A TEST OF THE SNOWBALL MODEL

A Dissertation Presented to

the Faculty of the Department of Biology & Biochemistry

University of Houston

In Partial Fulfillment

of the Requirements for the Degree

Doctor of Philosophy

By

Ata Kalirad

November 2016

A TEST OF THE SNOWBALL MODEL

Ata Kalirad

APPROVED:

Ricardo B. R. Azevedo, Chairman
Dept. of Biology & Biochemistry

Kevin E. Bassler
Dept. of Physics

Timothy Cooper
Dept. of Biology & Biochemistry

Elizabeth Ostrowski
Dept. of Biology & Biochemistry

Rebecca Zufall
Dept. of Biology & Biochemistry

Dean, College of Natural Sciences and Mathematics

Acknowledgements

My special thanks goes to Dr. Ricardo B. R. Azevedo, without whom this thesis and my love for Wagnerian music dramas would have been impossible. I would like to thank the rest of my thesis committee: Dr. Kevin E. Bassler, Dr. Timothy Cooper, Dr. Elizabeth Ostrowski, and Dr. Rebecca Zufall. Their continued efforts and support played a central role in the completion of this thesis.

I need to specially thank Dr. Tony Frankino for his help and encouragements. My sincere thanks also goes to my first PI, the inimitable Dr. Amy Sater.

I have to thank my colleagues in that most auspicious of places, Azevedo lab: Sunil Guharajan, Kedar Karkare, Grimaldo Elias Ureña, Bingjun Zhang, and Hao Zhang. I have to thank members of Cooper lab, especially Fen Peng, Yinhua Wang, and Dr. Andrea Wünsche for helping me with the valuable, though ultimately futile, cell fusion assay.

I have to thanks the staff of the department of Biology & Biochemistry for their helpfulness. Special thank goes to: Yonia Pulido, Mallory Travis, and Rosezelia Jackson.

Many thanks to friends and colleagues, past and present, in the division of Ecology & Evolution. I am resisting to name them all to avoid the possible embarrassment of leaving out someone by mistake.

A TEST OF THE SNOWBALL MODEL

An Abstract of a Dissertation

Presented to

the Faculty of the Department of Biology & Biochemistry

University of Houston

In Partial Fulfillment

of the Requirements for the Degree

Doctor of Philosophy

By

Ata Kalirad

November 2016

Abstract

Genetic incompatibilities can emerge as a by-product of genetic divergence. According to Dobzhansky and Muller, alleles at different loci that have fixed in different genetic backgrounds may be incompatible when brought together in a hybrid. Orr showed that the number of Dobzhansky–Muller incompatibilities (DMIs) should accumulate faster than linearly—i.e., snowball—as two lineages diverge. Several studies have attempted to test the snowball model using data from natural populations. One limitation of these studies is that they have focused on predictions of the snowball model but not on its underlying assumptions. Here I use a computational model of RNA folding to test both predictions and assumptions of the snowball model. In this model, two populations are allowed to evolve in allopatry on a holey fitness landscape. I find that the number of DMIs involving pairs of loci (i.e., simple DMIs) does not snowball—rather, it increases approximately linearly with divergence. I show that the probability of emergence of a simple DMI is approximately constant, as assumed by the snowball model. However, simple DMIs can disappear after they have arisen, contrary to the assumptions of the snowball model. This occurs because simple DMIs become complex (i.e., involve alleles at three or more loci) as a result of later substitutions. I introduce a modified snowball model—the melting snowball model—where simple DMIs can become complex after they appear. The melting snowball model can account for the results of the RNA-folding model. I also find that complex DMIs are common and, unlike simple ones, do snowball.

Reproductive isolation, however, does not snowball because DMIs do not act independently of each other. I also test the snowball model at the population level using an individual-based model. Using this model, I show that recombination rate, gene flow, and ancestral polymorphism can slow down the snowballing of incompatibilities between diverging populations. These factors result in selection for mutationally robust genotypes, and genotypes that are more resistant to mutations are also more resistant to introgressions, which reduces the number of DMIs.

Contents

1 Incompatibilities: A Guide for the Perplexed **1**

1.1 Where do species come from? 2

1.2 The varieties of prezygotic isolation 5

1.3 The causes of postzygotic isolation 6

1.4 The Dobzhansky-Muller model of genetic incompatibilities 8

1.5 The snowball model 10

1.6 RNA-folding model 12

1.7 Evolving on a holey fitness landscape 14

1.8 Summary 17

2 Spiraling complexity: a test of the snowball effect **18**

2.1 The quest to validate the snowball model 19

2.2 Methods 22

2.2.1 Simulating the snowball using RNA folding 22

2.2.2 How to find DMIs 24

2.2.3 Proportion of single introgressions involved in a DMI 29

2.2.4 DMI network 29

2.2.5 Reproductive isolation 30

2.2.6 “Holeyness” of the fitness landscape 31

2.2.7 Direct simulation of the snowball model 31

2.3 Results . . . 33
2.3.1 Simple DMIs do not snowball in the RNA-folding model . . 33
2.3.2 The probability that a DMI appears is approximately constant in the RNA-folding model . . . 34
2.3.3 Simple DMIs do not persist indefinitely in the RNA-folding model . . . 38
2.3.4 The RNA-folding simulations agree with the melting snowball model . . . 44
2.3.5 Complex incompatibilities snowball in the RNA-folding model 49
2.3.6 Reproductive isolation does not snowball in the RNA-folding model . . . 51
2.3.7 The fitness landscape influences the parameters of the melting snowball model . . . 56
2.3.8 $\mathcal{P}_1$ and Welch's paradox . . . 57

