length of this importance. Unlike models that assume a complete network, it may take complex interactions a very long time to outnumber digenic ones, depending on the density of the network. Letting $\bar{K}_n$ be the number of fixed substitutions when there are as many $n-$connected components as digenic ones, we find that:

$$\alpha_2 \binom{\bar{K}}{2} = \alpha_n \binom{\bar{K}}{n}$$

This relationship immediately implies that $\bar{K} \propto \frac{\alpha_2}{\alpha_n}$ at the very least, and this ratio can be rather large for scarce networks.

We can therefore predict that while digenic interactions will eventually be outnumbered by more complex incompatibilities, speciation "early" in divergence is unlikely to be caused by complex BDMIs. Furthermore, what constitutes "early" depends very heavily on both the density of the network and the assumed probabilities of BDMIs. While anecdotal evidence for such a trend, a recent study in *S. cerevisiae* found that a digenic BDMI rapidly occurred under strong directive selection Anderson *et al.* (2010).

## Relative Importance of More Complex BDMIs

We find that incompatibilities of average complexity are less likely than either digenic, or more complex interactions throughout divergence. This is seen both in the expected number of such interactions (Figure 2) and in the relative contribution to the probability of speciation. This is because $\alpha_n$ is smaller for components of intermediate size than it is for either digenic interactions or very complex ones (Figure 1). At the same time, in the case of flat $p_n$ the number of complex interactions far outnumbers that of these medium sized ones, whereas scaling $p_n$ drastically increases the contribution that digenic interactions perform.

Since there is not much data for BDMIs in general, there is currently not enough empirical evidence to confirm or deny this trend, however it does present a testable hypothesis for speciation research.

## Relative Importance of Highly Complex BDMIs

Part of the goal of this research was to see how protein interaction network structure would affect the number and importance of complex incompatibilities. Many researchers have attempted to explain the relative frequency with which complex BDMIs are observed (Orr (1995), Welch (2004)). It remains difficult to ascertain the mechanisms by which highly complex BDMIs could cause speciation. Few authors argue that such incompatibilities are simply more likely than digenic ones. As noted earlier, it has been argued that there are simply more possible complex BDMIs (Orr (1995), Gavrilets (2003)). Furthermore, it has been speculated that the fraction of imaginable paths connecting a common ancestor to two incompatible genotypes that do not pass through inviable genotypes may be higher (Orr (1995),Welch (2004)). Furthermore, it has been shown that even at low frequency of viable genotypes, so long as genotypic space is large, there may be large connected components of viable genotypes with multiple highly complex incompatibilities between them (Gavrilets, 2004). In short, the general story has been that complex interactions are more numerous combinatorially, and are less constrained to evolve.

Our findings agree with most BDMI models in the prevalence of complex interactions. While the combinatorial nature of complex interactions certainly has an effect, we can also show that they are more likely due to some properties of networks.

There are three ways to justify the preponderance of complex interactions from a graph theoretical point of view: the sharp threshold of connectedness for random graphs, the probability that a component is connected, and the density necessary for a connected component to be possible.

While the equation is asymptotic, recall that a connected component a.s. exists in a random network if the density is greater than $\frac{\ln(n)}{n}$. If $\mathcal{M}(K)$ is a random-network, for large $n$ we can guarantee the connected component exists even if the density of the network is very low. That is, the density of the network does not need to be as high to guarantee complex connected components as it does to guarantee smaller ones. Even if it is not certain that the connected component exists, we know from Equation 4 that as $n \to \infty$, $\alpha_n \to 1$, if $\alpha > 0$. So, as seen in Figure 1, even if $\alpha_n$ decreases initially, we know that there will exist an inflection point such that more complex interactions will become more likely. Unlike the combinatorial results, which indicate that $\binom{K}{n}$ increases quite quickly when $n \leq \frac{K}{2}$, the probability results extend to all $n \leq K$.

Both of these results rely on a random-network structure, but we can argue intuitively why even on other graph structures complex interactions should be more likely. The minimum number of interactions necessary for an $n-$connected component to exist is $n-1$, while the possible number of interactions in such a component is $\binom{n}{2}$. If we think in terms of density, minimum density required for an $n-$connected component to exist is $\frac{2}{n}$. Of course, the probability that, given a random network, this density is sufficient to connect the components is fairly low, but it is non-zero nonetheless.

Thus there are inherent properties of networks that make interactions between many proteins more likely than interaction between several. There are therefore reasons other than the combinatorial nature of complex interactions for why complex BDMIs may occur more often.

1. It may be easier to reach two incompatible genotypes without going through an inviable genotype.
2. As the genotypes are diverging, complex interactions between many loci are more likely to exist.
3. There are simply more possible complex BDMIs.

Note that the interactions require the genotypes to be able to evolve, and that the number of possible complex interactions is bound by how many actual complex interactions there actually are, not by the possible number of such interactions, so all the explanations complement each other.

Since the number of possible complex BDMIs is limited by the interaction network of substitutions, it is worth noting under what cases we would expect very few complex BDMIs. As already noted, the number of $n-$connected components in $\mathcal{M}(K)$ can be no higher than the number of such components in $\mathcal{P}$. Thus, if a PPI network were found to have few complex interactions, we would also expect fewer complex BDMIs. Furthermore, if PPI-networks are truly similar to small-world/power law networks, no complex interactions can occur if there are no substitutions in the few proteins that are highly connected. There is some evidence that such substitutions may be less likely (Hahn and Kern (2005), Zotenko *et al.* (2008), Ramsay *et al.* (2009), Herbeck and Wall (2005)), however, it has also been argued that the predictive factor for low fixation rates is level of expression and not the number of interactions.

In short, the numerical results presented tend to agree with the assumption that complex interactions, and therefore complex BDMIs, highly outnumber simple ones as diverge continues. However, we present qualifications of under what kinds of parameters we expect to see this pattern, and when simple interactions, especially digenic, may remain the most important contributor to the development of RBIs.

# Conclusion

We have shown that it is possible to model the evolution complex BDMIs under the constraints of PPI networks. Most previous models of BDMI formation either heavily simplify or ignore the underlying architecture of these interactions.

Perhaps most surprisingly, our model leads us to strongly suspect that digenic BDMIs are likely to be much more important than previously thought. While under dense networks the contribution of digenic interactions to the probability of BDMIs is quickly overshadowed by more complex interactions, in scarce networks, and especially when the probability of complex BDMIs is much lower than that of digenic, we find that digenic interactions remain the major contributor to the probability of speciation, whether measured as the proportion of the probability that comes from such interactions alone, or as the probability that only such a BDMI occurs.

Furthermore, the network framework allows us to provide additional, intuitive explanations for why complex BDMIs may be found more frequently. While it is certainly the case that combinatorially such interactions are more numerous, we also find that they are more likely to occur on PPI networks, although the accuracy of such results is somewhat questionable. However, it is clear that not only are there more combinations of many loci, but that there are also more ways in which such loci may be connected, making highly complex interactions much more likely to exist on the PPI-network than digenic ones.

The model suffers mainly from lack of good parameter bounds. Not many PPI-networks have been investigated heavily, and too few BDMIs have been studied to allow for extrapolation on the magnitude of $p_n$, or the structure of $\mathcal{P}$. Furthermore, unless the precise selection regime is known, $\mathcal{M}(K)$ can only be modeled as a random network of some type. As evident even within our small parameter space, the results of the model are highly sensitive to both $\alpha_2$ and $p_n$. We cannot therefore present any predictions of what rates of BDMIs should be seen occurring in nature in general. However, so long as we have a good idea of $\alpha_2$, we may still provide at least some hypothetical boundaries on the fastest that a RBI would develop by assuming a flat $p_n$ of reasonable magnitude.

In the future the model can be extended to accommodate for a more complex view of speciation and more varied assumptions about the diverging populations (see Appendix III for some examples). While many such changes could be implemented, some of the most exciting prospects likely lie with modeling direction of selection via different probabilities of nodes mutating. Such modeling may even be able to account for the discrepancy seen in the literature concerning the snowballing strength of isolation (Gourbière and Mallet (2010)).

