Reproductive Barriers as a Byproduct of Gene Network Evolution

Chia-Hung Yang[1] **and Samuel V. Scarpino**[1,2,3,4,5*]

***For correspondence:**
s.scarpino@northeastern.edu

[1]Network Science Institute, Northeastern University, Boston, United States; [2]Department of Marine and Environmental Sciences, Northeastern University, Boston, United States; [3]Department of Physics, Northeastern University, Boston, United States; [4]Department of Health Science, Northeastern University, Boston, United States; [5]ISI Foundation, Turin, Italy

Abstract Speciation in the absence of divergent selection remains a topic of active debate in evolutionary biology. Existing empirical and theoretical studies have linked the process of speciation to complex genetic interactions. Gene Regulatory Networks (GRNs) capture the inter-dependencies of gene expression and encode information for individual development on a molecular level, which form a feedback loop to learn both patterns and effects of hybrid incompatibilities. Here, we develop a pathway framework considers GRNs as a functional representation of coding sequences. We then simulated the dynamics of GRNs through a simple model integrating natural selection, genetic drift and sexual reproduction and uncovered reproductive barriers among allopatric population subjected to identical selection pressure. A minimal mechanism of how reproductive isolation emerged was identified by numerical counter-factual analyses. We discuss how many features of our results are able to account for observed empirical patterns, which are currently in opposition to classical models of speciation. This study adds support for the central role of gene networks in speciation and their potential to shed light on as yet largely unexplained patterns in evolution.

Introduction

Over the past 100 years, the role of reproductive isolation due to genetic differences between populations has received considerable attention in both the empirical and theoretical literature on speciation (***Rieseberg et al., 1996***; ***Coyne and Allen Orr, 1998***; ***Marques et al., 2019***; ***Satokangas et al., 2020***). Through this work, it is widely accepted that divergent selection between geographically isolated populations can facilitate speciation due to the accumulation of genetic incompatibilities (***Bateson, 1909***; ***Dobzhansky, 1936***; ***Muller, 1942***). Despite well-established examples from *Drosophila* (***Brideau et al., 2006***), *Xiphophorus* (***Wittbrodt et al., 1989***; ***Powell et al., 2020***), *Oryza* (***Yamamoto et al., 2010***), *Arabidopsis* (***Bikard et al., 2009***), and *Mus* (***Davies et al., 2016***), the genetics and evolutionary history of incompatibilities are typically far more complex than suggested by early models (***Marques et al., 2019***).

Classically post-zygotic, genetic isolation is thought to arise due to epistatic interaction between loci, where alleles arise and fix in allopatry prior to secondary contact, i.e., the Bateson-Dobzhansky-Muller (BDM) model (***Bateson, 1909***; ***Dobzhansky, 1936***; ***Muller, 1942***). However, many incompatibilities uncovered using high-throughput molecular analyses (***Kuzmin et al., 2018***) and quantitative traits loci mapping (***Turner et al., 2014***; ***Chae et al., 2014***), do not conform to the processes suggested by BDM model. In particular, in both natural populations and model organisms, studies have found reproductive barriers exist between allopatric populations experiencing similar selection pressures (***Schluter, 2009***) and many of the alleles underlying genetic incompatibility predate the allopatric split of populations (***Marques et al., 2019***). Both of which are clear violations of the BDM model. As a result, why and how genetic incompatibilities arise without divergent selection and involve alleles that pre-date the allopatric separation of populations remains one of the most profound questions in evolutionary biology ***Marques et al.*** (***2019***).

Analytical and computational models have proposed theoretical explanations for the observed patterns of complex genetic interaction underlying post-zygotic isolation. A collection of models considered *de-novo* allele substitutions on the population level and the accompanying accumulation of hybrid incompatibilities. For example, ***Orr*** (***1995***) predicted that the number of incompatibilities should increase faster than linearly with the number of substitutions. The study by ***Orr*** also suggested higher prevalence of complex genetic interactions than simple pairwise incompatibilities. This so-called “snowballing" effect has been further extended by incorporating protein-protein interaction and RNA folding (***Livingstone et al., 2012***; ***Kalirad and Azevedo, 2017***).

The substitution-based approaches nevertheless are largely incompatible with emerging data on the evolutionary history of alleles involved in reproductive isolation (***Marques et al., 2019***). In addition, many models make an implicit assumption that two allopatric lineages only differed by fixed alleles, which does not capture the empirical diversity among individuals' gene expression in natural populations (***Gould et al., 2018***). More importantly, substitutions originating from *de-novo* mutations fail to explain the recent evidence that ancient alleles underlying reproductive barriers often predate speciation events (***Sicard et al., 2015***; ***Meier et al., 2017***; ***Nelson and Cresko, 2018***; ***Wang et al., 2019***; ***Duranton et al., 2019***; ***Marques et al., 2019***).

Another class of computational approaches focused on the overall regulation structure that is potentially accountable for complex genetic interactions, whose evolution then creates a feedback loop to generate hybrid incompatibilities. Gene regulatory networks (GRNs) describe inter-dependencies between gene expression and encode information of individual development on the molecular level. ***Johnson and Porter*** (***2000***) simulated a single linear regulatory pathway as a sequence of matching functions for binding sites, which resulted in reduced hybrid fitness compared to non-epistatic models. ***Palmer and Feldman*** (***2009***) explored the developmental process where the expression of gene products was iteratively determined through the regulatory networks. Diverse dynamics of hybrid incompatibilities was revealed which suggested the role of neutral gene regulatory evolution on speciation. Recently, ***Schiffman and Ralph*** (***2018***) modeled gene networks as linear control systems and demonstrated that reproductive isolation can be a consequence of parallel evolution of GRNs with equivalent mechanism.

The implications from gene network evolution are not mere outcomes of incorporating complexity into existing computational models. Instead, it is natural to consider GRNs to study evolutionary processes due to their close relation to coding sequences. Ideally, and hypothetically given “omniscience" over the genomes including comprehension of every fundamental interaction between molecules, one can reconstruct inter-dependencies among genes and thus obtain the GRN from a bottom-up approach. Of course, this ambition is far from practical and even sounds like a fantasy. Yet, it shows that GRNs are essentially a direct abstraction of the genome sequence. Furthermore, this abstraction has been proposed as the heart of the omnigenic perspective of complex traits (***Boyle et al., 2017***), which aims to ultimately map genotypes to phenotypes. GRNs therefore bridge the gap between inheritance factors and physiological traits, whose dynamics over generations then becomes a candidate to understand speciation.

Moving beyond substitution-based approaches, models that consider the evolution of GRNs are more flexible and can embrace recent observations such as the rich genetic variation in natural populations and the that incompatible alleles often far predate speciation events. That modern genetic details on incompatibilities are often opposed to existing theory is well articulated by ***Marques et al.*** (***2019***) who suggested that these two lines of empirical evidence can be consolidated into a “combinatorial view" of speciation. The combinatorial mechanism proposes that, if there was a past admixture event or if standing genetic variation persists, the reassembly of these old genetic variants can facilitate rapid speciation and adaptive radiation. Here, we integrate the combinatorial view and the evolution of GRNs. Specifically, we study the inherited molecular pathways encoded in GRNs, which are established upon genetic elements and propagate chemical signals that produce physiological traits. These pathways amplify a gene networks' potential to disentangle the genotype-phenotype map in light of epistasis.

Specifically, we propose a pathway framework for studying the evolution of genetic interactions that considers GRNs as a functional representations of coding sequences. The pathway framework takes a network-science approach to model how a current generation's GRNs bring forth the next generation's GRNs. Presuming ancestral variation as in the combinatorial view of speciation, the dynamics of individuals' gene networks was simulated through a naive model integrating independent assortment during sexual reproduction, genetic drift resulting from finite population size and natural selection on gene network functionality. We observed emergence of reproductive barriers among allopatric populations under identical selection pressure, where early evolutionary divergence between lineages was critical for barriers to arise. We concluded that it was the functional degeneracy nature of GRNs that accommodated potential lethal pathways in a diverse genetic background and leaded to reproductive barriers.

Results

The Pathway Framework: Networks as a Functional Representation of Genetic Interactions

Gene interactions networks are conventionally built such that genes are "nodes" and interactions between genes are "edges" or links, for examples see ***Tong et al.*** (***2004***); ***Schlitt and Brazma*** (***2007***); ***Langfelder and Horvath*** (***2008***). Here we propose an alternative methodology, termed the *pathway framework*, for constructing gene interaction networks. The key idea is to conceptualize genes, or alleles of genes, as "black boxes" that describe their expression behavior. More precisely, the pathway framework transforms alleles of genes into directed edges pointing from nodes that are activator/repressor molecules, e.g., transcription factors, and nodes that represent gene products, e.g., proteins. For example, in ***Figure 1*** we show how: a.) a gene is activated by a transcription factor and generates a protein product (top-right), b.) two genes interact via a transcription factor created by one gene that activates the other (middle-right), and c.) genes can interact via shared transcription factors (bottom-right). As a result of its flexibility, arbitrarily complex genetic interactions can be encoded as "pathways" through a gene interaction network.

Importantly, while our proposed representation is closely related to conventional gene interaction networks (and a direct mapping between the two always exists when considering interactions mediated by a single class of molecules, e.g., proteins), the pathway framework is often either a more compact or informative representation. For example, anytime a gene is regulated by a protein product from another gene, the conventional framework usually show redundancy that does not appear in the pathway framework, and the pathway framework will capture information not present in the conventional construction, e.g., see Box 1. Because the computational complexity of network analyses often scales non-linearly with the number of edges, switching to the pathway framework can facilitate a more robust exploration of model space.

The pathway framework further highlights how phenotypes are a product of both genetics and the environment (not all nodes in the pathway framework need be gene products). Concentrating on the molecular basis of physiological traits, a phenotype can be thought of as the biochemical status of a universal collection of nodes in the pathway framework, e.g., gene products such as proteins or environmental stimuli. Therefore, under the pathway framework, the development of a phenotype can be viewed as an iterative process of chemical signals propagating through woven pathways built from a groups of "inherited metabolisms", namely the functionality of genes, and external signals from the environment. As a result, the pathway framework can readily capture genetic, environment, and gene x environment effects in the same network.