3 How do populations affect the accumulation of incompatibilities? 63

3.1 Segregating DMIs . . . 64
3.2 Methods . . . 67
3.2.1 The individual-based model . . . 67
3.2.2 How to find DMIs . . . 71
3.2.3 Ancestral polymorphism . . . 72
3.2.4 Gene flow . . . 72
3.2.5 Gene diversity . . . 72
3.2.6 Hamming distance . . . 73
3.2.7 Proportion of inviable sequences . . . 73
3.2.8 Reproductive isolation . . . 73
3.3 Results . . . 74
3.3.1 Incompatibilities snowball in the individual-based model . 74

3.3.2 Recombination slows down the snowballing of incompatibilities . . . 75
3.3.3 Recombination suppresses the emergence of segregating DMIs . . . 76
3.3.4 Ancestral polymorphism affects the accumulation of DMIs but not that of segregating DMIs . . . 78
3.3.5 Gene flow slows down the snowballing of incompatibilities but does not eliminate them entirely . . . 79
3.3.6 Higher mutation rates result in more robust genotypes . . . 80

4 Insights from studying speciation in a RNA world 93

4.1 What the RNA-folding model teaches us about DMIs . . . 94
4.1.1 The RNA-folding model supported the central prediction of the snowball model . . . 94
4.1.2 A possible resolution for Welch's paradox . . . 96
4.1.3 Complex DMIs are more abundant in the RNA-folding model . . . 97
4.1.4 How simple incompatibilities become complex . . . 98
4.2 How populations shape the accumulation of incompatibilities . . . 101
4.2.1 Selection for robustness can affect the accumulation of DMIs within and between populations . . . 101
4.2.2 Ancestral polymorphism acts against the snowballing of incompatibilities . . . 104
4.2.3 DMIs can persist in spite of gene flow . . . 105
4.3 Philosophical obstacles to a complete understanding of DMIs . . . 105

List of Figures

1.1 An example of a simple DM incompatibility between *D. melanogaster* and *D. simulans*. 9

1.2 The accumulation of a simple DMI follows from the divergence between two lineages. 11

1.3 RNA folding and base pair distance 15

1.3 Evolution on a holey fitness landscape. 17

2.1 Detecting DMIs. 25

2.2 Sequence evolution in a direct simulation of the snowball model showing the first $k = 4$ substitutions. 32

2.3 Fitting the snowball model to the RNA-folding model. 35

2.4 Distributions of the parameters of the snowball and linear models in the RNA-folding simulations 36

2.5 Evolution of the number of simple DMIs and Evolution of the probability, p_k. 38

2.6 A single substitution can dramatically rearrange the network of potential DMIs. 40

2.7 Jaccard index 42

2.8 The RNA-folding simulations agree with the melting snowball model. 44

2.9 Distributions of the parameters of the melting snowball model in the RNA-folding simulations. 45

2.10 The RNA-folding model behaves as expected under the melting snowball model. 48

2.11 Evolution of the probability, q, that a simple DMI becomes complex. 49
2.12 Complex DMIs snowball in the RNA-folding model. 50
2.13 The effect of multiple substitutions on the accumulation of DMIs . 52
2.14 Complex DMIs snowball in the RNA-folding model 53
2.15 Reproductive isolation does not snowball in the RNA-folding model. 54
2.16 RI does not reflect the number of DMIs 55
2.17 The fitness landscape influences the parameters of the melting snowball model. 59
2.18 Holeyness decreases with the value of α (A) and increases with the number of base pairs. 60
2.19 The accumulation of DMIs as a function of α 61
2.20 $\mathcal{P}_1$ 62

3.1 Segregating DMI in *C. elegans* 66
3.2 How to generate an ancestral population 69
3.3 Recombination at the population level 70
3.4 Incompatibilities snowball in the individual-based simulation. . . . 82
3.5 The effects of recombination on the individual-based model 84
3.6 Recombination affects the snowballing of incompatibilities (AIC) . 85
3.7 The recombination load 86
3.8 The effects of ancestral polymorphism on the individual-based simulation. 88
3.9 Ancestral polymorphism affects the snowballing of incompatibilities (AIC) 89
3.10 Gene flow can slow down the snowballing of incompatibilities but it does not eliminate them 90
3.11 Gene flow affects the snowballing of incompatibilities (AIC) 91
3.12 Mutational robustness as a function of mutation rate 92

List of Tables

1.1 **Isolating barriers** 4

2.1 **Properties of the 10^3 ancestors used in the simulations with $\alpha = 12$** 23

4.1 **The monomorphic model provides mixed support for the snowball model** 95

Chapter 1

Incompatibilities: A Guide for the Perplexed

> The view generally entertained by naturalists is that species, when intercrossed, have been specially endowed with the quality of sterility, in order to prevent the confusion of all organic forms.
>
> Darwin (1859, p. 246)

> It is quite possible to think of a world in which species do not exist but are replaced by a single reproductive community of individuals.
>
> Mayr (1942, p. 282)

1.1 Where do species come from?

To ask why there are species is to ask one of the most fundamental questions in evolutionary biology (Sherratt and Wilkinson 2009). While the very first step in answering this multifaceted question, i.e., to define species, has been vigorously discussed and debated over the years (Dobzhansky 1937; Mayr 1942; Sokal and Crovello 1970; van Valen 1976; Wiley 1978; Templeton 1980; Cracraft 1987), many have, either explicitly or implicitly, used the biological species concept, it being one of the more pragmatic definitions. The biological species concept turns the quandary of speciation into a much more modest, yet still challenging, question about the emergence of reproductive isolation (RI).[1] To presume an intimate association between species and RI dates back to the era before Darwin. Even the unknown author writing on *Espèce* in *Encyclopédie* define the species as "... nothing else than a constant succession of similar individuals that reproduce [among] themselves." (Diderot and d'Alembert 1751, vol. 5, p. 957)