## Acknowledgments

I would like to thank my research advisers Dr. Kevin Livingstone and Dr. Peter Olofsson for their help and patience, and my lab-mates (in alphabetical order: Garner Cochran, Ali Elaydi, Karen MacPherson, Carter Payne and Kerry Seitz) for help in much of the foundational work for this research. Funding for this research was provided by Trinity University, HHMI, and NSF.

## Appendix

### I. Protein Networks as Graphs

This section serves as a short introduction to graph theory for biologists unfamiliar with the topic, and gives some important definitions. Graph theory deals with abstractions of graphs and networks, consisting of nodes (also called vertices) and edges between the nodes. Thus, a graph $\mathcal{G}$ is generally defined as a set of $\mathbf{V}$ and $\mathbf{E}$, which are respectively the sets of nodes and edges. Biologically, the nodes represent abstract loci, either genes or proteins, and an edge between the nodes represents an interaction, either physical or genetic (in the case of the dataset we use, we consider both types of interaction). Thus, the graph $\mathcal{P} = \{\mathbf{G}, \mathbf{I}\}$ is a protein interaction network with $|\mathbf{G}|$ proteins.

$\mathbf{E}$ can be a rather complicated set, as edges can have very many properties. However, in case of our model edges are tuples of the form $\{i, j\}$ s.t. $i, j \in \mathbf{V}$, indicating that there is an edge between vertices $i$ and $j$.

In this paper we only consider undirected graphs - that is, the edges $\{i, j\}$ and $\{j, i\}$ are equivalent.

This is because under our model, an incompatibility between $i$ and $j$ is directionless, and so we ignore the complications brought on by directed edges. Furthermore, in an undirected graph we know that the number of interactions is the magnitude of set of edges, $|\mathbf{E}|$, which makes looking at the number of possible interactions much simpler as we do not have to consider the possibility that several interactions may occur between two proteins. Some graphs may also contain multiple parallel edges, or edges from a node to itself, but these complications are also unnecessary for a model of Protein Interaction Networks when considering BDMIs.

Lastly, edges may have weights representing the strength of an interaction between nodes. We do not consider possible weights to the edges, and assume that interactions either exist or do not, and are not weighted by likelihood/frequency/strength of interaction. While adding this layer of complexity can be useful for some biological models that deal with graphs, in our case we are using the graph simply to ask whether an interaction between two loci exists.

An important topic in graph theory is the classification of different types of networks. The simplest and most direct type of graph is a complete network. A complete network is a network where all possible edges exist. Thus, the average path length between any two nodes is always 1, and any $n-$connected component exists. Most of the early work on graphs was done on random-networks, known as the ErdősRényi model. There are several ways to interpret a random network that turn out to be equivalent - it is a network where the probability of any edge existing is uniform, or it can be seen as a network with a fixed number of edges that are distributed between the nodes at random. Most of the results for this model are for the typical random network, which is the average random-network. More recently researchers have investigated more complex network types. One type is small-world networks, where the average distance from any two nodes scales as the logarithm of the total number of nodes. That is, the distance between nodes is relatively small even when the networks are large. When large, such networks tend to have a power-law distribution of degrees. That is, there are very few nodes of high degree, while most nodes only have several edges.

## II. Hybrid Fitness

Some speciation studies look at hybrid fitness (and the genetic load separating the two populations) rather than assuming that a BDMI will cause reproductive isolation instantly. Following Orr (1995), let $L(K)$ be the strength of reproductive isolation such that $L(K) = 1 - w(K)$, where $w(K)$ is the fitness of the hybrid after $K$ mutations have been fixed in either population. Since we are looking at the cumulative effect of all interactions between substitions, we assume that incompatibilities of different complexities affect the fitness independently, so let $w_n(K)$ be the fitness of the hybrid $K$ as affected by interactions between $n$ loci only, and let $w(K) = \prod_{n=2}^{K} w_n(K)$. Following Orr's model again, we assume that different incompatibilities of the same complexity also act independently to reduce fitness and that their effects are relatively small. Let $r_n$ be the average reduction in the fitness of the hybrid by an incompatibility between $n$ genes. After $K$ substitution are fixed, on average there will be a total of $E[CC_n(K)]$ interactions between $n$ substitions giving:

$$w_n(k) \approx (1 - r_n)^{E[CC_n(K)]}$$

Now looking at the cumulative effect over interaction between 2 to $K$, we get:

$$L(K) = 1 - w(K) \approx 1 - \prod_{n=2}^{K} (1 - r_n)^{E[CC_n]} \tag{7}$$

Assuming that $\mathcal{M}(K)$ is random-network, we can also expand the above to:

$$L(K) \approx 1 - \prod_{n=2}^{K} (1 - r_n)^{\alpha_n \binom{K}{n}}$$

Which is equation 6 with $r_n$ instead of $p_n$, so the two expressions are equivalent whenever $p_n = r_n$ for all $n$.

While some of the variables differ, the overall expression of the increase in hybrid load depends on roughly the same parameters. In fact, if $r_n = p_n$, the results are exactly the same, and so any findings for $S(K)$ are applicable to $L(K)$ as well. Note that the assumptions regarding independence may also be more fitting for such a model - it seems biologically sound to separate reduction in fitness by a digenic interaction independently of a trigenic one it is part of. However, the similarities, at least numerically, between the two allow us to focus on either only fitness or probability of speciation. We will therefore mainly look at the probability of speciation, with the assumption that the results for $S(K)$ can be translated to represent hybrid load, while $R_n(K)$ and $Q_n(K)$ can be seen as properties specific to the load from interactions between $n$ loci. Lastly, $I(K)$ and $I_n(K)$ can be seen as the reduction in hybrid fitness if the effects are additive as opposed to independent. While all of these interpretations are possible, none of the results will be presented under the lens of hybrid fitness.

## III. Extending the Model

While results in this paper have been given for some limited assumptions, the model framework is currently advanced enough to allow for the inclusion of more complex assumptions.

In its most abstract form, the model deals with some populations ($A, B, C, \ldots$, although generally constrained to $A$ and $B$), the protein interaction network of the ancestor $\mathcal{P}$, and the protein interaction network that would appear in a hybrid $\mathcal{M}(K)$. The probability of speciation depends on the structure of $\mathcal{M}(K)$ which in turn depends on both the structure of $\mathcal{P}$ and which substitutions $A$ and $B$ are fixing.

The process can therefore be condensed to two steps: picking a node in $\mathcal{P}$ to be added to $\mathcal{M}(K)$ and counting the number of $n-$connected components in $\mathcal{M}(K)$ to find the probability of a BDMI. In this paper we show that it is possible to gain some information about the probabilities of BDMIs by making certain assumptions about the structure of $\mathcal{P}$ and by assuming that loci are fixed at random. This works precisely because it gives us information about the structure of $\mathcal{M}(K)$. We could, however, simply assume $\mathcal{M}(K)$ takes some kind of structure without conditioning on $\mathcal{P}$ and in some ways this is what the original Orr model does by making the implicit assumption of $\mathcal{M}(K)$ being a complete network. More preferably, empirical studies of how the proteins in BDMIs interact, such as Burkart-Waco *et al.* (2012), will allow us to make predictions of the structure of $\mathcal{M}(K)$ directly. Nonetheless, a more general model would likely include both the selective forces on $A$ and $B$ as well as the resulting network that is produced from a specific $\mathcal{P}$. Thus, in the author's opinion, further models should primarily attempt to give a better account of the network between substituted loci.

In the first step (selection of substitutions), we have so far assumed that nodes are chosen uniformly from $\mathcal{P}$, and that no node could mutate twice.

**A. Selection:**

The model is actually fairly amiable to altering the assumption concerning random selection of nodes in $\mathcal{P}$. Let $X(Y|K_Y)$ be the next locus to fix a substitution in population $Y$ given the set $K_Y$ of the previously fixed substitutions in $Y$. In the current model, fixation of genes can be seen as proceeding under genetic drift, and so $X(A|K_A), X(B|K_B) \sim unif(N-K)$, where $N$ is the number of loci as defined previously and $K = |\mathbf{G}'(K)|$ is the magnitude of the set of loci in $\mathcal{M}(K)$, equivalent to $K_A + K_B$ if only one substitution in either population is allowed per locus. Here I present a case example of how selection could be incorporated instead (i.e $X(A|K_A) \nsim unif$), although further work is required to polish the methodology, and no numerical results are presented.