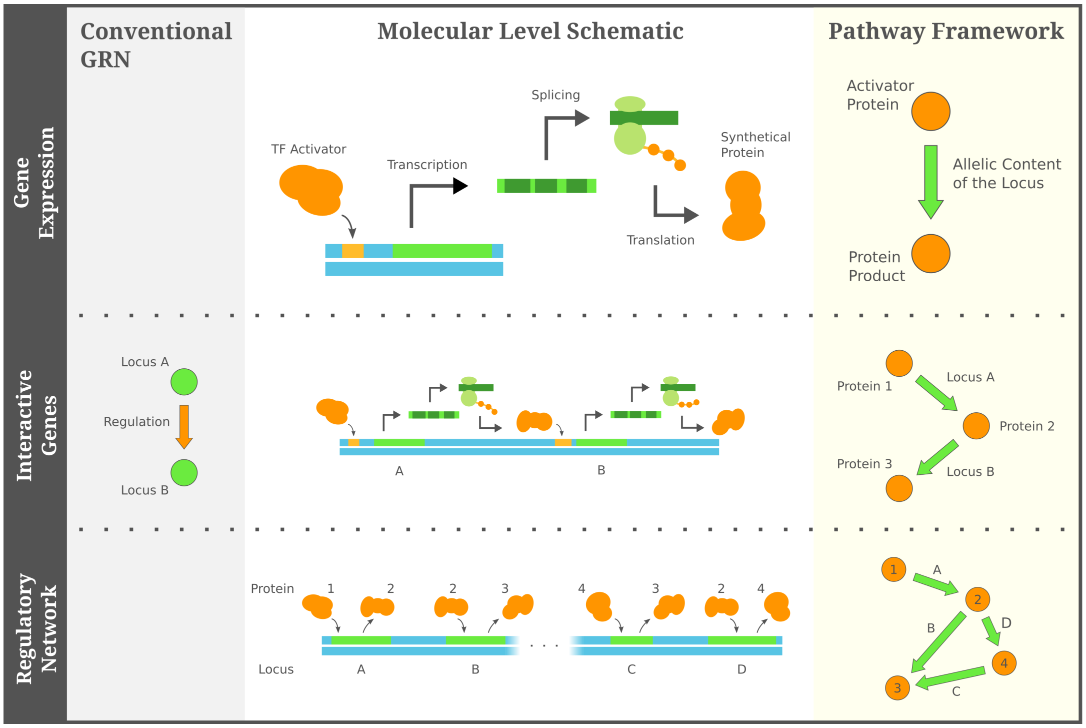
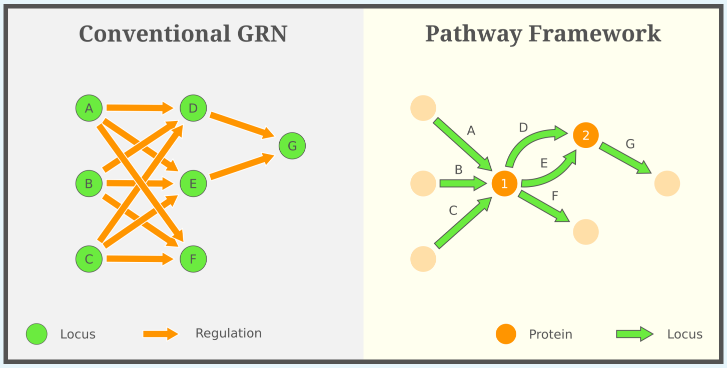


Figure 1. Pathway framework captures complex genetic interactions through consecutive regulatory pathways. In contrast to directly representing genetic interactions as in conventional GRN, the pathway framework abstracts genes as black boxes of their expression behavior. It turns alleles of genes into edges between the transcription factors and the protein products, and regulatory interactions between genes are encapsulated by consecutive pathways.

Box 1. Pathway framework is often a more compact representation

The pathway framework is usually more informative than the conventional construction of GRNs since it directly shows the expression behavior of genes. When considering genetic interactions that are mediated by a single class of molecules, e.g., one gene being regulated by the protein product of another, the pathway framework takes advantages of this information and presents genetic interactions in a compact pathway format. On the contrary, a conventional GRN lacks the specific regulatory context, and thus it has to present all pairs of interacting genes as individual edges rather than summarizing them by a smaller set of protein mediators. More technically, the pathway framework and a conventional GRN correspond to the first- and second-order de Bruijn graph (*De Bruijn, 1946*) respectively, and higher-order de Bruijn graphs usually tackle combinatorial problems at the cost of introducing redundant elements.



Evolutionary Mechanisms under the Pathway Framework

Although in its most abstract state, the pathway framework can include nodes that are not proteins and not directly involved in gene regulation, we focus here on the evolution of GRNs where all nodes are proteins directly involved in transcriptional regulation. To model and simulate the evolution of GRNs, the pathway framework translates evolutionary mechanisms, such as mutation, independent assortment, recombination, and gene duplication, into graphical operations on the gene networks[1]. Because mutation of a locus can potentially alter its protein product and/or the transcription factor binding region(s), we consider mutation as rewiring process where the incoming and/or outgoing directed edges are re-directed to point from or to different nodes (***Figure 2***, top-right). Independent assortment during meiosis can be modeled via edge-mixing of parental GRNs such that an offspring acquires alleles, i.e., edges in the GRN, from both parents (***Figure 2***, bottom). Similar to mutation, recombination is an edge-rewiring process that is constrained to swapping binding sites or transcription factors at the same locus. Finally, gene duplication is equivalent to adding a parallel edge that represents the identical allelic content of a duplicated locus.

An individual's viability subjected to natural selection is a response to the molecular phenotypic status, which, under the pathway framework, can be modeled as a fitness function associated with the collective state of nodes in the gene network. For example, one could study the time-varying concentration of each protein, attach a continuous dynamic or a stochastic reaction to every allele and define fitness as a function of the high-dimensional concentration vector, etc.. On the other extreme, we instead consider Boolean networks, which have been shown to effectively portray many of the relevant dynamical features of empirical regulatory systems (***Davidich and Bornholdt, 2008***). In this minimal scenario, each protein is assigned to a Boolean state — present or absent. External environmental signals stimulate the existence of some proteins in the organism. The logical states then cascade through the genetic pathways, where given the presence of a gene's transcription factor, its allele turns on and generates a protein product. The phenotype of a GRN is thus the “reachability" from the environmental stimuli, whose binary survival is defined via a sharp fitness landscape over plausible collective Boolean states (***Figure 2***, top-left).

We further adopt the Boolean-state assumption of GRNs because it readily sheds light on the formation of hybrid incompatibilities. A hybrid incompatibility is a combination of alleles that were separated in parental lineages but are present in hybrids and cause fatalities. Moreover, the combination is minimal in the sense that the lack of any of its allelic elements will not lead to an inviable hybrid. In the pathway framework, suppose that the binary viability only depends on a set of lethal proteins, i.e. an individual will not survive selection if any of those protein are present, a combination of alleles that includes a pathway from a environmental stimulus to a lethal protein makes the GRN inviable. If the alleles exactly comprise a simple path, which contains no cycles, they become a minimal combination and thus form an incompatibility. Additionally, The complexity of genetic interactions can be characterized by the number of alleles involved, which is called the order of hybrid incompatibility and related to the length of the simple pathway[2].

Simulating the Evolution of GRNs

Briefly, we first consider a Wright-Fisher model of evolution with natural selection, i.e., constant population size, no mutation, no migration, non-overlapping generations, and random mating. Selection occurs during the haploid stage of the life-cycle, which fuse randomly after selection, i.e., create diploids, and undergo meiosis to generate the subsequent generation (simulations are further detailed in the Methods). Populations are seeded such that each individual has a randomly generated GRN and evolve until a single GRN fixes in the population.

Figure 3a shows the proportion of individuals in the population that survive natural selection.

[1] These graphical operations particularly focus on edges in the GRNs, while remaining the underlying node set constant because the nodes represent all *possibly existing* proteins in the organism.

[2] Since for $n \geq 1$, $n+1$ alleles form an nth-order incompatibility, the order of genetic interaction is then the path length minus one.

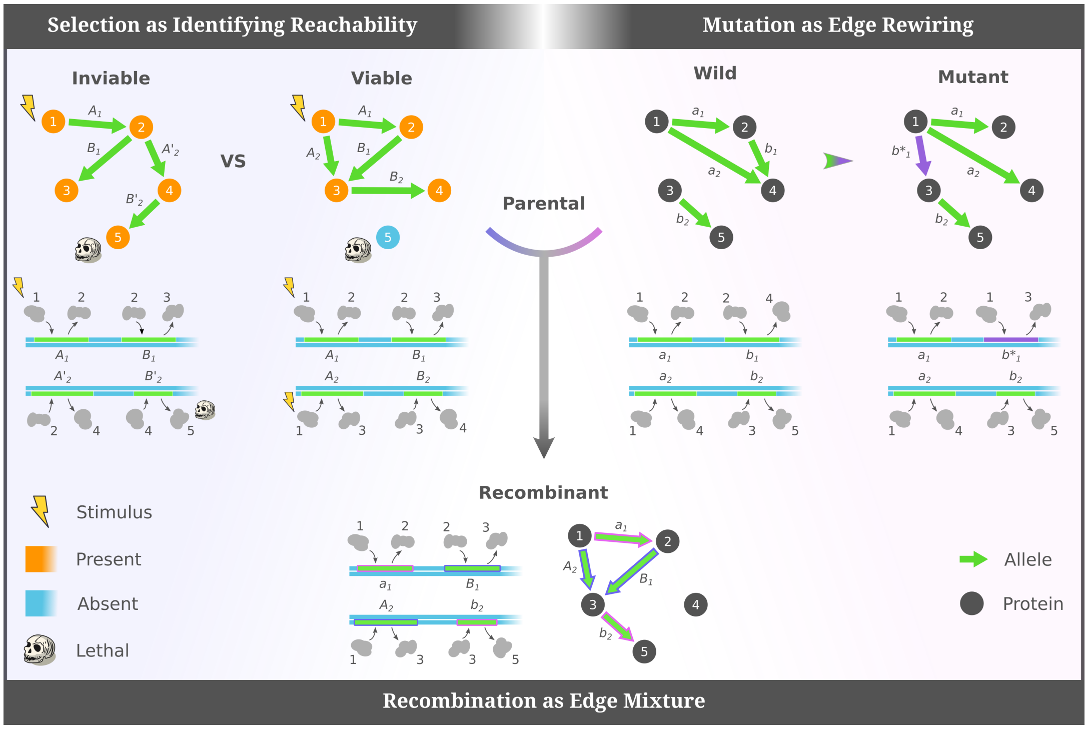

Selection as Identifying Reachability
Mutation as Edge Rewiring
Inviable
VS
Viable
Wild
Mutant
Parental
Recombinant
Stimulus
Present
Absent
Lethal
Allele
Protein
Recombination as Edge Mixture

Figure 2. Pathway framework turns evolutionary mechanisms into graphical operations on the GRNs. Since the pathway framework directly models the functionality of alleles of genes as edges, mutation and recombination can be modeled as edge-rewiring and edge-mixing respectively, while a minimal selection scenario of binary fitness can be modeled as identifying reachability in the GRNs.