Using RI as our yardstick, we can envision speciation as the gradual accumulation of divergent genetically based characteristics in different populations. Some of these divergent characteristics, known as reproductive isolating barriers,

[1]While Mayr is usually credited with introducing the concept of biological species concept, he himself cited Poulton (1908) and Jordan (1905), since they had already introduced this species concept (Mallet 2004).

decrease the level of interbreeding between populations (Table 1.1). As populations diverge, isolating barriers accumulate, and the level of RI among populations increases (Coyne and Orr 1989, 1997; Sasa *et al.* 1998; Edmands 2002; Fitzpatrick 2002; Presgraves 2002; Lijtmaer *et al.* 2003; Mendelson *et al.* 2004; Bolnick and Near 2005; Johnson 2006; Gourbière and Mallet 2010; Giraud and Gourbière 2012). Eventually RI reaches a point where two of these populations are considered distinct species.

For Darwin the presence of RI, in the form of inviable and/or sterile hybrids, posed a serious challenge: why would natural selection favor a trait as seemingly disadvantageous as hybrid inviability/sterility? While Darwin has on many occasions, been accused of bungling his attempt to address the question of speciation, thus failing to live up to the title of his magnum opus, not least by his defender par excellence, Huxley (1863, p. 148), a close reading of his chapter on hybrids in *On the origin of species* reveals his attempt at deciphering this mystery of mysteries.[2] But in spite of his best efforts coupled with his characteristic meticulousness, his description of hybrid sterility/inviability is a mishmash of ecological and genetic isolating barriers. The lack of a genetic understanding of RI meant that Darwin could not offer a more crystalline conclusion than "... the degree of difficulty in uniting two species, and the degree of sterility of their hybrid-offspring should generally correspond, though due to distinct causes; for both depend on the amount of difference of some kind between the species which

[2]Reznick (2010) even suggests that relating Darwin's writing on species to our current research program is foolhardy since Darwin did not distinguish between varieties and species as we do.

Table 1.1: Isolating barriers

Category
Prezygotic
Potential mates live in the same place, but do not meet either due to habitat isolation or temporal isolation.
Behavioral isolation (potential mates meet, but choose not to mate.)
Mechanical isolation (mating is not possible due to morphological differences.)
Gametic isolation (male gametes are transferred, but egg is not fertilized.)
Postzygoric
Zygote dies early in embryogenesis.
F_1 hybrids are inviable or sterile.
F_2 hybrids are inviable or sterile.

After Dobzhansky (1937); Coyne and Orr (2004); Johnson (2006); Barton *et al.* (2007); Ptacek and Hankison (2009).

are crossed" [p. 278]. Although this conclusion is as good an explanation as one could hope for in the 19th century (Sloan 2008), Darwin's explanation does not provide a mechanism that could explain the occurrence of hybrid inviability and/or sterility in nature.

The first step in resolving the dilemma of hybrid inviability and/or sterility is to disentangle the different isolating barriers that result in the defective hybrids (Table 1.1).

1.2 The varieties of prezygotic isolation

The prezygotic isolating barriers include an array of different barriers, all acting before the formation of the zygote. If two different species inhabit two different environments, then the probability that a member of one species would stumble upon someone form the other species is greatly diminished. Genetic divergence can cause such **habitat isolation** by making a species inept at functioning in the habitat occupied by another species. Competition can also result in "habitat balkanization" (Coyne and Orr 2004, p. 182). Wang *et al.* (1997) shows this habitat isolation in two subspecies of *Artemisia tridentata*, where each subspecies is more adapted to its native habitat.

Breeding at different times can result in **temporal isolation**. This type of isolation can be caused by different responses to an event in the environment such as tides or it can be identical responses to different events. Habitat differentiation can also indirectly cause temporal isolation (Coyne and Orr 2004, p. 206).

Behavioral isolation can occur in animals and hinges on behavioral cues and mismatch between behavioral signals can prevent mating and prevent gene flow between species (Kaneshiro 1980).

One of the more obvious manifestations of prezygotic isolation is **mechanical**. In its simplest form, mechanical isolation stems from male and female genitalia of two different species not fitting in one another. In damselflies (Coenagrionidae family), a male grasps at a female's thorax with its abdominal appendages. In several species of damselfly, interspecific mating fails because a male cannot

secure a firm grip on the female's thorax (Paulson 1974).

It is possible for isolating barriers to happen after mating but before the formation of the zygote. Historically, **gametic isolation** had been neglected since it is difficult to study, but in recent decades this mode of isolation has been the subject of several studies (Coyne and Orr 2004, p. 232).

1.3 The causes of postzygotic isolation

A number of genetic mechanisms can result in postzygotic isolation (Maheshwari and Barbash 2011):

> **Chromosomal rearrangements.** Mating between parents that differ in their karyotypes can result in hybrids with aneuploidy, inversions, or meiotic defects. For instance the infertility of mule (♀ horse × ♂ donkey) and hinny (♀ donkey × ♂ horse) is caused by aneuploidy since both mule and hinny have $2n = 63$ whereas their parents, horse and donkey, have $2n = 64$ and $2n = 62$, respectively (Allen and Short 1997).[3]
>
> **Haploinsufficiency.** Hybrids that lack copies of necessary genes are doomed to sterility or inviability. Masly *et al.* (2006) show that crossing *Drosophila melanogaster* (*mel*) females with *D. simulans* (*sim*) males can result in hybrids that lack a single copy of *JYAlpha* gene. This gene, which encodes the alpha

[3]For a comprehensive look at the history of the chromosomal speciation theory and its different manifestations in nature see Brown and O'Neill (2010).

subunit of Na^+/K^-ATPase, resides on the fourth chromosome of *mel* but in *sim* it is found on the third chromosome.This transposition means some of the hybrids carry just a single copy of *JYAlpha*, causing them to become sterile.