Suppose $A$ and $B$ are populations of equivalent fixed size, and each is undergoing equally strong directional selection, although not necessarily in the same direction. We will still make no assumptions as to how quickly new alleles are fixed. While the mathematics of the process may seem quite complex as a verbal model, considering network theory actually makes the results relatively painless, and the fact that there is directional selection actually tells us quite a lot about what $\mathcal{M}(K)$ should look like.

As the populations are undergoing equally strong selection, we can assume that $K_A = K_B = K/2$, and the next substitution is equally likely to be fixed in either population. Since we assume strong selection, we know that loci that are responsible for specific aspects of the phenotype will undergo selection. Intuitively, these genes should be more likely to interact than genes chosen at random, and a mutation on one locus should make it easier for the loci it interacts with to fix novel variants as well (see Palmer and Feldman (2009) for another model with similar assumptions). In other words $P(X(A|K_A) = n) \sim d_n(K_A)w(n)$, where $w(n)$ is the relative fitness of the substitution at $n$ and $d_n(K_A)$ is the degree of $n$ in the graph $(\mathcal{P} \setminus \{K_A \cup n\})^C$. Verbally, the probability locus $n$ fixes a substitution should be proportional to the number of interactions $n$ has with previously substituted nodes as well as the relative fitness of this substitution. Let $\lambda(K)$ be the fixation probability for all nodes $n \in \mathbf{G} \setminus \mathbf{G}'(K-1)$, given as $\frac{1}{2}(P(X(A|K_A) = n) + P(X(B|K_B) = n)$. It may seem impossible to gleam any information from this, however if selection acts on a sub-set of the network we already know quite a lot.

Consider loci $i$ and $j$, where $d_i(K_A) \gg d_j(K_A)$. A substitution at $i$ is then fixed with appropriately higher probability. Under such assumptions we would predict that $\mathcal{M}(K)$ would actually have higher density than under the assumptions of uniform probability presented in this paper. If selection is acting on a sub-network $\mathcal{A}$ with density $\alpha_{\mathcal{A}}$ in population $A$, and sub-network $\mathcal{B}$ with density $\alpha_{\mathcal{B}}$ in $B$, we would expect $\alpha_{\mathcal{M}(K)}$, the density of $\mathcal{M}(K)$, to simply be the average of the two densities. For proof, consider the average number of edges in $\mathcal{M}(K)$. We know that around $\alpha_{\mathcal{A}}\binom{K_A}{2} + \alpha_{\mathcal{B}}\binom{K_B}{2}$ edges exist in $\mathcal{M}(K)$, with further possible edges between loci in $\mathcal{A}$ and $\mathcal{B}$. Since $K_A, K_B \approx \frac{K}{2}$ (round to nearest integer for binomial):

$$\alpha_{\mathcal{M}(K)} \approx \frac{\alpha_{\mathcal{A}}\binom{\frac{K}{2}}{2} + \alpha_{\mathcal{B}}\binom{\frac{K}{2}}{2}}{\binom{K}{2}} \approx \frac{\alpha_{\mathcal{A}} + \alpha_{\mathcal{B}}}{2}$$

Thus, even if the selection process for the next substitution is complicated, so long as we know the densities of the sub-networks undergoing selection we can find the average number of interactions in $\mathcal{M}(K)$. Given what we know of PPI networks, it also makes intuitive sense to claim the density of sub-networks related functionally should be higher than the overall density. Note that for complex interactions we would still need to make some kind of stronger assumption about the structure of either $\mathcal{P}$ or $\mathcal{M}(K)$. Interestingly, assuming that $\mathcal{P}$ is an ErdősRényi graph would imply that $E[\alpha_A] = E[\alpha_B] = \alpha_2$ where $\alpha_2$ is the density of $\mathcal{P}$, so selection would have no effect.

Further assuming that the rates of fixation differ in the populations would only make the expression for $\alpha_{\mathcal{M}(K)}$ be weighted by the ratios $f = \frac{K_A}{K}$ and $(1-f) = \frac{K_B}{K}$. We can therefore vary what kind of selection either population is experiencing, as this would primarily change the fixation rates, giving:

$$\alpha_{\mathcal{M}(K)} \approx \frac{\alpha_A \binom{fK}{2} + \alpha_B \binom{(1-f)K}{2}}{\binom{K}{2}}$$

Once again, if $\mathcal{P}$ is an ErdősRényigraph, the density remains the same. However, given a small-world structure, it is likely that the densities of $\mathcal{A}$ and $\mathcal{B}$ are both higher than the average density. So, the extended model should also predict faster rates of speciation when populations are undergoing selection as opposed to drift, which seems to fit our intuition of how divergence occurs. Importantly we do not consider the *time* until the development of an RBI, only the number of fixed substitutions, which proves to be sufficient for faster divergence under certain circumstances.

Alternatively, or in complement with this addition, $p_n$ may be scaled by the strength of selection as the substitutions are having a stronger fitness effect. Incorporating selection also brings in stronger concerns on the assumption that all loci are at most diallelic. If selection is occurring on the same set of loci in both populations, could we not see multiple alleles become fixed subsequently in the same locus?

**B. Multiple alleles at each locus** The assumption that all loci are diallelic can be removed by allowing edges of the form $\{i, i\}$, and modifying our choice of nodes to reflect the possibility that the same protein fixes mutation multiple times. We can modify our previous definitions as follows. Given $\mathbf{G}'(K-1)$ and $\mathbf{X}(K-1)$ a mutation is fixed on node $i \in \mathbf{G}$. Then,

$$\mathbf{G}'(K) = \begin{cases} \mathbf{G}'(K-1) \cup \{i\} & \text{if } i \notin \mathbf{G}'(K-1) \\ \mathbf{G}'(K-1) & \text{else} \end{cases}$$

and

$$X(K) = \begin{cases} X(K-1) \cup \{\{i,j\} : \{i,j\} \in \mathbf{I}\} & \text{if } i \notin \mathbf{G}'(K-1) \\ X(K-1) \cup \{i,i\} & \text{else} \end{cases}$$

Modeled this way, $\mathcal{M}(K)$ has two interesting new properties. First, the magnitude of $\mathbf{G}'(K)$ is not necessarily the same as $K$ (so, even in a complete network, we have no guarantee that $CC_n(K) = \binom{K}{n}$. It may therefore be more appropriate to define $K$ not as the number of fixed substitutions, but as the number of loci with a substitution fixed in either population. Assume this is true and let $K'$ be the number of fixed substitutions. Second, unless the selection method favors alleles on previously unaltered loci, there will exist a point when each successive mutation will be less likely to hit a new node than one with a pre-existing mutation. Therefore, for some large $K$, $P(CC_2(K) - CC_2(K-1) \leq 1) \gg 0$. Since the probability of BDMIs directly depends on $CC_n$, which in turn depends on the density of $\mathcal{M}(K)$ defined by $\frac{CC_2}{\binom{K'+1}{2}}$, we will see a slow down in the increase of $S(K)$ unless the probability of at least one BDMI existing is already 1 by that point.

### C. Variable Fixation Rates:

Somewhat similar to the issue of selection is the influence of $\mathcal{P}$ on the probability of specific loci fixing mutations. Several studies have found a negative correlation between the degree of loci in protein-networks and the rate of their evolution, although the relationship may not be causative (Zotenko *et al.* (2008), Hahn and Kern (2005), Wuchty and Almaas (2005), Ramsay *et al.* (2009)). It may be preferable, therefore, to lower the probability of nodes of higher degree to fix substitutions. Intuitively, this would lower the density of $\mathcal{M}(K)$ no matter its structure. It is difficult, however, to either find a good estimate of the strength of this effect, or an analytical solution for its impacts in a biological PPI network.