Initially, the fraction of viable individuals differed dramatically between simulations with different initial conditions due to the variation of randomly seeded GRNs. As the gene networks evolved, the population's viability increased and quickly reaches a state where every individual survives selection (dashed line). During this 100% survival stage, natural selection was no longer effective and the population evolves to fixation via genetic drift. Not surprisingly, our results demonstrate that GRNs can rapidly evolve from a heterogeneous population with low average viability to "match" and imposed environment.

In addition to achieving 100% survival, populations always fix for a single GRN. ***Figure 3***b plots the number of structurally-distinct GRNs in each generation. The decreasing trend demonstrates that, although various GRNs have equal survival probability, it became more and more likely that individuals shared a common GRN. Moreover, the populations always fixed a single GRN (dotted line) after a sufficiently long period of time. This phenomenon can be intuitively explained by the mechanism of sexual reproduction. In our model, parents with identical GRN would lead to offspring of the same GRN, since any two corresponding groups of segregated alleles retrieved the parental gene network. Thus once there was a majority gene network in the population, it has a higher chance to retain its genetic configuration in the next generation rather than being replaced by shuffled variants.

Lastly, to better understand how parallel lineages evolve, we consider a scenario where multiple allopatric populations are seeded with the *same* initial conditions. Similarly, each allopatric population rapidly achieves 100% survival and then fixes a single GRN. However, across allopatric populations, seeded from the same initial conditions, many different GRNs fixed. ***Figure 4*** presents the distribution of fixed GRNs for a smaller-scale simulation (Setup 2 in Methods). We see that the fixed gene networks were diverse and non-uniformly distributed. Despite being under identical selection forces and having the same initial condition, lineages evolving from a common ancestral population fixed alternative GRNs. This result demonstrates that a broad range of GRNs can survive

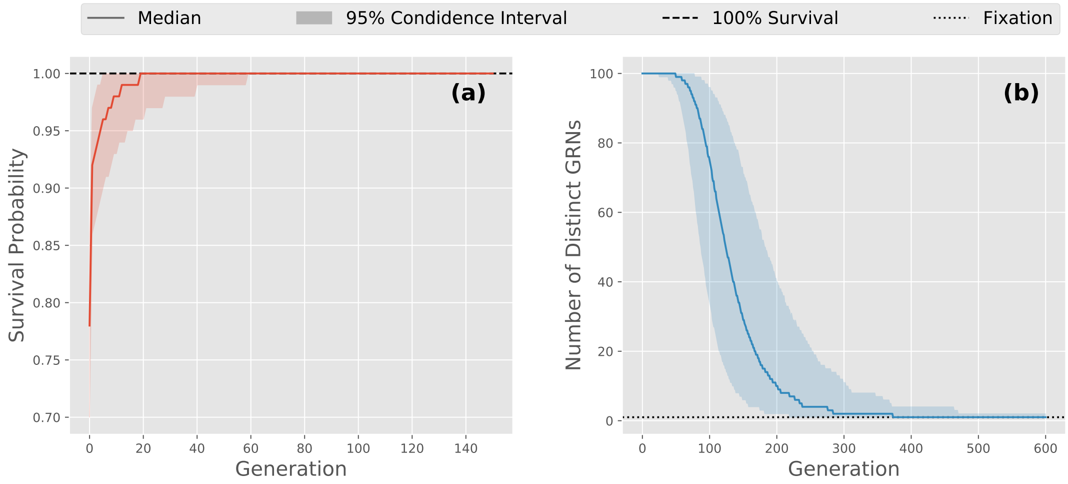


Figure 3. Populations adapted to the environment and then fixed a single GRN. Here we show, for every generation of GRN evolution across multiple allopatric populations with different initial conditions: **(a)** the survival probability of an individual and **(b)** the number distinct GRNs in each population, where two individuals' GRNs were deemed effectively identical if they were isomorphic. The average viability of each population increased over time and rapidly achieved 100% survival, which indicates that evolution of GRNs drove adaptation toward the imposed environment. We also observe decreased variation of GRNs as they evolved, with individuals in the same allopatric population, i.e., simulation run, eventually fixing for the same GRN.

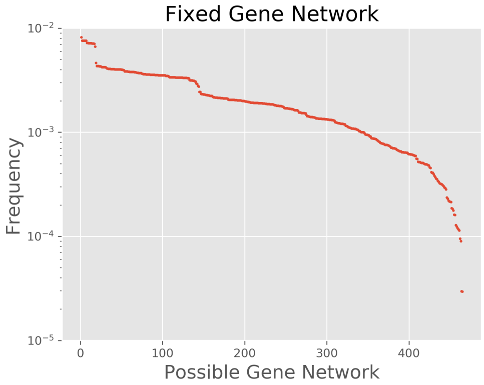


Figure 4. Fixation of parallel lineages resulted in a wide range of GRN structures. We simulated isolated populations from the same intial conditions until they reached fixation. In this case Setup 2 in Methods was applied in order to tractably enumerate all plausible GRN, and the ancestral populations were chosen such that the fixation was unbiased by the initial allele frequencies. The 10^7 acquired GRNs were categorized into 465 viable structures and the fixation frequency of each structure was plotted in a descending order. The distribution shows that isolated lineages fixed alternatives gene networks, some among which were more favorable under our model of GRN evolution.

the given selection pressure. Furthermore, none of the viable GRN structures had a zero fixation probability, indicating an thorough exploration of evolution in the space of possible GRNs. That so many different GRNs fixed suggests that evolution was less governed by a definite trajectory, but instead it occurs via an uncertain realization among all the possibilities constrained by the ancestral population and the selection pressure.

Reproductive Barriers Arose Rapidly as Gene Networks Evolved

If the survival probability and fitness of GRNs were identical, the distribution of fixed networks should be uniform over all viable conformations. Because we observe a strongly non-uniform distribution (see ***Figure 4***) some other form of selection is likely operating on the GRNs. We note that during random mating, even between two parents with viable GRNs, some of their shuffled offspring can be inviable. Coupled with the observation that different allopatric populations, i.e., simulation runs, fix alternative GRNs from the same initial conditions, we hypothesized that some degree of reproductive isolation may exist between these fixed populations.

To test for the presence of reproductive isolation, we performed a "hybridization" experiment between parallel lineages that had reached fixation. Starting with lineages branched from a common ancestral population, two fixed lineages were randomly selected and interbred. Hybrids were generated and the reproductive isolation metric (RI) between the parental populations was computed (see Methods). By repeating this procedure, we obtained a distribution of reproductive isolation, as demonstrated in ***Figure 5***a inset. Despite a large fraction of crosses resulting in nearly zero RI, we discovered pairs of lineages with positive reproductive isolation metric. Specifically, the RI distribution displays several regions of positive reproductive isolation such that a high percentage of hybrid offspring are inviable. Thus, we conclude that reproductive barriers between fixed lineages, derived from the same initial population and experiencing identical selection, exist.

Given noticeable reproductive barriers between fixed lineages, we further studied when those barriers first manifested during GRN evolution. Note that because our simulations did not contain mutation, incompatibilities arise because of shuffling during meiosis. Here, instead of waiting until GRN fixation, we instead evolve lineages for T generations and then cross them to generate hybrids as described above. By varying T, a series of reproductive isolation distributions were acquired. ***Figure 5***a collects and displays them in a heat map. A vertical slice represents a RI distribution as in the inset panel, but crosses were made after T generations rather than waiting for lineages to reach fixation. We see that the regions of high incompatibility noted in ***Figure 5***a inset becomes bands in the heat map, which allows us to trace the emergence of reproductive barriers.

Initially the reproductive isolation distribution was relatively symmetric around zero. However, As GRNs evolved, the range of RI broadened and its extreme value in the positive tail increased. The trend towards higher levels of RI decelerated after 100 generations; it then stabilized and formed a band structure, where crosses cluster around certain levels of reproductive isolation. ***Figure 5***a hence reflects that reproductive barriers existed at low levels as soon as the lineages started evolving independently and peaked at a time prior to GRN fixation. By assumption, the alleles underlying RI were present in the ancestral population, but we further conclude that RI peaked well before fixation of GRNs.

Next, for incompatible hybrids generated in our crossing experiment, we determine how complex the underlying mechanism of RI was. Specifically, ***Figure 5***b shows how frequently an inviable hybrid resulted from an incompatibility of a certain order. We see that hybrid incompatibilities spanned over a broad range of interaction orders. Importantly, the simple two-allele interaction was only slightly more common than incompatibilities resulting from three or four interacting alleles and that interactions above forth order made up 2.79 percent of all incompatibilities. However, we note that the frequencies of incompatibility order varied depending on the ancestral population.

The pattern of complex genetic interactions provides insights on the distribution of reproductive isolation. Based on the independent assortment mechanism in our model–and assuming that multiple incompatibilities rarely occurred between two parental GRNs–we conclude that hybrid incompatibilities quite often involved higher order interactions, which did not arise as a result of selection, but simply were an expected consequence of GRNs being high order (***Appendix 1***). Further, the discrete characteristic of hybrid incompatibilities led to a higher likelihood at certain RI levels. The band structure in ***Figure 5***a agrees with this prediction (***Appendix 1***), which suggests that reproductive barriers are strongly influenced by the concealed hybrid incompatibilities and are coupled with the genetic interaction pattern shown in ***Figure 5***b.