Sequence Divergence. The mere existence of molecular differences at the DNA level between the parents can potentially disrupt crossing-over during meiosis and cause hybrid incompatibility. Two species of *Saccharomyces* (*S. cerevisiae* and *S. paradoxus*) have been used to demonstrate that mere sequence divergence can, in fact, cause RI because the mismatch repair system, which involves resolving heteroduplex structures during meiosis, cannot operate fully if the two parental sequences are greatly divergent (Greig *et al.* 2003).

Transposable Elements. There are multiple instances of transposable elements or noncoding repeats causing RI (Michalak 2008). The cross between two species of wallabies, *Macropus eugenii* and *Wallabia bicolor*, can activate hitherto dormant transposable elements in the hybrids (O'Neill *et al.* 1998).

Dosage Imbalances. Josefsson *et al.* (2006) used *Arabidopsis thaliana* and *A. arenosa* to show that only carrying certain proportions of parental genomes results in functional hybrids. Josefsson *et al.* (2006) suggest that the cause of the hybrid incompatibility might be related to insufficient amount of maternal or paternal regulatory elements needed to compensate for the excess or scarcity of certain genes in the hybrids.

The last cause of hybrid incompatibility, commonly referred to as **the Dobzhansky-Muller (DM) model** is at the center of this thesis and thus demands to be explained fully.

1.4 The Dobzhansky-Muller model of genetic incompatibilities

The inception of the Dobzhansky-Muller (DM) model of genetic incompatibilities and its relation to speciation can be traced back to Dobzhansky (1937) and his seminal work, *Genetics and the origin of species* (Orr 1996; Gavrilets 2004; Johnson 2009). Lewontin (1974) described the study of the genetics of speciation as a "methodological impossibility", given the difficulties involved in studying the hybrids. But a discovery in 1922 made it feasible for the fly geneticists to circumvent this "methodological impossibility". Donald Lancefield realized that crossing *Drosophila persimilis* with *Drosophila pseudoobscura* yields fertile female hybrids, in addition to the sterile male hybrids. This discovery was the panacea geneticists interested in studying speciation so desperately needed (Orr 1996). Dobzhansky's insight into the genetics of speciation, later bolstered by Muller (1942), can be traced to Lancefield's breakthrough.[4]

[4]Orr (1996) argues that Bateson deserves credit for the DM model, as he suggested the possibility that "complementary factors" between species can result in the hybrid sterility. Conversely, Forsdyke (2011), who is one of the authors of a comprehensive biography on William Bateson (Cock and Forsdyke 2008), suggests that a careful reading of Bateson reveals the non-genic nature of his ideas that cannot be equated with the type of epistatic interaction the DM model is based upon.

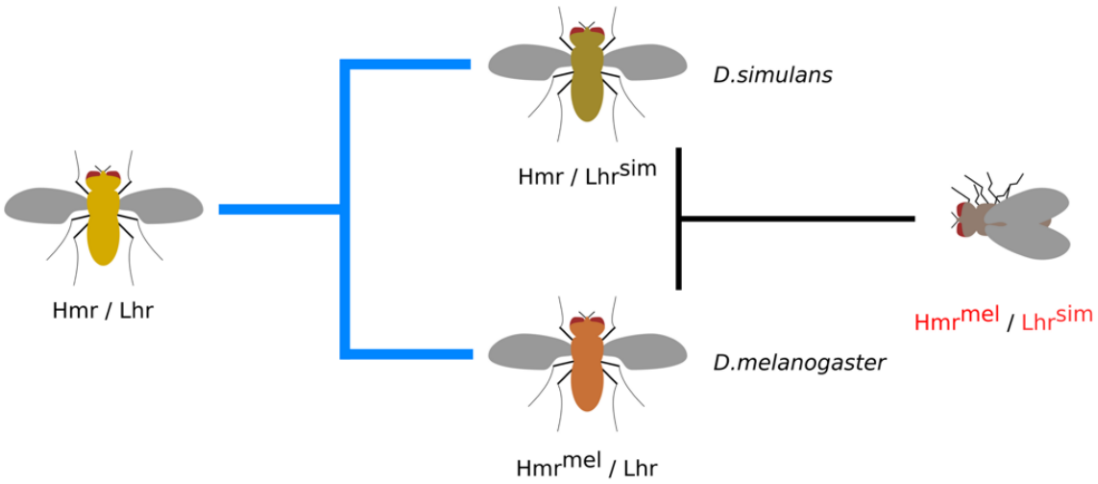

Figure 1.1: An example of a simple DM incompatibility (DMI) between *D. melanogaster* and *D. simulans*. The two derived versions of ancestral alleles , *Lhr* and *Hmr*, which are located at different loci,are functional in *D. simulans* and *D. melanogaster* respectively, but are deleterious when brought together in the hybrid.

The elegance of the DM model lies in its simplicity. In figure 1.1, two derived versions of ancestral alleles arose in *Drosophila simulans* and *D. melanogaster*. While selection demands the derived version of *Lhr* to be compatible with the genetic background in which it arose, i.e., the *D. simulans* genome, there is no guarantee that the same mutations would be beneficial or neutral in a different genetic background, such as that of *D. melanogaster*. In this scenario, the two derived are indeed incompatible when brought together in the hybrid. *Lhr* and *Hmr* encode proteins that are involved in suppression of transposable elements and satellite DNAs. Hence hybrid lethality is caused by overexpression of transposable elements related to heterochromatin regulation (Satyaki *et al.* 2014).