We can, however, consider its effects in a toy model. Under the assumption that $\mathcal{P}$ is an ErdősRényirandom network, there are very few nodes with degrees much lower or higher than $N * \alpha_2$. On average, the probability that a randomly selected node $i$ is of degree $n$ is:

$$P(d_i = n) = \alpha^n (1-\alpha)^{N-n}$$

Now, define $f(n) = P(\text{node of degree } n \text{ fixes a substitution}) = \frac{1}{n+1}$ as an arbitrary scaling for the probability that a substitution is fixed when it occurs in a locus of degree $n$. Lastly, let mutations occur randomly on any locus with uniform probability. Since fixation events are independent of each other we can obtain the following relationships. After $M$ mutations, on average $M\alpha_2^n(1-\alpha_2)^{N-n}$ loci of degree $n$ that had a mutation. Each of these mutations is fixed with probability $\frac{1}{n+1}$, giving the expected number of loci of degree $n$ with fixed substitutions as:

$$E[d_n|M] \approx \frac{M\alpha_2^n(1-\alpha_2)^{N-n}}{n}$$

Now the number of nodes that have fixed substitutions after $M$ mutations, $K$, and the degree in $\mathcal{P}$ of loci in $\mathcal{M}(K)$ $(X_M)$ will be:

$$E[K|M] = \sum_{n=1}^{N} E[d_n|M] \approx \sum_{n=1}^{N} \left( \frac{M\alpha_2^n(1-\alpha_2)^{N-n}}{n+1} \right)$$

$$E[X_M] = \sum_{n=1}^{N} nE[d_n|M] \approx \sum_{n=1}^{N} M\alpha_2^n(1-\alpha_2)^{N-n}$$

Since this is still a random network, we know that while the total number of edges in $\mathcal{M}(K)$ will be lower than $E[X_M]$, the average degree in $\mathcal{M}(K)$ is the same as the average density in $\mathcal{P}$ of the loci in $\mathcal{M}(K)$, i.e.

$$\alpha_{\mathcal{M}(K)} = \frac{E[X_M]}{\binom{N}{2}}$$

. Thus, the model set-up, by itself, is not terribly difficult. Unfortunately, here the roadblock is precision: even for modest $N$ we will soon find that most numeric simulations will not be able to handle the tiny probabilities. Future work with arbitrary precision numerical simulations my provide more data here.

## IV. Compatibility with other models

As proof of concept as to how known models could be extended to account for PPI structure using our model, we present several cases of models in the literature that may be amiable to including PPI-network structure. The models chosen here do not produce novel results, but show that including network structure does not change the impact of many other variables considered in the literature.

### A. Palmer and Feldman 2009:

When considering divergence, Palmer and Feldman (2009) note that Orr's results rely on $K$ increasing continuously. However, fitness is only considered to depend on it since $K$ is used as a proxy of the divergence between the two populations - and this may decrease under certain circumstances. Instead, they propose letting $D$ be the number of diverged loci between the populations. Essentially, $D$ allows the model to account for both introgression of the alleles if the populations are not allopatric or parallel evolution. This is a relatively simple change roughly equivalent to removing nodes from $\mathcal{M}(K)$ under certain circumstances. Unless these nodes are highly specific, we can act on the assumption that $D \approx K$ and so we would predict many of the same patterns in the development of RIBs.

Additionally, Palmer and Feldman look at the effects of a finite number of interacting loci, thus a comparison with our model might be very interesting. They define the incompatibility between the two populations analogously to the genetic load equation in Orr (1995), so we can use the results of Appendix II to show how PPI networks play a role in such models. Furthermore, as shown in Appendix III, we can incorporate selective pressures, allowing us to generate a model on many of the same assumptions. However, as few analytical results are presented in this model it is difficult to say exactly how our results would be different. Palmer and Feldman make many assumptions about both the selective regime in the populations and the fitness impact of each substitution, as well as how it enables further substitutions. Unfortunately, as they do not present analytical results we cannot easily extend their model with ours.

**B. Welch 2004:**

A slightly more interesting results is considered by Welch (2004), when looking at the number of genes introgressing in the hybrid, $m$, in order to account for asymmetric BDMIs. In this model the expected number of incompatibilities involving $n$ loci is given as $E[I_n] = p_n E[C(K,n)]$, where $p_n$ is the probability of an incompatibility between $n$ loci, similar to our model and $E[C(K,n)]$ is the expected number of unproven combinations of $n$ loci, given by $E[C(K,n)] = \binom{K}{n}(2^n - n - 1)$. This is derived as the product of the total number of possible combinations of $K$ loci in groups of $n$ and the number of combinations of the $n$ loci ($2^n$) minus the number that have to have been for the $K$ loci to have obtained substitutions ($n+1$: the $n$ loci and the ancestral state). Welch goes on to also look at how this number is affected by the number of introgressing genes. While, we have not tried to perform a similar step in our model, but it is easily evident that the results would mainly change the number of unproven combinations while the remainder of the model would simply incorporate this modified value. Assuming that $\mathcal{P}$ is a random network, we know at the very least that the number of interactions between $n$ loci is approximately $\alpha_n \binom{K}{n}$. Incorporating interactions into the number of untested loci is a bit more difficult: we know that on average $\alpha_2$ of the loci actually interact, so that there are on average $\alpha_2 n + 1$ tested combinations of the loci, out of a total possible number of $2^{\alpha_2 n}$. Thus, given PPI-network structure, we let

$$E[C(K,n)] = \alpha_n \binom{K}{n} (2^{\alpha_2 n} - \alpha_2 n - 1)$$

.

Network density is expected to play a similar result when considering the introgression of $m$ loci out of the $K$ loci with fixed substitutions. The central quantity Welch considers here is the expected number of combinations of $n$ amino acids that may be formed by introgressing $m$ loci and counting only combinations where all $m$ loci introgress, denoted by $E[C(K,n,m)]$, where

$$E[C(K,n,m)] = \frac{\binom{K}{n}}{\binom{K}{m}} \left[ \binom{n}{m} - 1 \right]$$

Given our previous reasoning,, since $\binom{K}{n}$ is a stand in for the number of interactions between $n$ loci, it seems like all that would be necessary for this portion of the results is to scale $E[C(K,n,m)]$ by $\alpha_n$. The expected number of BDMIs formed by $m$ introgressing genes is then

$$E[I_{n,m}] = p_n \alpha_n E[C(K,n,m)] = \frac{E[I_n(K)]}{\binom{K}{m}} \left[ \binom{n}{m} - 1 \right]$$

, $E[I_n]$ as defined in Model and Methods. Welch lets $P_m$ be the probability that a BDMI of any size occurs with $m$ genes introgressing and assumes $E[I_{n,m}] \ll 1$. Doing the same, we find:

$$P_m \approx 1 - \prod_{n=m}^{K} 1 - E[I_{n,m}] = 1 - \prod_{n=m}^{K} \left( 1 - \frac{E[I_n]}{\binom{K}{m}} \left[ \binom{n}{m} - 1 \right] \right)$$

This leads us to many of the same conclusions: divergence still snowballs even if it does so considerably slower, but the model can still account for the findings that DMIs should be asymmetrical.

### C. Orr and Orr 1996:

Orr and Orr (1996) consider more than two populations in the process of speciation. Given $b$ allopatric populations of total fixed size $N$, such that each population is of size $\frac{N}{b}$. Two metrics are considered: expected number of substitutions until the first of the $\binom{b}{2}$ pairs of populations produces hybrids with BDMIs, or the time until the average of these pairs is reproductively isolated via a BDMI. We will only look at the latter. Unlike the general model, we need to distinguish between incompatibilities between two substituted loci (called $DD$ incompatibilities) and an incompatibility between a substituted locus in one population and an ancestral locus it has not encountered (called $DA$ incompatibilities). Orr and Orr show that the expected number of each is the same and for large $K$ is approximately $\frac{p_2 K^2}{b^2}$. Therefore, the total number of expected incompatibilities is twice that. Accounting for the expected number of interactions given a PPI-network structure, we instead obtain the result:

$$I(K) = I_{DD} + I_{DA} \approx \frac{4 p_2 \alpha_2 \binom{K}{2}}{b^2} = \frac{2 p_2 \alpha_2 K(K-1)}{b^2}$$

### D. Snowball Theory:

A recent concern in the study of speciation has been the validation of the snowball effect, with studies finding no snowballing relationship between divergence and magnitude of hybrid sterility or inviability. In fact Gourbière and Mallet (2010) find that a model that assumes slowing divergence rather than snowballing fits the data much better. If we modify our model as in Appendix III.B, slowing in the rate of divergence becomes a consequence of fewer novel interactions, thus it may provide some theoretical backing for the possibility of slowing divergence. Presgraves (2010) has responded that the magnitude of inviability/sterility is not equivalent to Orr's prediction that the number of incompatibilities should snowball, and cites Moyle and Nakazato (2010), Matute *et al.* (2010) as cases where the snowball effect is confirmed when looking at the number of incompatibilities rather than the strength of isolation. While incorporating multiple substitutions at a single locus in our model would lead to an slowdown in the expected number of incompatibilities eventually, at least early in divergence, as is the case with both of the cited studies, the snowball effect should still be evident. It is worth noting that it could be argued that each successive mutation is more diverged and therefore more likely to cause speciation. While mathematical treatment of this may be difficult, it may counter the slowdown seen in the raw number of possible BDMIs. The conditions under which slow down rather than snowballing would be observed need to be better defined if our model is to be tested for accuracy in this realm as well.