Early Divergence between Lineages was Critical for Reproductive Barriers to Emerge

To further study the emergence of reproductive barriers in our model, we investigated the relative importance of various evolutionary forces in generating the observed patterns of RI. In particular, were the barriers attributed to selection pressure, random genetic drift, or both? We designed two “control scenarios” that were based upon the previously simulated model, but contained modifications to remove the effects of either selection or drift. Comparing the strength and pattern of RI resulting from the two control scenarios, i.e., the removal of drift or selection, to the original GRN dynamics, which contain both evolutionary forces, provides an assessment of the removed component’s role in shaping the observed pattern of RI.

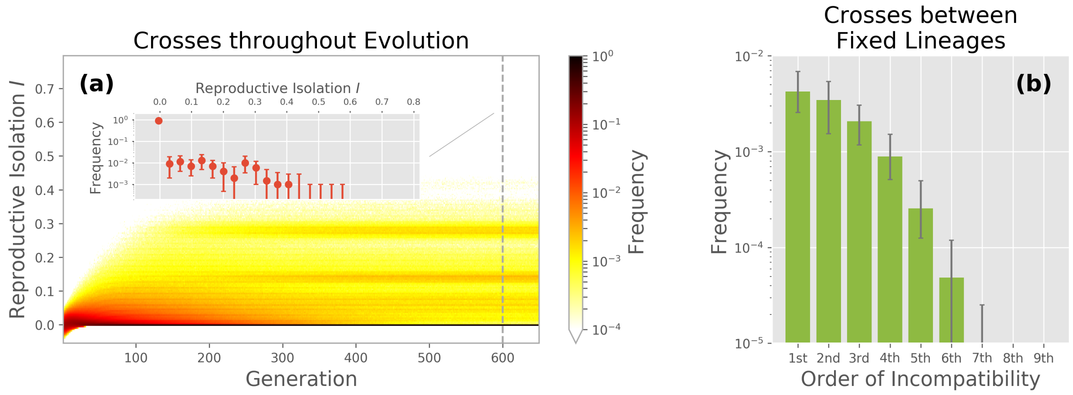


Figure 5. Reproductive barriers arose rapidly between allopatric populations. (a, Inset) Distribution of reproductive isolation between pairs of fixed lineages. A non-negligible fraction of crosses led to positive reproductive isolation, which reflects the occurrence of inviable hybrids and indicates reproductive barriers between fixed lineages. **(a)** We crossed allopatric populations at every generation during GRN evolution and stacked the RI distributions into a heat map. A vertical slice in this heat map represents the RI distribution at a given time, similar to the inset, but where the color shows the mean frequency for each bin. The growing level of positive RI indicates that reproductive barriers arose at the early stage of evolution. **(b)** Frequency that incompatibilities with various order were observed among hybrids between fixed lineages. We see that the order of incompatibilities included a broad range and that the simple pairwise interaction did not significantly dominate over more complex incompatibilities. Moreover, hybrid incompatibilities are consistent with the clustered level of RI and hence sheds light on the observed RI distribution (***Appendix 1***). In both the inset and panel (b), the plots show the statistic of the distribution among multiple groups of allopatric populations, specifically the median frequency and the 95% confidence interval.

Removing the effect of natural selection is straightforward to simulate. In this control scenario, populations simply evolve in a selectively neutral environment where all GRNs are viable. Thus, all individuals survived and genetic drift became the only remaining evolutionary force. Of course, this neutrality concurrently made the RI metric ill-defined. We avoided this issue in the crossing experiments to calculate RI by placing the parental populations under the same non-neutral environment in the original model, so the hybrids would be generated from survivors subjected to selection pressure. The reproductive isolation metric could then be computed with respect to the non-neutral environment. This ensures comparability between the model and the “no selection" control scenario since the survivability of hybrids was evaluated under the same environment and was not biased by the otherwise inviable parents.

Figure 6a shows the contrast of barriers observed in the original GRN evolution model (red) and in the scenario with no selection (blue). We traced the leading reproductive isolation over time, defined as the 99th percentile of the RI distribution, which is a sufficient indicator of reproductive barriers between lineages. We discovered that in both the model and the control scenario, the leading RI I^* increased and then saturated. Furthermore, the growth in I^* decelerated after a similar number of generations in both scenarios. That RI occurs at a higher level in the control experiment indicates that selection did not “cause" the fixation of barriers between allopatric populations, but instead suggests that selection was actually limiting chances for incompatibilities to occur in hybrids. We hypothesize that–although restricted as compared to drift–selection operating on incompatibilities likely induced the observed disconnect between viability and fitness seen in ***Figure 4***.

We next turned to the contribution of genetic drift to the emergence of reproductive barriers. The control scenario, however, was less straightforward due to technical difficulties associated with directly removing random genetic drift from the model. Neither abandoning sexual reproduction nor simulating an infinite population would result in non-trivial and/or computationally tractable GRN evolution. Alternatively, we designed a control scenario where the evolutionary influence

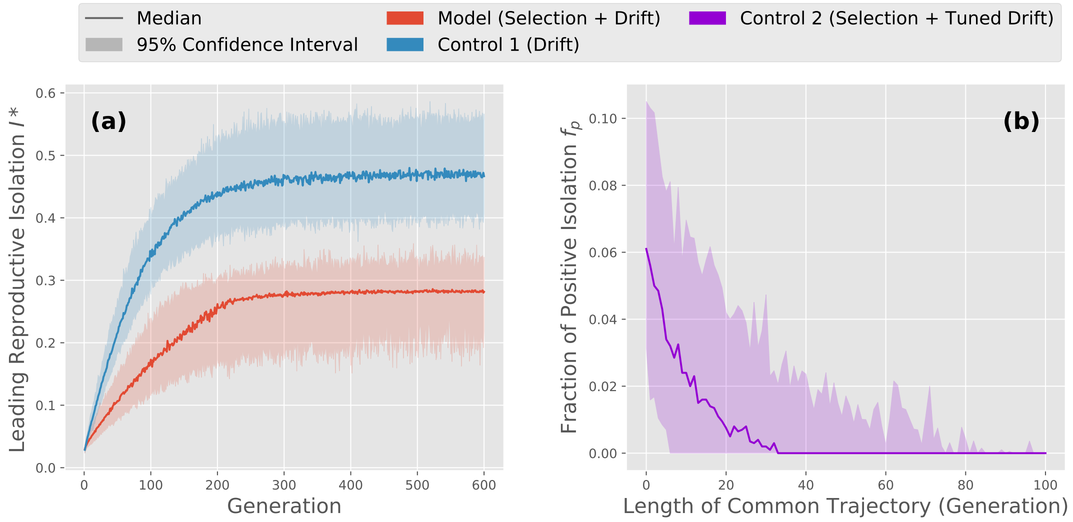


Figure 6. Early divergence of evolutionary trajectories between lineages was necessary for reproductive barriers to arise. Here we compare a statistic, termed leading reproductive isolation I^* (99th percentile of the RI distribution), measuring the degree of reproductive barrier in the original model and two designed control scenarios. Control scenarios were simulated with the same group of ancestral populations as the model, where lineages were then crossed to generate hybrids. **(a)** Leading reproductive isolation I^* among allopatric populations over time, where positive values indicate the existence of reproductive barriers. We plot the original model in red and the control scenario with a neutral environment in blue. The increasing and larger I^* uncovered in the control scenario implies that reproductive barriers were still observed when the selection forces were silenced. **(b)** Long-term fraction of positive RI f_p when the influence of random genetic drift was tuned. We simulated the evolution of lineages, but first confine them to a common trajectory of length L, which was realized by evolving a single population from the ancestors for L generations, and then simulated allopatric evolution from this now less diverse ancestral population. The original model corresponds to the case where $L = 0$, and for any positive L the effect of drift were lessened. We obtained the f_p metric when lineages evolved for 600 generations, where $f_p = 0$ suggests no barriers among populations. That f_p decreased with L to 0 shows that reducing the effect of drift diminished reproductive barriers. As a result, it implies the criticality of divergence among evolutionary trajectories for barriers to emerge.

of drift could be tuned and limited. Genetic drift results in stochasticity and causes populations to experience diverse trajectories. On the other side of the coin, if two lineages show similar evolutionary trajectories, one would say that drift effectively leads to less divergence between them. We restricted the influence of genetic drift by first confining lineages in a common trajectory for L generations, and then freed the populations and let them evolve independently, i.e., in allopatry. Varying the length of the common trajectory L tunes the overall similarity among lineages. L hence quantitatively reflects the strength of genetic drift.

Figure 6b demonstrates the long-term fraction of positive reproductive isolation introduced in Methods, termed f_p, as we varied the length of the common trajectory. Despite substantial variation in f_p in the original model, which corresponds to the case where $L = 0$, a decline of f_p was uncovered as early evolutionary confinement was extended. We discovered 50% of the experiments showed a zero f_p in the after lineages were evolved together for 40 generations, and as the length of common trajectory exceeded 80 generations positive reproductive isolation was hardly found between lineages. More importantly, ***Figure 6***b suggests that as the evolutionary influence of genetic drift was mitigated, RI was weakened and eventually vanished. Namely, restricting early divergence among populations due to genetic drift diminished reproductive barriers. This control scenario consequently suggests that, instead of the selection pressure, divergence between lineages, coupled with high diversity in the ancestral population, is critical for reproductive barriers to arise.

Intra-lineage Incompatibilities were Eliminated Stochastically While Inter-lineage Incompatibilities Persisted and Led to Reproductive Barriers

To better understand how reproductive barriers might be removed within a lineage, but persist between lineages, we computed two quantities from the underlying genetic pool. First, the size of the genetic pool, which determines how many possible genotypes a population contains. This measure captures the potential genetic diversity in the population. Second, we count the number potential incompatibilities in the underlying genetic pool, which are lethal allelic combinations that could potentially be realized in the next generation. These incompatibilities compose the source of inviable offspring and RI between allopatric populations. However, because even for small GRNs searching for all possible incompatibilities quickly becomes computationally intractable, we developed a novel algorithm (summarized in Methods) to compute their number in the genetic pool.