The interest in the DM model over the years has resulted in a extensive catalog of genes that negatively interact with each other and thus result in hybrid

inviability or sterility (reviewed in Presgraves 2010b; Rieseberg and Blackman 2010; Maheshwari and Barbash 2011). Although the varieties of molecular symptoms that result from DM interactions are impressive, one has to wonder about the number of these incompatibilities between different lineages with varying degrees of divergence. It turns out that the DM model can provide us with testable predictions regarding this point.

1.5 The snowball model

Let us start with an ancestral sequence and allow two lineages starting from the same ancestor to accumulate mutations. While the first mutation cannot result in an incompatibility since it arose in the ancestral background and ought to be compatible with it, the second mutation can result in a pairwise incompatibility (Figure 1.2).

The genetic incompatibilities can be classified according to the number of loci involved in them. A **simple** DMI is an incompatibility between two different loci. A **complex** DMI arises when more than two loci participate in its formation where all the participating alleles are needed to cause the incompatibility. It is needless to say that complex incompatibilities come in varying orders, e.g., three-way, four-way, etc., but for the moment I will focus on simple incompatibilities.

The new substitution k can potentially be incompatible with alleles at $k-1$ diverged loci. Any of these potential incompatibilities can be a DMI according

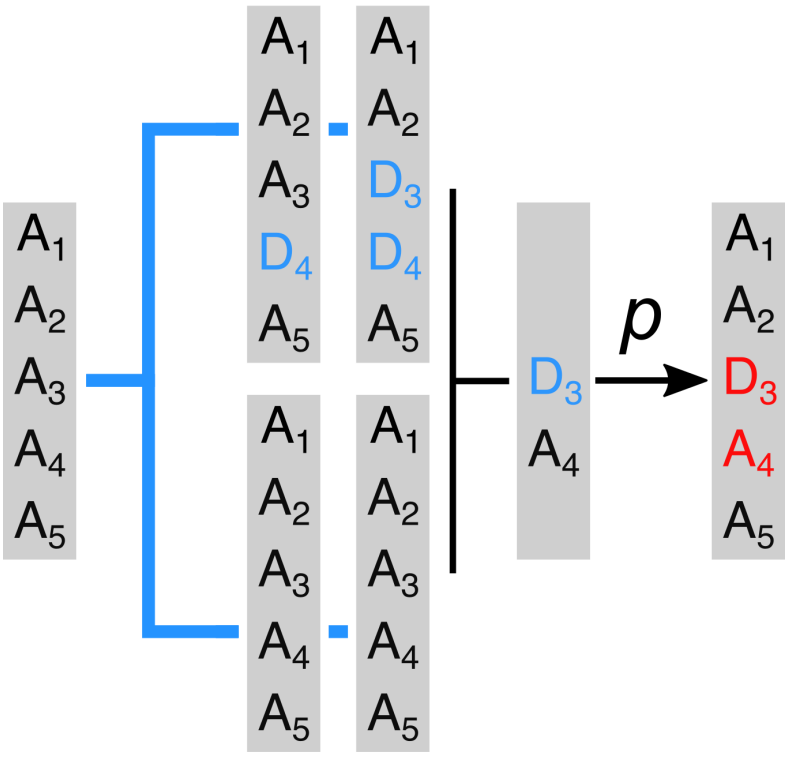

Figure 1.2: The accumulation of a simple DMI follows from the divergence between two lineages. The first mutation ($A_4 \rightarrow D_4$) cannot cause a genetic incompatibility since it arose in the ancestral background and has to be compatible with the ancestral alleles. However, the second mutation ($A_3 \rightarrow D_3$) arose in a background where A_4 did not exist. There is a probability p that the potential incompatibility between D_3 and A_4 does in fact result in a DMI.

to the probability p. So the total number of incompatibilities after k substitutions will be

$$I_k = I_{k-1} + (k-1)p \tag{1.1}$$

where I_{k-1} is the number of simple incompatibilities that accumulated after $k-1$ substitutions. Assuming $I_1 = 0$, the solution to Equation 1.1 is

$$I_k = \frac{k(k-1)p}{2} \quad . \tag{1.2}$$

Equation 1.2 predicts that the number of simple incompatibilities will accumulate faster than linearly as a function of divergence, a pattern Orr (1995) described as "snowballing." This prediction assumes that p remains constant as populations diverge. These simple incompatibilities can occur between an ancestral and a derived allele or between two derived alleles.

Complex incompatibilities are also expected to snowball but following different relationships from that in Equation 1.2: incompatibilities of order n are expected to accumulate at a rate approximately proportional to k^n (Orr 1995; Welch 2004).

1.6 RNA-folding model

RNA has certain properties that makes it suitable to study epistatic interactions: Firstly, there is a direct relationship between the genotype, i.e., the string of nucleotides, and the phenotype, i.e., the secondary structure of the RNA. Secondly,

computational algorithm can predict the secondary structure of a RNA sequence. The RNA-folding model has been used to study other evolutionary consequences of epistasis, including robustness (van Nimwegen *et al.* 1999; Ancel and Fontana 2000), evolvability (Wagner 2008; Draghi *et al.* 2010), and the rate of neutral substitution (Draghi *et al.* 2011).