# V. Figures

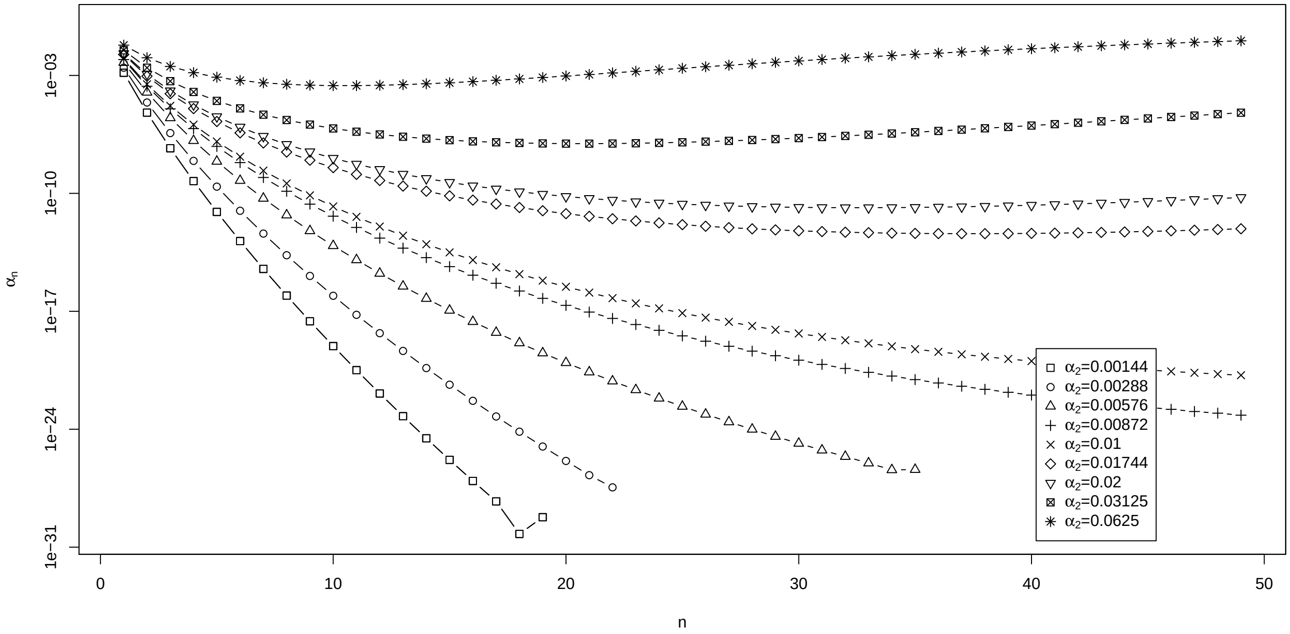

Figure 1: The probability of an $n-$ connected component in a random graph with average density given by $\alpha_2$. For scarce graphs, our numerical implementation of Equation 4 is not able to compute probabilities of components larger than 20. This is one of the reasons why we limit our numerical results to at most BDMIs between 10 proteins.

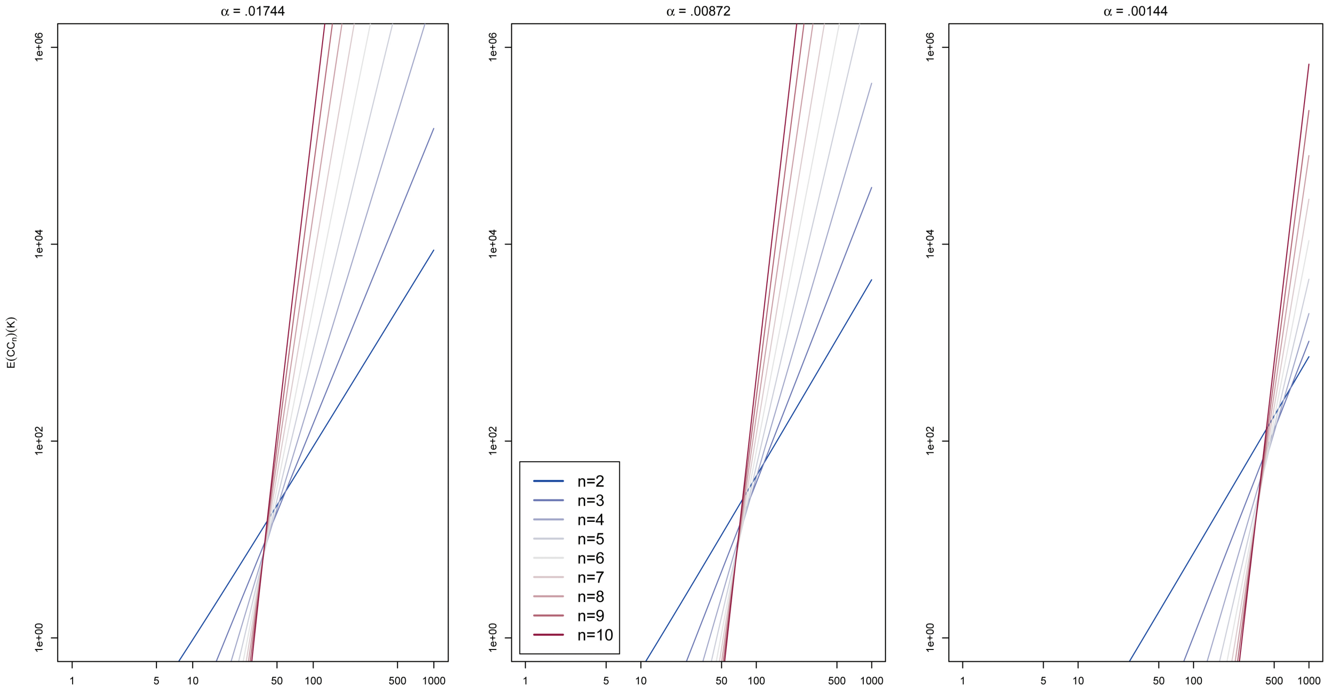

Figure 2: The expected number of connected components of sizes $n = 2 \ldots 10$, for the three densities used to obtain the remaining results. Gilbert's equation does not counter the impact of the combinatorial nature of complex interactions - for all $n > 2$, once $CC_n > CC_2$, the number of $n-$connected components rapidly outpaces the number of digenic ones.