Because our model does not contain mutation, one would expect the size of the underlying genetic pool to decline in our simulated gene network evolution. Any allele in an individual was inherited from its parents, and thus it must appear in the parental generation as well. Additionally, a parental allele might not persist in the offspring for two possibilities: either it was not transmitted because of finite population size of the progeny generation and the stochasticity during sexual reproduction, i.e. drift, or it formed a lethal pathway along with other inherited alleles which made the offspring inviable, i.e. selection.

Figure 7a demonstrates the size of genetic pool over time, where we compare simulations in the original model (red) and in the control scenario without selection pressure, i.e., only genetic drift will reduce the size of the genetic pool (blue). A rapid decline of genotypic diversity was witnessed under both models. More intriguingly, little difference was found between the GRN evolution model and the control scenario under a neutral environment. The two median curves nearly overlaps, and for any given generation, the pool size in the original model was not significantly smaller than the control counterpart. Therefore, we find additional support for our earlier finding that although both natural selection and random genetic drift decreased genotypic diversity, drift was the dominant driving force. However, while the effect of drift reduced diversity within a lineage, it increased the divergence among lineages.

Figure 7b shows the number of potential incompatibilities within a lineage's underlying genetic pool (orange). We found that the amount of incompatibilities embedded in a population also

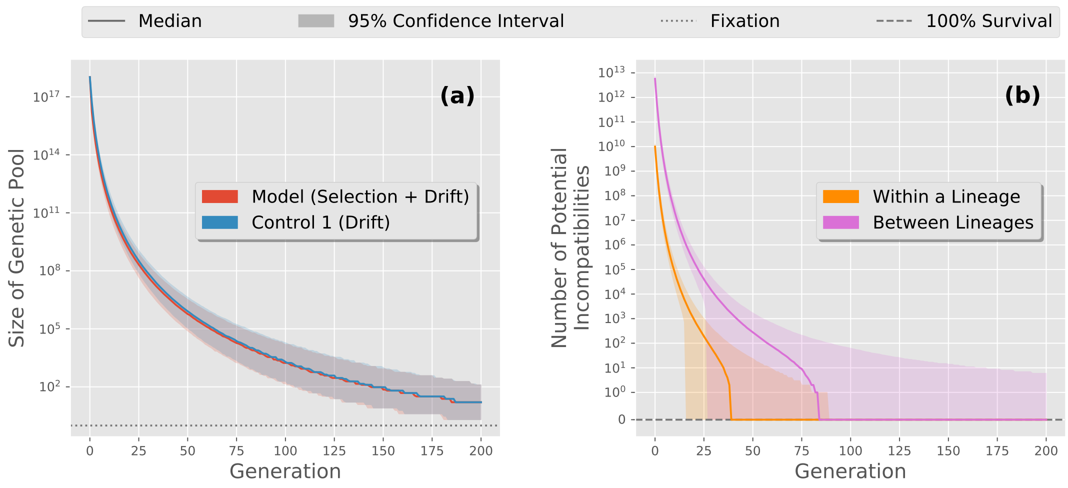


Figure 7. The underlying genetic pool lost alleles and eliminated potential incompatibilities within allopatric populations, whereas inter-lineage incompatibilities persisted. (a) Size of the underlying genetic pool for each generation, where we plot the original model in red along with the no selection control scenario in blue. Both cases show a similar reduction in the genetic pool. The similarity of these curves suggests that the continual losses of allelic diversity within a lineage was dominated by random genetic drift. **(b)** Number of potential intra-lineage (orange) and inter-lineage (pink) incompatibilities for each generation in the original model. We found that the number of potential incompatibilities also decreased as GRNs evolved, which is explained by the reduced allelic diversity in the genetic background. The vanishing intra-lineage incompatibilities implies disappearing sources of inviable hybrids, and it provides a mechanistic understanding of how a genopytically rich populations adapted to the imposed environment. Contrarily, the intra-lineage incompatibilities remained during GRN evolution. It was the persistent potential incompatibilities between allopatric populations that led to evident reproductive barriers.

Figure 7–Figure supplement 1. Inter-lineage incompatibilities were sustained throughout GRN evolution.

decreased over time. This phenomenon is understood by the continual loss of allelic diversity, since removing an allele from the underlying pool always restricts the possibilities to form a lethal pathway in the GRN. Furthermore, the number of potential incompatibilities fell rapidly until no potential incompatibilities remained. The elimination of potential incompatibilities illuminates how a population adapted to the imposed environment when GRNs evolved, as shown in ***Figure 3***a. Random genetic drift drove the loss of a lineage's genotypic diversity, and along with the guidance of selection, it eliminated probable lethal pathways in the genetic background. Once all the potential incompatibilities were eliminated, no source of inviable offspring existed and consequently the population reached 100% survival. Again, this result supports our earlier finding that natural selection was operating against incompatibilities within a lineage, but that drift was nevertheless the dominate force in structuring incompatibilities between lineages.

Finally, we investigated incompatibilities between underlying pools of lineages, which we call the "inter-lineage" incompatibilities, as compared to potential lethal allelic combinations within a population termed "intra-lineage" incompatibilities. ***Figure 7***b presents the number of inter-lineage incompatibilities over generations (pink). We observed more incompatibilities between allopatric populations than those within a population, i.e., sympatric RI, and similarly their amount dropped as allelic diversity decreased. In contrast, inter-lineage incompatibilities were removed at a slower pace compared to intra-lineage incompatibilities. The sustained confidence interval further suggests that some inter-lineage incompatibilities persisted, which was also the case after populations reached fixation (***Figure 7–Figure Supplement 1***). The persistence of these potential incompatibilities qualitatively explain the inviable hybrids revealed after GRN evolution. In spite of lineages adapting to the same imposed environment, hybrdiziation can "resurrect" a lethal combination of alleles, which was eliminated in either lineages yet remained in their joint genetic background. This explanation also supports the stronger barriers uncovered in the neutrally evolving control in ***Figure 6***a, since inter-lineage incompatibilities would be more persistent without the constant selection pressure (***Figure 7–Figure Supplement 1***).

Discussion

In this work, we propose a path-oriented construction of GRNs where alleles are labeled and presented by their functionality. The pathway framework brings a natural perception of GRNs considering how a genotype can give rise to a phenotype, and it allows us to apply network science analyses to study the process of speciation. We simulated the generational dynamics of gene networks via a model incorporating natural selection, segregation and random sampling. With the presumption of ancestral genetic variants, a population adapted to the imposed environment and fixed a single GRN, whereas parallel allopatric populations resulted in alternative regulatory structures. More importantly, we discovered reproductive barriers that arose rapidly among allopatric lineages even under the same selection pressure.

We also provide a mechanistic illustration of how reproductive isolation emerged as GRNs evolved. Early evolutionary divergence of lineages, particularly the way they lost accessible alleles in their genetic background, established the base of reproductive barriers. Despite that allopatric populations adapted to the imposed environment whose genetic background no longer contained lethal allelic combinations, potential incompatibilities could persist in the joint background of two parallel lineages. Interbreeding them might therefore resurrect previous removed incompatibilities and led to inviable hybrids.

The persistence of inter-lineage incompatibilities implies co-occurrence of many GRNs with negative reproductive interaction under the same selection force. This “functionally degenerate" characteristic of GRNs reflects the concept of genetic redundancy (***Nowak et al., 1997***; ***Láruson et al., 2020***), and it resonates with earlier studies that suggested alternative regulatory structures to achieve the same phenotype (***True and Haag, 2001***; ***Wagner and Wright, 2007***; ***Schiffman and Ralph, 2018***). Our pathway framework illustrates why degenerate genotypes can naturally arise. Once the alleles are presented as functional pathways connecting a underlying group of proteins, the conjunction between genetic factors and physiological traits is no longer a bipartite mapping; the phenotype, as the collective chemical status of proteins, is a convolution of active signals and external stimuli propagating on the network consisting of genetic pathways. The pathway configuration that satisfies an acknowledged environmental input and phenotypic output is, as a result, not unique. One could find numerous functionally degenerate gene network structures fulfilling the input-output pair, as what ***Figure 4*** demonstrates, whereas mixing edges between two GRNs possibly leads to a fatal pathway and hence an inviable offspring. Therefore, we evidence that the pathway framework underlines the role of GRNs in speciation processes through the innovative edge-and-node interpretation between genotypes and phenotypes.

Our minimal model of GRN evolution encapsulates selection through binary viability, which is essentially a special of holey adaptive landscapes (***Gavrilets, 1997***). ***Gavrilets and Gravner*** (***1997***) introduced a multi-locus model where each genotype was independently assigned to one of the two fitness level. The study suggested that reproductive isolation could arise from the high dimensionality of the genotype space, which bypassed and connected seemingly disjoint genotypic regions. In a similar spirit, our model further ties the high dimensionality of genotypes to complex genetic interactions; under the pathway framework, inviability originates at the mechanism of hybrid incompatibilities, i.e., allelic combinations that form lethal pathways in a GRN. The pathway framework also features flexibility, and in future works it can be combined with other fitness landscapes that have been investigated in the speciation literature. For example, ***Barton*** (***2001***) demonstrated that stabilizing selection can generate reproductive isolation, and the pathway framework can be easily embedded into such a continuous fitness landscape.

Our work endorses the latent connection between speciation processes and ancestral genetic variation. Ancient polymorphisms not only confound inference of evolutionary processes that can drive genomic divergence (***Guerrero and Hahn, 2017***), but they have also been hinted as a potentially good substrate for rapid speciation through the combinatorial mechanism (***Marques et al., 2019***). In particular ***Marques et al.*** reviewed that old genetic variants had underwent selection and thus likely to be beneficial, they would have higher allele frequency than *de-novo* mutations, and they could enrich large-effect haplotypes and more. Alternatively, we demonstrate that stochasticity of losing accessible pathways in GRNs relatively thrived selected functional regulatory structures among ancestral polymorphisms. Segregating these regulatory structures may notwithstanding upraise deadly pathways. Our pathway framework hence adds theoretical supports to findings of substantial inheritable polymorphism in hybrid incompatibilities, as reviewed in ***Cutter*** (***2012***). We suggest to consider evolution of regulatory pathways as a parallel mechanism with which ancestral genetic variation can facilitate appearance of new species.