Predicting the minimum free-energy (MFE) of a RNA sequence using a computationally efficient algorithm is possible if we assume that the effect of a base-pair on MFE is solely dependent on its nearest neighbors. The MFE structure then is simply the most probable structure given the combined MFE effects of its base-pairs (Zuker and Stiegler 1981; Do *et al.* 2006; Hamada *et al.* 2009; Bindewald *et al.* 2010; Sato *et al.* 2011; Swenson *et al.* 2012) (Figure 1.3A). But how reliable are these MFE predictions? Bernet and Elena (2015) introduce a set of single and double substitutions into 5′-UTR of Tobacco etch virus. They proceeded to measure the fitness effect of these mutations in a test tube. Bernet and Elena (2015) found that the fitness effects measured experimentally do correlate positively with the effects predicted by RNA-folding algorithm.[5]

I used the ViennaRNA package 2.1.9 (Lorenz *et al.* 2011) with default parameters to compute the minimum free-energy (MFE) secondary structure of each sequence. The similarity between two MFE structures can be quantified using the base-pair distance between the two structures. Base-pair distance is defined as the number of base-pairs required to convert one MFE structure into another

[5]Bernet and Elena (2015) utilized the RNAfold program from the ViennaRNA package version 1.6.4 and the LocARNA webserver

(Figure 1.3B). I used the base-pair distance to calculate the fitness of RNA sequence i using the step function:

$$w_i = \begin{cases} 1 & \text{if} \quad \beta_i > \alpha \quad \text{and} \quad \delta_i \leqslant \alpha \\ \\ 0 & \text{otherwise} \end{cases} \tag{1.3}$$

where β_i is the number of base pairs in the secondary structure of sequence i, δ_i is the base-pair distance between the structure of sequence i and the reference structure, and α is an arbitrary threshold. Unless otherwise stated we used $\alpha = 12$.

1.7 Evolving on a holey fitness landscape

The fitness function in Equation 1.3 specifies a neutral network (Schuster *et al.* 1994; van Nimwegen *et al.* 1999), a type of holey fitness landscape (Figure 1.3). A holey fitness landscape, first introduced by Gavrilets (2004), is an approximation of a rugged landscape where valleys are replaced by holes and the all the viable genotypes form ridges. As Gavrilets (2004) points out, speciation can occur on a holey landscape, where two species which started from a common ancestor are now separated by holes on the fitness landscape.

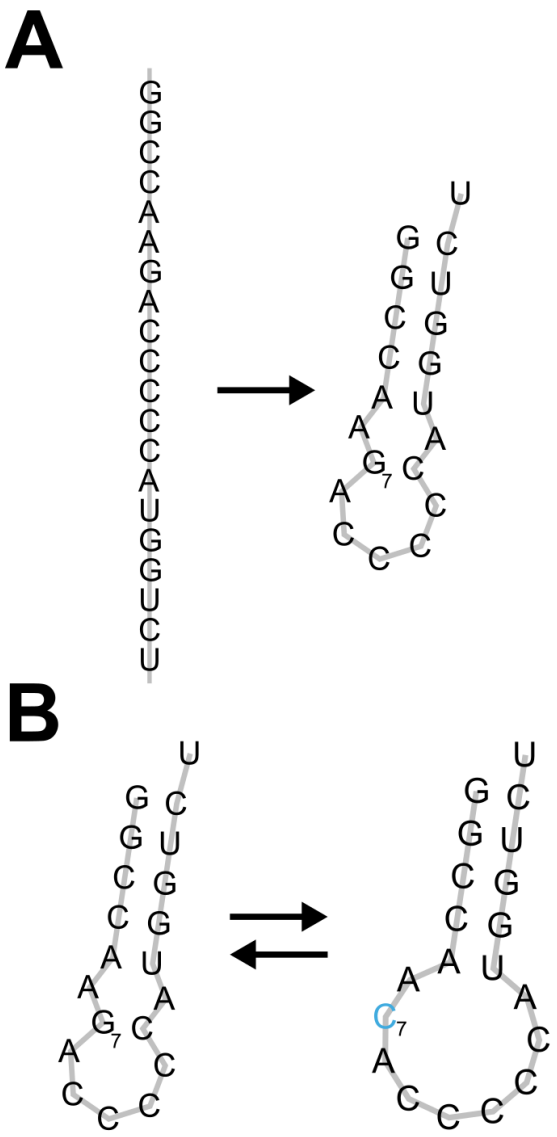

Figure 1.3: The secondary structure of a RNA sequence can be determined using a computationally efficient algorithm such as vienna RNA package used here (A). A single substitution ($G_7 \rightarrow C_7$) results in a MFE structure (right) that is one base-pair distance away from the original structure (B).