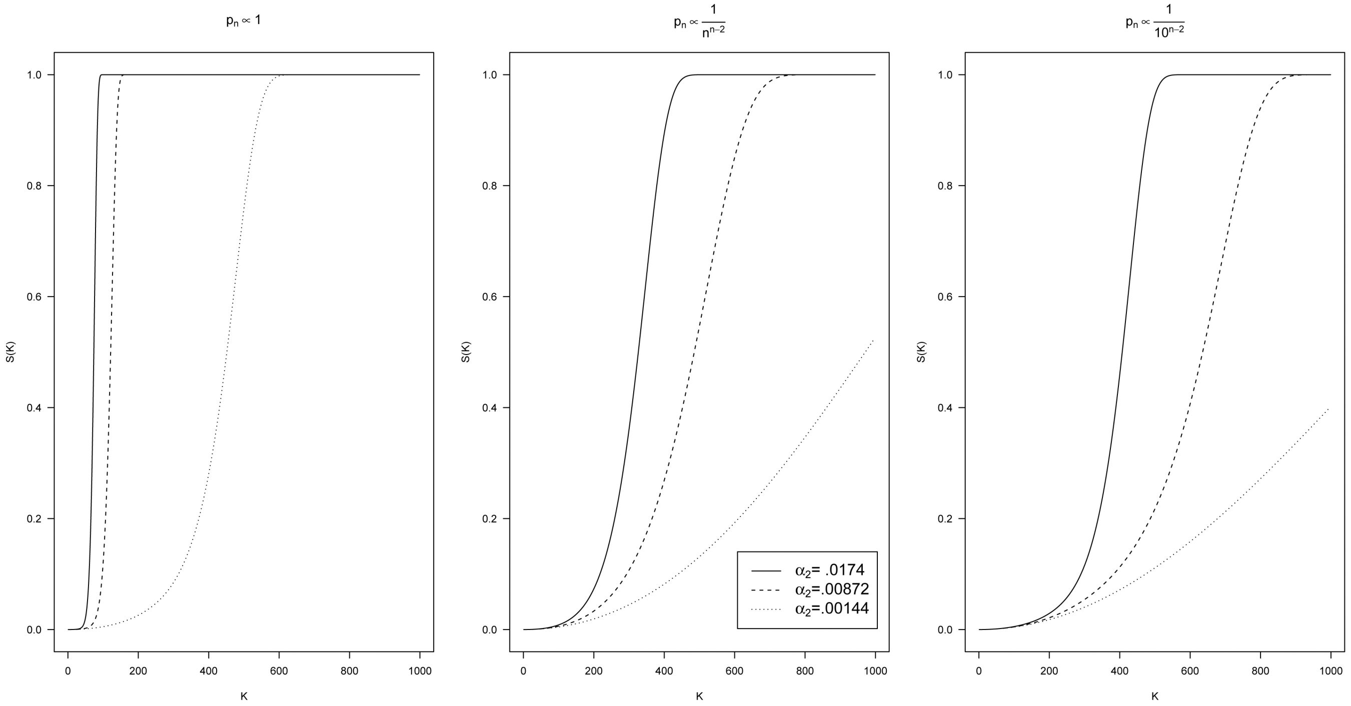

Figure 3: The probability of a BDMI between $n = 2 \ldots 10$ loci after $K$ mutations. As in our previous model Livingstone *et al.* (2012), the density of the network has a strong effect on the probabilities of speciation. This effect is stronger when considering scaling probabilities of BDMIs, giving rise to a wide range of speciation rates based on 2 parameters.

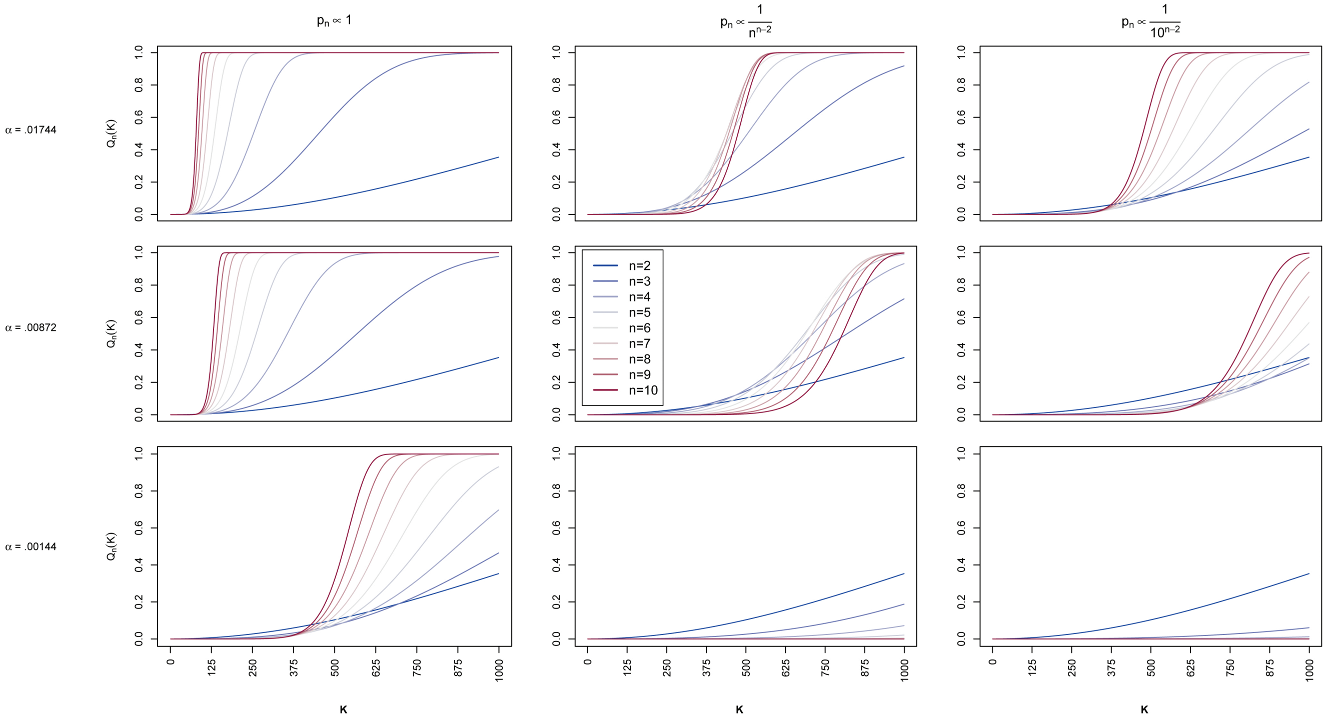

Figure 4: The probability that an interaction between $n$ genes causes a BDMI, for $n = 2 \ldots 10$. Both the scaling of probability of a BDMI, and the average density of the network have a strong impact on the shape of these curves. Note that $Q_2$ is fixed for all parameter choices, allowing for comparisons on the effect of structure and scaling on the development of complex incompatibilities.

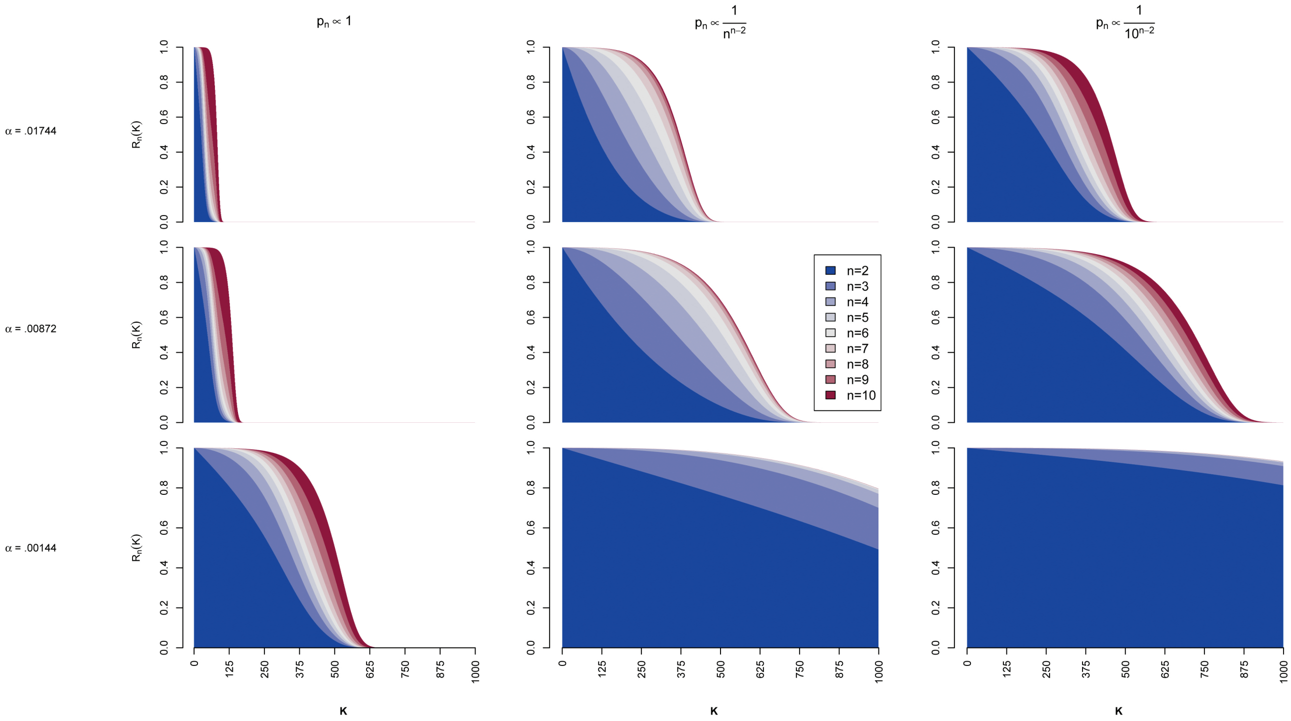

Figure 5: Stacked graphs of the proportion of the probability of speciation that comes from only $n-$connected components occurring. Can be seen as the probability that only an $n-$connected component