In principle, any group of ancestral polymorphisms that encodes a lethal regulatory pathway induces a non-zero chance of reproductive isolation. We numerically assessed the strength of reproductive barriers reflecting on the tuned ancestral variation (***Appendix 2***). For finite-size allopatric populations, there appeared a critical amount of variants to observe evident barriers. Further theoretical efforts are required to quantitatively comprehend the strength of barriers and its relation with the extent of ancestral variation. First, one needs more advanced analyses than ***Appendix 1*** to evaluate the survival probability of hybrids given multiple incompatibilities embedded in parental GRNs. Second, the likelihood that a certain incompatibility lies between two parental GRNs depends on the balanced distribution of regulatory structures, for instance ***Figure 4*** as the case at fixation. The skewed patterns of fixed GRNs sketches that some regulatory structures are more favorable than others under evolution. Understanding the balance between gene regulation is necessary to model the dynamics of hybrid incompatibilities.

Methods

Numerical Simulations

General Schema and Assumptions

In this work we simulated evolution GRNs in allopatric populations. Throughout evolution, we assumed that individuals had a constant number of loci and thus a fixed number of edges in their GRNs. The underlying set of nodes in GRNs also remained unchanged as we reasoned in Results. We further introduced different categories of nodes/proteins to concrete the space of plausible alleles. Some proteins were presumed to only be present with the environmental stimuli, which were not products of any locus; on the other hand, some other proteins were presumed to have mere physiological effects, and thus they were not capable of activating gene expression. We called them source proteins and target proteins respectively. A plausible allele was therefore labeled by a non-target protein that could activate its expression and a non-source protein that would be synthesized. In our simulations we supposed only one source protein and one target protein.

We considered a naive model of GRN evolution incorporating natural selection, independent assortment and random genetic drift. The environmental condition was set fixed over time and across populations. We assumed that the environment stimulated presence of one protein and it specified another protein with a lethal effect[3]. Viability of individuals was presumably equated to the reciprocal binary state of the lethal protein. Hence given the current generation, individuals were selected such that whoever did not possess a pathway from the environmental stimulus to the lethal protein survived and were able to reproduce.

The survivors then randomly mated and formed the next generation with independent assortment. Here we assumed individuals with haploid-dominant life cycles, where the multicellular haploid stage is evident[4]. Supposed even segregation during meiosis of the diploid zygotes, we modeled the process of independent assortment as follow. Two parental individuals were randomly

[3] Specifically, they reconciled with the source and the target protein respectively.

[4] During reproduction, specialized haploid cells from two individuals combined and formed a diploid zygote. The zygote experienced meiosis and generated haploid spores, which then developed into multicellular-haploid-stage individuals through mitosis.

sampled from the survivors. The set of loci was first randomly partitioned into two groups of equal sizes. The offspring inherited alleles of one group of loci from one of its parents and alleles of the remaining loci from the other parent. Hence half of the edges in the offspring's GRN came from one parent's GRN and the rest was acquired from the other. This procedure was repeated until the next generation had the same constant population size as their predecessors.

Simulations and Parameter Setups

Here we summarize the two different parameter setups in our simulations:

Setup 1: We assumed 11 possibly existing proteins in the organism. A generation was composed of 100 individuals with 10 loci each. We generated 100 ancestral populations where individuals' GRNs were randomly sampled from all plausible genotypes. For every ancestral population, we in parallel ran 100 simulations from it, which were regarded as lineages evolving in isolated geo-locations.

Setup 2: We assumed 5 possibly existing proteins in the organism. A generation was composed of 16 individuals with 4 loci each. We generated 10^4 ancestral populations induced from a genetic pool[5] containing all plausible alleles for each locus. For every ancestral population, we in parallel simulated 10^3 lineages from it.

The randomly generated ancestral populations encapsulate our assumption of ancestral genetic variation, which reflect divergence of gene regulation that has been found in empirical studies (***Gould et al., 2018***). Setup 2 aimed to examine how broadly, in terms of fixed GRNs, evolution can explore in all possibilities. Thus it consisted of a larger amount of simulations starting with unbiased ancestral populations that were induced from a maximal genetic pool. If not otherwise specified, simulations shown in Results were run under Setup 1.

When we inspected reproductive barriers between allopatric populations by interbreeding them, we first sampled 1000 pairs of lineages and then each generated F_1 1000 hybrids. The survival probability of hybrids can then be obtained for all crosses. The same sampling procedure was also applied when we computed the number inter-lineage potential incompatibilities between pairs of allopatric populations.

Metrics of Reproductive Isolation

We introduce a quantitative measure of reproductive isolation between lineages which evolved from a common ancestral population. Given a group of lineages and a chosen pair among them, the reproductive isolation between the pair is defined as the relative difference of hybrid survival

$$I = \frac{p_c - p_h}{p_c} \tag{1}$$

where p_h is the survival probability of F_1 hybrids, and p_c denotes the average of survival probabilities of all lineages' next generation. A positive value of reproductive isolation I implies that the hybrids have less survivability than the expectation of the offspring. In the extreme case where no hybrid lives, $I = 1$. It therefore serves as an indicator of reproductive barriers between two lineages.

Strengths of reproductive barriers among the group of lineages are described through a distribution of reproductive isolation, which can be obtained by sampling pairs of lineages and computing their reproductive isolation I. We further introduce two indicators for the existence of reproductive barriers. A quantity named leading reproductive isolation I^* is defined as the 99th percentile of the reproductive isolation distribution. It signals that there is one percent of crosses with reproductive isolation equal or larger than I^*. We would also like to raise a caveat that $I^* > 0$ is sufficient for the existence of reproduction barriers but not a necessary condition, due to the possibility of positive I in the distribution even if $I^* \leq 0$. The leading reproductive isolation metric hence summarizes a high level of reproductive barriers that can be found among the lineages. On the other hand,

[5] We refer a population induced from a genetic pool to a sample among all possible populations that own the same underlying genetic pool.

the fraction of positivity in the reproductive isolation distribution serves as a necessity indicator for reproductive barriers, which we denote as f_p. The zero-value of f_p implies that none of the crosses generate inviable hybrids more than the anticipation of the offspring and thus the absence of reproductive barriers. Contrarily, a positive f_p does not satisfy existence of barriers considering small reproductive isolation subject to noise. These two indicators are beneficial for us to identify the responsible part of the model to the observed evolutionary consequences.

Potential Incompatibilities within and between Genetic Pools

An intra-lineage incompatibility is a group of alleles in its genetic pool, each of a unique locus, that generates a lethal pathway. In our model those incompatibilities are the only source of inviability, and hence the number of potential incompatibilities provides information about reproductive barriers. Nevertheless, counting the number of potential incompatibilities within a genetic pool through a brute-force manner is computationally intractable. Here we suggest a relatively efficient algorithm when the total number of loci is small. Our strategy is to turn the task into solving a graph problem. The genetic pool can be transformed to an edge-colored network where nodes once more represent possibly existing proteins in the organism. The edges correspond to available alleles in the pool, which are colored by their according loci. A potential incompatibility then becomes a simple path from an environmental input signal to a lethal protein node, with an additional constrain that no edges on the path have the same color. We call such a path an edge-colorful simple path (ECSP).

The proposed algorithm, as demonstrated in ***Appendix 3*** Algorithm 1, counts the number of ECSPs from the source nodes to the targets nodes by having agents propagate on the edge-colored network iteratively. An agents is capable of keeping information of the trajectory, including its current position on the network, the colors of edges it has traversed and the nodes that it has visited[6]. Initially we deploy one agent on each source node. At every iteration, each agent is substituted by all of its possible successors who are a hop away, such that the hop along with the agent's memory obeys an edge-colorful simple path. Those successors can be deduced from the agent's trajectory information as shown in ***Appendix 3*** Algorithm 2. The cautiously-designed rule of agent propagation guarantees that the total number of agents locating on the target nodes at the nth iteration equals to the number of the desired ECSPs of length n. Moreover, since the order of an potential incompatibility is bounded above by the number of genes in the organism, iterations as many as the amount of edge colors in the network are sufficient to obtain a computationally feasible count of all potential incompatibilities. The efficiency of the algorithm can be further improved by, instead of keeping track of numerous agents, monitoring the distribution of agent states over iterations.

The same algorithm can be applied to count the number of inter-lineage incompatibilities as well. In this case the underlying genetic pools of both lineages are transformed into a single edge-colored network, whose edges then consist of alleles in the two pools and are again colored by their according loci. A ECSP on this composite network either only traverses through edges from one of the genetic pools, or it contains alleles from the two different pools. These two scenarios correspond to a incompatibility within and between genetic pools respectively. Therefore, by counting the number of ECSPs on the composite network, and subtracting by the number of potential incompatibilities within the two genetic pools separately, we can compute the number of incompatibilities between the two underlying genetic pools.

References

Barton NH. The role of hybridization in evolution. Molecular ecology. 2001; 10(3):551–568.

Bateson W. Heredity and variation in modern lights. Darwin and modern science. 1909; .

[6] In Algorithm 1, the NEW-AGENT procedure creates an agent instance given its position, visited colors and nodes accordingly. This trajectory information is also accessible fields of the agent instance.

Bikard D, Patel D, Le Metté C, Giorgi V, Camilleri C, Bennett MJ, Loudet O. Divergent evolution of duplicate genes leads to genetic incompatibilities within A. thaliana. Science. 2009; 323(5914):623–626.

Boyle EA, Li YI, Pritchard JK. An Expanded View of Complex Traits: From Polygenic to Omnigenic. Cell. 2017; 169(7):1177 – 1186. http://www.sciencedirect.com/science/article/pii/S0092867417306293, doi: https://doi.org/10.1016/j.cell.2017.05.038.

Brideau NJ, Flores HA, Wang J, Maheshwari S, Wang X, Barbash DA. Two Dobzhansky-Muller genes interact to cause hybrid lethality in Drosophila. science. 2006; 314(5803):1292–1295.