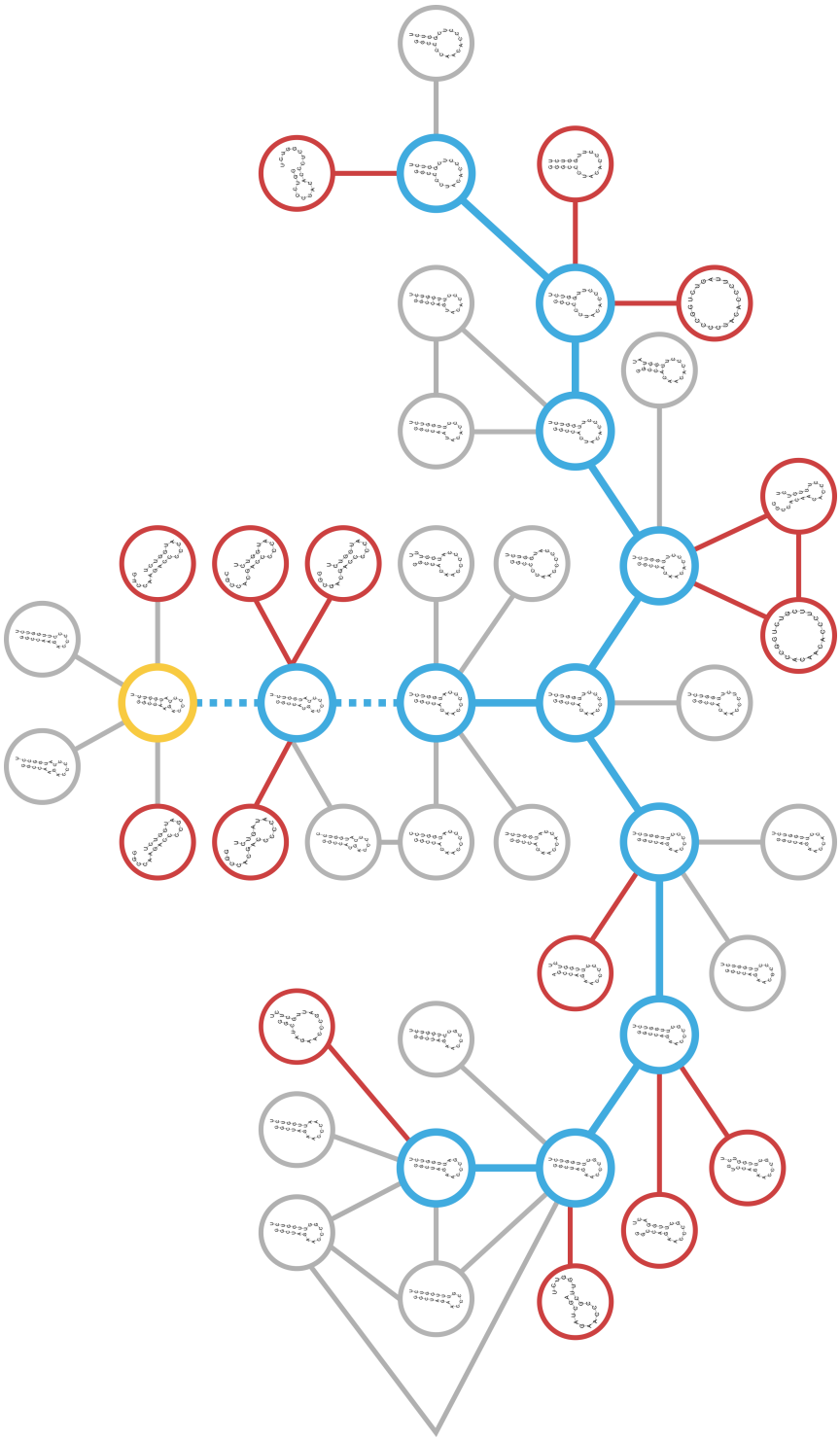

Figure 1.3 ***(previous page)*****:** Evolution on a holey fitness landscape. Mutational network of RNA sequences. Lines connect sequences of 20 nucleotides that can be reached by a single nucleotide substitution. Only a tiny fraction of the entire mutational network of $\sim 10^{12}$ sequences is shown. Furthermore, only a few of the 60 mutational neighbors of each sequence are shown. A sequence is viable (yellow, blue or gray circles) if its secondary structure both has more than $\alpha = 2$ base pairs and is at most $\alpha = 2$ base pairs away from the reference structure (yellow circle); a sequence is inviable otherwise (red circles) (Equation 1.3). Each simulation starts with a burn-in period where a sequence with the reference structure undergoes 3 neutral substitutions (dashed blue lines). After that, the resulting sequence is used as the ancestor of two lineages that alternately accumulate neutral substitutions until they have diverged at $k = 8$ sites (solid blue lines).

1.8 Summary

In the next chapter I attempt to test the assumptions of the snowball model. Using the RNA-folding model described above, combined with an exhaustive introgression approach, I test if the number of simple DMIs in fact snowballs.

In the third chapter I investigate the effects of recombination, migration, and polymorphism on the patterns of accumulation of genetic incompatibilities.

The final chapter is an attempt to make sense of the results from the monmorphic model and the population level model, and to juxtapose them with what is known from nature.

Chapter 2

Spiraling complexity: a test of the snowball effect

> To expect that the intricacies of science will be pierced by a careless glance, or the eminences of fame ascended without labour, is to expect a particular privilege, a power denied to the rest of mankind; but to suppose that the maze is inscrutable to diligence, or the heights inaccessible to perseverance, is to submit tamely to the tyranny of fancy, and enchain the mind in voluntary shackles.
>
> Samuel Johnson, *The Rambler*, 9 July 1751

The snowball model has been the center of several studies since its conception. In this chapter, I will expand some of the famous studies which attempted to corroborate the snowball model and then use the RNA-folding model to test the assumptions of the snowball model.

2.1 The quest to validate the snowball model

Several studies have attempted to test the snowball model. They have employed three different approaches. First, using postzygotic RI as a proxy for the number of DMIs. For example, Larcombe *et al.* (2015) measured the strength of hybrid incompatibility between *Eucalyptus globulus* and 64 species of eucalypts. They observed a faster than linear increase in RI with genetic distance, consistent with the prediction of the snowball model. Results from other studies using a similar approach have provided little support for the snowball model (Coyne and Orr 1989, 1997; Sasa *et al.* 1998; Fitzpatrick 2002; Presgraves 2002; Lijtmaer *et al.* 2003; Mendelson *et al.* 2004; Bolnick and Near 2005; Gourbière and Mallet 2010; Giraud and Gourbière 2012), leading some to pronounce the snowball "missing" (Johnson 2006; Gourbière and Mallet 2010). However, this indirect approach cannot provide a strong test of the snowball model because it relies on the untested ancillary assumption of a linear relationship between the number of DMIs and RI. This assumption will not be met if, for example, DMIs do not act independently on RI (Orr 1995; Welch 2004).