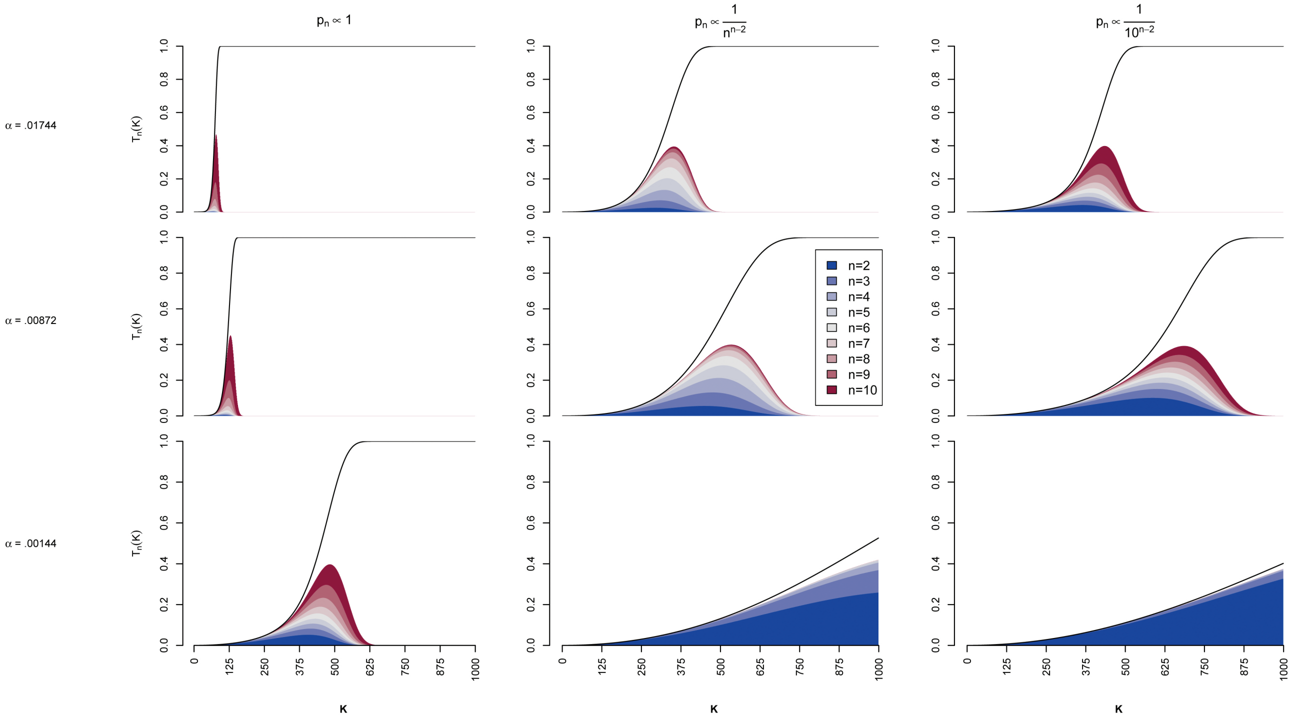

Figure 6: Stacked graphs of the probability that only an $n-$connected component causes a BDMI.

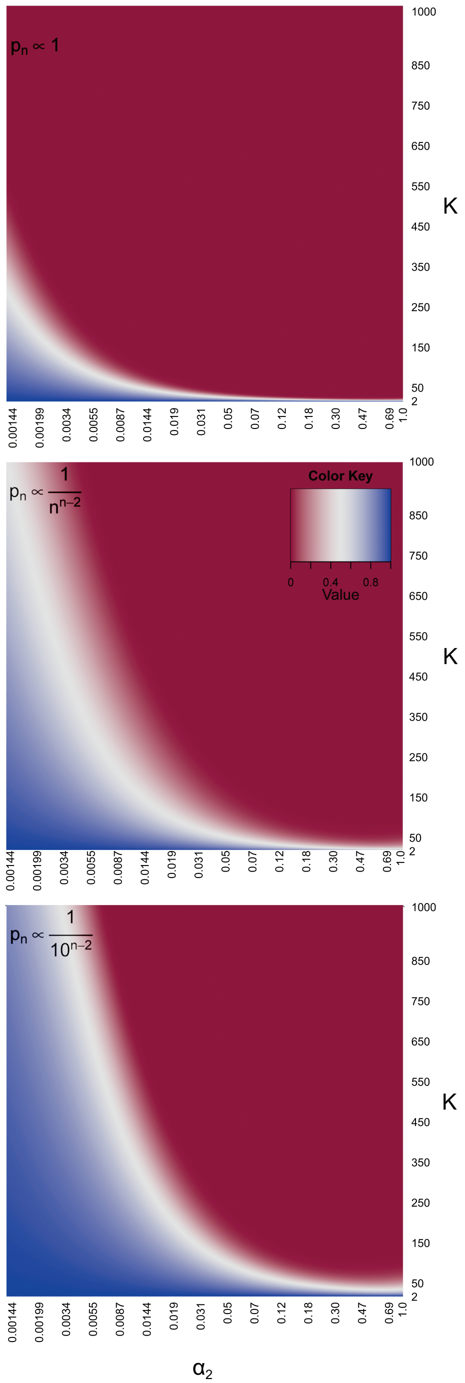

Figure 7: The proportion of the probability of speciation that is only digenic interactions for a variety of network densities. While digenic interactions become unimportant quite quickly in dense networks, in scarcer networks their importance may linger for a while.

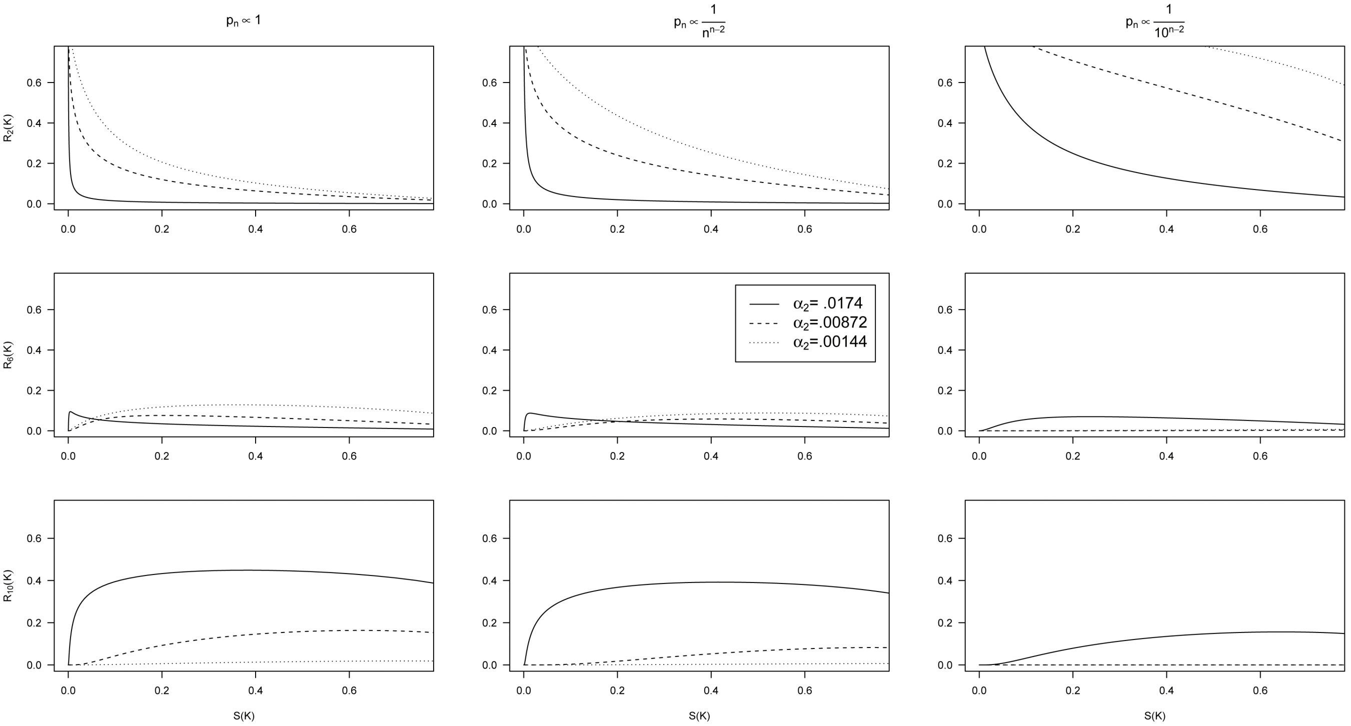

Figure 8: The relative proportion of the probability of speciation different sized complexes contribute with increasing probability of speciation.