Chae E, Bomblies K, Kim ST, Karelina D, Zaidem M, Ossowski S, Martín-Pizarro C, Laitinen RE, Rowan B, Tenenboim H, Lechner S, Demar M, Habring-Müller A, Lanz C, Rätsch G, Weigel D. Species-wide Genetic Incompatibility Analysis Identifies Immune Genes as Hot Spots of Deleterious Epistasis. Cell. 2014; 159(6):1341 – 1351. http://www.sciencedirect.com/science/article/pii/S0092867414013762, doi: https://doi.org/10.1016/j.cell.2014.10.049.

Coyne JA, Allen Orr H. The evolutionary genetics of speciation. Philosophical Transactions of the Royal Society of London Series B: Biological Sciences. 1998; 353(1366):287–305.

Cutter AD. The polymorphic prelude to Bateson–Dobzhansky–Muller incompatibilities. Trends in ecology & evolution. 2012; 27(4):209–218.

Davidich M, Bornholdt S. The transition from differential equations to Boolean networks: A case study in simplifying a regulatory network model. Journal of Theoretical Biology. 2008; 255(3):269 – 277. http://www.sciencedirect.com/science/article/pii/S0022519308003652, doi: https://doi.org/10.1016/j.jtbi.2008.07.020.

Davies B, Hatton E, Altemose N, Hussin JG, Pratto F, Zhang G, Hinch AG, Moralli D, Biggs D, Diaz R, et al. Re-engineering the zinc fingers of PRDM9 reverses hybrid sterility in mice. Nature. 2016; 530(7589):171–176.

De Bruijn NG. A combinatorial problem. In: *Proc. Koninklijke Nederlandse Academie van Wetenschappen*, vol. 49; 1946. p. 758–764.

Dobzhansky T. Studies on hybrid sterility. II. Localization of sterility factors in Drosophila pseudoobscura hybrids. Genetics. 1936; 21(2):113.

Duranton M, Allal F, Valière S, Bouchez O, Bonhomme F, Gagnaire PA. The contribution of ancient admixture to reproductive isolation between European sea bass lineages. BioRxiv. 2019; p. 641829.

Gavrilets S. Evolution and speciation on holey adaptive landscapes. Trends in ecology & evolution. 1997; 12(8):307–312.

Gavrilets S, Gravner J. Percolation on the fitness hypercube and the evolution of reproductive isolation. Journal of theoretical biology. 1997; 184(1):51–64.

Gould BA, Chen Y, Lowry DB. Gene regulatory divergence between locally adapted ecotypes in their native habitats. Molecular Ecology. 2018; 0(0). https://onlinelibrary.wiley.com/doi/abs/10.1111/mec.14852, doi: 10.1111/mec.14852.

Guerrero RF, Hahn MW. Speciation as a sieve for ancestral polymorphism. Molecular ecology. 2017; 26(20):5362–5368.

Johnson NA, Porter AH. Rapid speciation via parallel, directional selection on regulatory genetic pathways. Journal of Theoretical Biology. 2000; 205(4):527–542.

Kalirad A, Azevedo RBR. Spiraling Complexity: A Test of the Snowball Effect in a Computational Model of RNA Folding. Genetics. 2017; 206(1):377–388. http://www.genetics.org/content/206/1/377, doi: 10.1534/genetics.116.196030.

Kuzmin E, VanderSluis B, Wang W, Tan G, Deshpande R, Chen Y, Usaj M, Balint A, Mattiazzi Usaj M, van Leeuwen J, Koch EN, Pons C, Dagilis AJ, Pryszlak M, Wang JZY, Hanchard J, Riggi M, Xu K, Heydari H, San Luis BJ, et al. Systematic analysis of complex genetic interactions. Science. 2018; 360(6386). http://science.sciencemag.org/content/360/6386/eaao1729, doi: 10.1126/science.aao1729.

Langfelder P, Horvath S. WGCNA: an R package for weighted correlation network analysis. BMC bioinformatics. 2008; 9(1):559.

Láruson ÁJ, Yeaman S, Lotterhos KE. The Importance of Genetic Redundancy in Evolution. Trends in Ecology & Evolution. 2020; .

Livingstone K, Olofsson P, Cochran G, Dagilis A, MacPherson K, Seitz KA. A stochastic model for the development of Bateson–Dobzhansky–Muller incompatibilities that incorporates protein interaction networks. Mathematical Biosciences. 2012; 238(1):49 – 53. http://www.sciencedirect.com/science/article/pii/S0025556412000491, doi: https://doi.org/10.1016/j.mbs.2012.03.006.

Marques DA, Meier JI, Seehausen O. A combinatorial view on speciation and adaptive radiation. Trends in ecology & evolution. 2019; .

Meier JI, Marques DA, Mwaiko S, Wagner CE, Excoffier L, Seehausen O. Ancient hybridization fuels rapid cichlid fish adaptive radiations. Nature communications. 2017; 8:14363.

Muller H. Isolating mechanisms, evolution, and temperature. In: *Biol. Symp.*, vol. 6; 1942. p. 71–125.

Nelson TC, Cresko WA. Ancient genomic variation underlies repeated ecological adaptation in young stickleback populations. Evolution Letters. 2018; 2(1):9–21.

Nowak MA, Boerlijst MC, Cooke J, Smith JM. Evolution of genetic redundancy. Nature. 1997; 388(6638):167–171.

Orr HA. The population genetics of speciation: the evolution of hybrid incompatibilities. Genetics. 1995; 139(4):1805–1813. http://www.genetics.org/content/139/4/1805.

Palmer ME, Feldman MW. DYNAMICS OF HYBRID INCOMPATIBILITY IN GENE NETWORKS IN A CONSTANT ENVIRONMENT. Evolution. 2009; 63(2):418–431. https://onlinelibrary.wiley.com/doi/abs/10.1111/j.1558-5646.2008.00577.x, doi: 10.1111/j.1558-5646.2008.00577.x.

Powell DL, García-Olazábal M, Keegan M, Reilly P, Du K, Díaz-Loyo AP, Banerjee S, Blakkan D, Reich D, Andolfatto P, et al. Natural hybridization reveals incompatible alleles that cause melanoma in swordtail fish. Science. 2020; 368(6492):731–736.

Rieseberg LH, Sinervo B, Linder CR, Ungerer MC, Arias DM. Role of gene interactions in hybrid speciation: evidence from ancient and experimental hybrids. Science. 1996; 272(5262):741–745.

Satokangas I, Martin S, Helanterä H, Saramäki J, Kulmuni J. Multi-locus interactions and the build-up of reproductive isolation. arXiv preprint arXiv:200513790. 2020; .

Schiffman JS, Ralph PL. System drift and speciation. bioRxiv. 2018; https://www.biorxiv.org/content/early/2018/01/26/231209, doi: 10.1101/231209.

Schlitt T, Brazma A. Current approaches to gene regulatory network modelling. BMC bioinformatics. 2007; 8(S6):S9.

Schluter D. Evidence for Ecological Speciation and Its Alternative. Science. 2009; 323(5915):737–741. http://science.sciencemag.org/content/323/5915/737, doi: 10.1126/science.1160006.

Sicard A, Kappel C, Josephs EB, Lee YW, Marona C, Stinchcombe JR, Wright SI, Lenhard M. Divergent sorting of a balanced ancestral polymorphism underlies the establishment of gene-flow barriers in Capsella. Nature communications. 2015; 6:7960.

Tong AHY, Lesage G, Bader GD, Ding H, Xu H, Xin X, Young J, Berriz GF, Brost RL, Chang M, et al. Global mapping of the yeast genetic interaction network. science. 2004; 303(5659):808–813.

True JR, Haag ES. Developmental system drift and flexibility in evolutionary trajectories. Evolution & development. 2001; 3(2):109–119.

Turner LM, White MA, Tautz D, Payseur BA. Genomic Networks of Hybrid Sterility. PLOS Genetics. 2014 02; 10(2):1–23. https://doi.org/10.1371/journal.pgen.1004162, doi: 10.1371/journal.pgen.1004162.

Wagner A, Wright J. Alternative routes and mutational robustness in complex regulatory networks. Biosystems. 2007; 88(1-2):163–172.

Wang B, Mojica JP, Perera N, Lee CR, Lovell JT, Sharma A, Adam C, Lipzen A, Barry K, Rokhsar DS, et al. Ancient polymorphisms contribute to genome-wide variation by long-term balancing selection and divergent sorting in Boechera stricta. Genome biology. 2019; 20(1):126.

Wittbrodt J, Adam D, Malitschek B, Mäueler W, Raulf F, Telling A, Robertson SM, Schartl M. Novel putative receptor tyrosine kinase encoded by the melanoma-inducing Tu locus in Xiphophorus. Nature. 1989; 341(6241):415–421.

Yamamoto E, Takashi T, Morinaka Y, Lin S, Wu J, Matsumoto T, Kitano H, Matsuoka M, Ashikari M. Gain of deleterious function causes an autoimmune response and Bateson–Dobzhansky–Muller incompatibility in rice. Molecular Genetics and Genomics. 2010; 283(4):305–315.

Appendix 1

Hybrid Inviability against a Single Incompatibility

Here we analytically evaluate the probability that a hybrid is inviable presuming that multiple incompatibilities are rarely embedded in two parental gene regulatory networks. In addition, this naive analysis explains the pattern of RI distribution, ***Figure 5***a in the main text.

Assume that there is only on incompatibility $\mathcal{I}$ between the two parental gene networks G_1 and G_2. For convenience we suppose there are an even number of loci in the organisms, denoted by $2m$, and let the incompatibility $\mathcal{I}$ be of order $k-1$ so it consists of k alleles to form a lethal combination. We also suppose that, among the k alleles in $\mathcal{I}$, k_1 of them come from G_1 and the other k_2 alleles are from G_2.

Following the rule of recombination between haploid GRNs in our model, the hybrid is generated by randomly segregating alleles of m loci from G_1 and then mixing with alleles of the other m loci from G_2. Hence if $m < k_1$ or $m < k_2$, then there is no chance that the incompatibility $\mathcal{I}$ appears in the hybrid. Otherwise, among all plausible segregation, we can compute the number of achievable ways that the k_1 and k_2 alleles from G_1 and G_2 respectively are sorted into the hybrid. The probability that the hybrid is inviable due to the only incompatibility $\mathcal{I}$ is thus

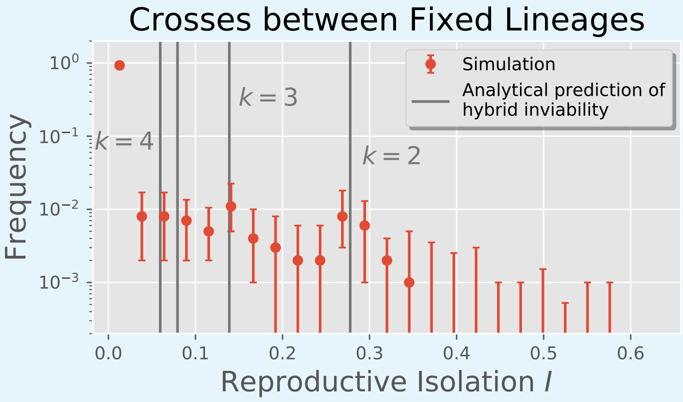
$$P(\mathcal{I}) = \begin{cases} \frac{\binom{2m-k}{m-k_1}}{\binom{2m}{m}}, & \text{if } k_1, k_2 \leq m \\ 0, & \text{otherwise} \end{cases} \tag{2}$$

If we further assume that $m \gg 1$ and $m \gg k$, applying the Stirling's approximation we have an estimate of the hybrid inviability

$$P(\mathcal{I}) = \frac{m!m!(2m-k)!}{(m-k_1)!(m-k_2)!(2m)!} \approx 2^{-k} \tag{3}$$

This plain derivation shows that, should there be only one incompatibility concealing between two parental GRNs, the survivability of a hybrid is predominantly determined by the order of the incompatibility.

Here ***Figure 1*** shows good agreement between our analytical prediction of hybrid inviability and the “bulges" from the observed RI distribution. Our simple derivation explains the higher likelihood of certain RI levels relative to their neighboring regions. It also manifests how the discreteness nature of hybrid incompatibilities shapes the RI distribution and that this characteristic has major effects on the strength of reproductive barriers.



Appendix 1 Figure 1. Comparison between the uncovered RI distribution in our simulations and the predicted hybrid inviability ***Equation 2***.

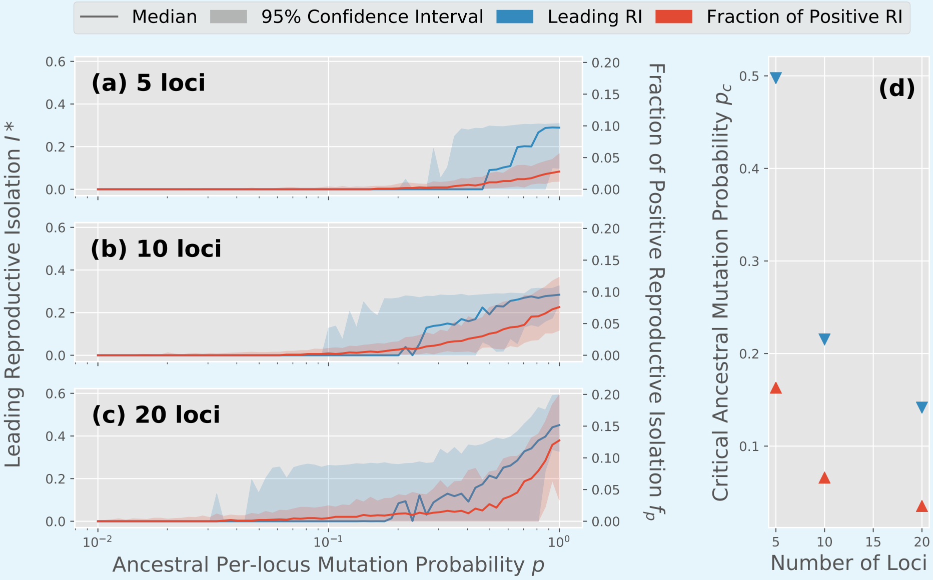
Appendix 2

Reproductive Barriers and Ancestral Genetic Variation

Here we demonstrate our examination on how the extent of ancestral genetic variation influences the appearance and strength of reproductive barriers. To begin with, we designed a pipeline to produce ancestral populations whose amount of genetic variation are tunable. A fixed population was first obtained from our GRN evolution model starting with randomly generated individual GRNs. For every locus, the allele might then mutate into any other possible allele with a per-locus mutation probability p. The resulting population was regarded as the ancestral population, where the mutation probability p became a tunable parameter to assess the degree of ancestral variation.

We followed the same methodology to simulate generational dynamics of GRNs and to compute reproductive isolation between allopatric lineages as in the main text. ***Figure 1***a-c shows, for different number of loci, the reproductive barriers consequent to the varying ancestral mutation probability p. Here we present two indicators of barriers: the leading RI (blue, left axis) and the fraction of positive RI (red, right axis). On a first glance the simulations evince that, for a organism with a larger number of loci, emergence of barriers only required a smaller ancestral mutation probability yet more apparent barriers were observed.

Figure 1a-c furthermore suggest some critical level of ancestral variation associated with the constant population size, such that reproductive barriers would hardly appear between lineages evolving from an ancestral population with less polymorphisms. We quantify the critical level of genetic variation through a critical mutation probability p_c; this is the smallest ancestral mutation probability with which a barrier indicator has non-zero median value. Nevertheless, due to the lack of a both sufficient and necessary indicator, we could only estimate the interval that this critical level fell into. The critical level of ancestral variation would be bounded above by p_c for the leading RI (a sufficient indicator of barriers) and bounded below by one for the fraction of positive RI (a necessary indicator of barriers). ***Figure 1***d presents the interval estimation that the critical ancestral variation fell into for organisms with different number of loci.



Appendix 2 Figure 1. Varying the extent ancestral variation and its corresponding strength of reproductive barriers. The GRN evolution was simulated under Setup 1 described in Methods. **(a-c)** Indicators of barriers for 5, 10 and 20 loci. **(d)** Estimation of their critical level of ancestral variation.

Appendix 3

Algorithms of Counting Potential Incompatibilities

Algorithm 1 COUNT-ECSP

Require: A set of source nodes S; a set of target nodes T; a map I from nodes to their incident outgoing edges; a set of path lengths of interests L.

Ensure: A map C from L to the number of edge-colorful simple paths from S to T, which are of the corresponding length.

```
C ← an empty map
l_max ← the largest element of L
A ← an empty list                                          ▷ Initialize agents.
for all node $s \in S$ do $A$.INSERT(NEW-AGENT($s$, $\emptyset$, $\{s\}$))
end for
for $l \leftarrow 1$ to $l_{max}$ do        ▷ Iterate over the number of hops agents have made from the source nodes.
    $n \leftarrow 0$
    $N \leftarrow$ an empty list                           ▷ Update the list of agents.
    for all agent $a \in A$ do
        for all agent $a' \in$ NEXT-POSSIBILITIES($a$, $I$) do
            $N$.INSERT($a'$)
            if $a'.position \in T$ then $n \leftarrow n + 1$
            end if
        end for
    end for
    $A \leftarrow N$
    if $l \in L$ then $C$.INSERT($l$, $n$)                 ▷ Update counting.
    end if
end for
return $C$
```

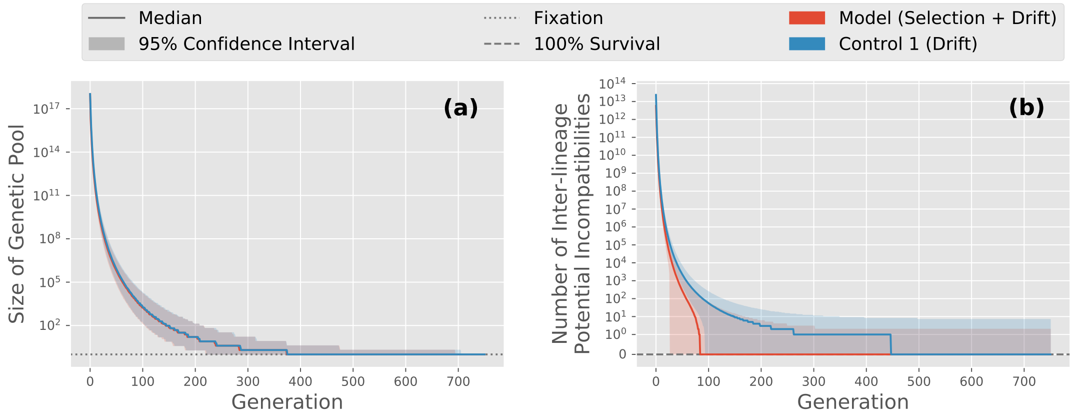
Algorithm 2 NEXT-POSSIBILITIES

Require: An agent a; a map I from nodes to their incident outgoing edges.

Ensure: A set P of agents who are of all the possible states that can be reached through a hop from the given agent a, such that

1. The hop only goes through an edge of a color that has not been visited by the agent.
2. The position after the hop has not been visited by the agent.

```
$P \leftarrow$ an empty set
for all edge $e \in I$.GET($a$) do
    if $e.color \notin a.colors\text{-}visited$ and $e.target \notin a.nodes\text{-}visited$ then
        $a' \leftarrow$ NEW-AGENT($e.target$, $a.colors\text{-}visited \cup \{e.color\}$, $a.nodes\text{-}visited \cup \{e.target\}$)
        $P$.INSERT($a'$)
    end if
end for
return $P$
```



791

Figure 7–Figure supplement 1. (a) The size the underlying genetic pool continually shrank until there was only one accessible genotype. At this stage a population fixated a single GRN, and no significant difference was found between the model and the control scenario without selection, i.e., drift only. **(b)** In our model, inter-lineage incompatibilities persisted throughout evolution (red), which accounts for the sustained confidence interval of their abundance even after populations reach fixation. Interestingly, in the control scenario where natural selection was silenced, inter-lineage incompatibilities were eliminated at a slower pace. We hypothesize that due to the lack of guidance by selection, inter-lineage incompatibilities only became inaccessible through random genetic drift. This scenario led to fatal allelic combinations that were more persistent than those in the model and hence stronger reproductive barriers were observed.