‘

## References

Aiello, W., F. Chung, and L. Lu, 2000 A random graph model for massive graphs. Proceedings of the 32nd Annual ACM Symposium on Theory of Computing : 171–180.

Anderson, J. B., J. Funt, D. A. Thompson, S. Prabhu, A. Socha, *et al.*, 2010 Determinants of divergent adaptation and Dobzhansky-Muller interaction in experimental yeast populations. Current Biology **20**: 1383–1388.

Bansal, S., S. Khandelwal, and L. A. Meyers, 2009 Exploring biological network structure with clustered random networks. BMC Bioinformatics **10**: 405.

Bateson, W., 1909 Heredity and variation in modern lights, 85–101. In *Darwin and Modern Science*, editor, A. C. Seward. Cambridge University Press.

Bikard, D., D. Patel, C. Le Metté, V. Giorgi, C. Camilleri, *et al.*, 2009 Divergent evolution of duplicate genes leads to genetic incompatibilities within *A. thaliana.* Science **323**: 623–6.

Burkart-Waco, D., C. Josefsson, B. Dilkes, N. Kozloff, O. Torjek, *et al.*, 2012 Hybrid incompatibility in arabidopsis is determined by a multiple-locus genetic network. Plant Physiology **158**: 801–812.

Cabot, E. L., A. W. Davis, N. A. Johnson, and C.-I. Wu, 1994 Genetics of reproductive isolation in the *Drosophila simulans* clade: complex epistasis underlying hybrid male sterility. Genetics **137**: 175–189.

Christie, P., and M. R. Macnair, 1987 The distribution of postmating reproductive isolating genes in populations of the yellow monkey flower, mimulus guttatus. Evolution : 571–578.

Chung, F., and L. Lu, 2002 The average distances in random graphs with given expected degrees. Proceedings of the National Academy of Sciences **99**: 15879–15882.

Coyne, J., and H. Orr, 1989 Patterns of speciation in Drosophila. Evolution **43**: 362–381.

Coyne, J. a., and H. a. Orr, 1998 The evolutionary genetics of speciation. Philosophical transactions of the Royal Society of London. Series B, Biological sciences **353**: 287–305.

De Queiroz, K., 2007 Species concepts and species delimitation. Systematic Biology **56**: 879–886.

Dobzhansky, T., 1937 Genetic nature of species differences. American Naturalist **71**: 404–420.

Erdos, P., and A. Renyi, 1960 On the evolution of random graphs. Publications of the Mathematical Institute of the Hungarian Academy of Sciences **5**: 17–61.

Gavrilets, S., 2003 Models of speciation: what have we learned in 40 years? Evolution **57**: 2197–2215.

Gavrilets, S., 2004 *Fitness landscapes and the origin of species (MPB-41)*. Princeton University Press.

Gilbert, E., 1959 Random graphs. The Annals of Mathematical Statistics **30**: 1141–1144.

Gourbière, S., and J. Mallet, 2010 Are species real? The shape of the species boundary with exponential failure, reinforcement, and the "missing snowball". Evolution **64**: 1–24.

Hahn, M., and A. Kern, 2005 Comparative genomics of centrality and essentiality in three eukaryotic protein-interaction networks. Molecular Biology and Evolution **22**: 7–10.

Herbeck, J., and D. Wall, 2005 Converging on a general model of protein evolution. TRENDS in Biotechnology **23**: 485–487.

Jordan, I. K., Y. I. Wolf, and E. V. Koonin, 2003 No simple dependence between protein evolution rate and the number of protein-protein interactions: only the most prolific interactors tend to evolve slowly. BMC evolutionary biology **3**: 1.

Kao, K. C., K. Schwartz, and G. Sherlock, 2010 A genome-wide analysis reveals no nuclear Dobzhansky-Muller pairs of determinants of speciation between *S. cerevisiae* and *S. paradoxus*, but suggests more complex incompatibilities. PLoS Genetics **6**: e1001038.

Kirkpatrick, M., and V. Ravigné, 2002 Speciation by natural and sexual selection: models and experiments. The American Naturalist **159**.

Kondrashov, A., 2002 Dobzhansky - Muller incompatibilities in protein evolution. Proceedings of the National Academy of Sciences of the United States of America **99**: 14878–14883.

Livingstone, K., P. Olofsson, G. Cochran, A. Dagilis, K. MacPherson, *et al.*, 2012 A Stochastic model for the development of Bateson-Dobzhansky-Muller incompatibilities that incorporates protein interaction networks. Mathematical Biosciences : 49–53.

Matute, D. R., I. a. Butler, D. a. Turissini, and J. a. Coyne, 2010 A test of the snowball theory for the rate of evolution of hybrid incompatibilities. Science **329**: 1518–21.

Mayr, E., 1942 *Systematics and the origin of species, from the viewpoint of a zoologist*. Harvard University Press.

Moyle, L. C., and T. Nakazato, 2010 Hybrid incompatibility ”snowballs” between *Solanum* species. Science **329**: 1521–3.

Muller, H. J., 1942 Isolating mechanisms, evolution and temperature. Biological Symposia **6**: 71–125.

Orr, H., 1995 The population genetics of speciation: the evolution of hybrid incompatibilities. Genetics **139**: 1805–1813.

Orr, H. A., and L. H. Orr, 1996 Waiting for speciation: The effect of population subdivision on the time to speciation. Evolution **50**: pp. 1742–1749.

Orr, H. a., and M. Turelli, 2001 The evolution of postzygotic isolation: accumulating Dobzhansky-Muller incompatibilities. Evolution **55**: 1085–94.

Palmer, M. E., and M. W. Feldman, 2009 Dynamics of hybrid incompatibility in gene networks in a constant environment. Evolution **63**: 418–31.

Presgraves, D., and W. Stephan, 2007 Pervasive adaptive evolution among interactors of the Drosophila hybrid inviability gene, Nup96. Molecular Biology and Evolution **24**: 306–314.

Presgraves, D. C., 2003 A fine-scale genetic analysis of hybrid incompatibilities in drosophila. Genetics **163**: 955–972.

Presgraves, D. C., 2010 The molecular evolutionary basis of species formation. Nature Reviews Genetics **11**: 175–80.

Ramsay, H., L. H. Rieseberg, and K. Ritland, 2009 The correlation of evolutionary rate with pathway position in plant terpenoid biosynthesis. Molecular Biology and Evolution **26**: 1045–1053.

Stark, C., B.-J. Breitkreutz, T. Reguly, L. Boucher, A. Breitkreutz, *et al.*, 2006 BioGRID: a general repository for interaction datasets. Nucleic Acids Research **34**: D535–9.

Turelli, M., N. H. Barton, and J. A. Coyne, 2001 Theory and speciation. Trends in Ecology & Evolution **16**: 330–343.

Welch, J., 2004 Accumulating Dobzhansky-Muller incompatibilities: reconciling theory and data. Evolution **58**: 1145–1156.

Wuchty, S., and E. Almaas, 2005 Peeling the yeast protein network. Proteomics **5**: 444–9.

Zotenko, E., J. Mestre, D. P. O'Leary, and T. M. Przytycka, 2008 Why do hubs in the yeast protein interaction network tend to be essential: reexamining the connection between the network topology and essentiality. PLoS Computational Biology **4**: e1000140.