The second approach to testing the snowball model involves estimating the

number of DMIs directly. For example, Moyle and Nakazato (2010) used a QTL mapping approach to test the snowball model in species of *Solanum*. They introgressed one or a few genomic segments from one species to another. When an introgressed segment caused a reduction in fitness, they concluded that it participated in a DMI. They found that DMIs affecting seed sterility accumulated faster than linearly, in agreement with the prediction of the snowball model. However, DMIs affecting pollen sterility appeared to accumulate linearly, contrary to the snowball model. Studies following this second approach (Matute *et al.* 2010; Moyle and Nakazato 2010; Matute and Gavin-Smyth 2014; Sherman *et al.* 2014; Wang *et al.* 2015) are likely to underestimate the true number of DMIs for two reasons. First, the introgressed genomic segments typically contain many genetic differences. For example, the individual segments introgressed in Moyle and Nakazato (2010) included approximately 2–4% of the genome, and likely contained hundreds of genes. Second, individual alleles might participate in multiple DMIs, specially if complex DMIs are common.

The third and final approach is ingenious because it does not require the direct study of hybrids. Consider two species, 1 and 2, diverged at k loci. If an allele, X_2, at one of these loci (X) is known to be deleterious in species 1 but is fixed in species 2, then species 2 must carry compensatory alleles at one or more loci $(Y_2, Z_2, \dots)$ that are not present in species 1 (which carries alleles $Y_1, Z_1, \dots$ at those loci). In other words, there must be a DMI involving the X_2 and $Y_1, Z_1, \dots$ alleles.

Following Welch (2004), I define $\mathcal{P}_1$ as the proportion of the k fixed differences

between the species where the allele from one species is deleterious in the other species. If each allele participates at most in one DMI, then $\mathcal{P}_1 = I_k/k$. This relationship assumes that p is low. If, in addition, $\mathcal{P}_1$ is entirely based on simple DMIs, then it is expected to increase linearly with genetic distance according to the snowball model (Equation 1.2; Welch 2004)

$$\mathcal{P}_1 = \frac{(k-1)p}{2} \quad . \tag{2.1}$$

Kondrashov *et al.* (2002) and Kulathinal *et al.* (2004) estimated $\mathcal{P}_1$ in mammals and insects, respectively. Surprisingly, both studies reported that $\mathcal{P}_1 \approx 10\%$ and is constant over broad ranges of genetic distances (e.g., human compared to either nonhuman primates or fishes, Kondrashov *et al.* 2002). These results are inconsistent with the prediction of the snowball model (Welch 2004; Fraïsse *et al.* 2016).

The tests of the snowball model outlined above give inconsistent results. Specifically, the most direct approaches (i.e., the second and third) give opposite results, a paradox first noted by Welch (2004). One common limitation to all approaches is that they focus on testing predictions of the snowball model, without testing its assumptions (e.g., constant p). Furthermore, each approach makes additional assumptions that also go untested (e.g., DMIs act independently on RI).

2.2 Methods

2.2.1 Simulating the snowball using RNA folding

I begin by picking a random RNA sequence that meets the conditions of eq 1.3, define its secondary structure as the reference, and allow it to accumulate 200 random neutral substitutions sequentially, allowing multiple hits. The resulting sequence is used as the ancestor. Table 2.1 shows summary statistics for the ancestral sequences for $\alpha = 12$.

The burn-in period is necessary because the initial sequence is not representative for the fitness landscape. For example, it has the reference structure (i.e., $\delta_i = 0$ base pairs), whereas most sequences in the fitness landscape are $\delta_i \approx \alpha$ base pairs away from the reference structure.

The ancestor is used to found two identical haploid lineages. The lineages evolve by alternately accumulating a series of neutral substitutions without gene flow (allopatry) until they differ at $k = 40$ sites. At a given step, one of the evolving sequences is subjected to a random mutation. If the mutation is neutral, it is allowed to substitute; if it is deleterious, it is discarded and a new random mutation is tried. The process is repeated until a neutral mutation is found. At the next step, the other evolving lineage is subjected to the same process.

At each step, the only sites that are allowed to mutate are those that have not yet undergone a substitution in either lineage since the lineages have started to diverge from their common ancestor. This constraint implies that no more

Table 2.1: Properties of the 10^3 ancestors used in the simulations with $\alpha = 12$

Property	Mean	S.D.*
Sequence		
GC content	0.49	(0.05)
Hamming distance from the reference sequence	56.16	(5.30)
Structure		
Minumum free energy (kcal mol^{-1})	–22.56	(5.70)
Number of base pairs	24.91	(4.16)
Proportion of inviable single mutants	0.57	(0.11)
Number of potential DMIs per site	8.99	(5.70)
Base pair distance from the reference sequence	11.29	(1.03)
Ensemble		
Base pair distance between pairs of sequences	49.51	(6.01)

* Standard deviation

than two alleles are observed at each site during the course of evolution and that substitutions are irreversible, in agreement with the assumptions of Orr's (1995) model. All types of base-substitution mutations have equal probability. Insertions and deletions are not considered.

2.2.2 How to find DMIs

In this section I use the general terms genotypes, loci and alleles, instead of sequences, sites and nucleotides.

Two genotypes, 1 and 2, both have fitness $w = 1$ and differ at $k \geq 2$ loci. Loci are denoted by $A, B, C, \ldots$ The alleles of genotype 1 are indicated by a subscript 1 $(A_1, B_1, C_1, \ldots)$; the alleles of genotype 2 are indicated by a subscript 2 $(A_2, B_2, C_2, \ldots)$. Introgression of the A_1 and B_1 alleles from genotype 1 to genotype 2 is denoted $1 \xrightarrow{A,B} 2$.

Simple DMIs: There is a simple DMI between the A_1 and B_2 alleles if *all* of the following 6 conditions are met.

1. The single introgression $1 \xrightarrow{A} 2$ results in an inviable genotype (Figure 2.1, step I). On its own, this condition indicates that there is a DMI between the A_1 allele and one or more alleles from genotype 2 at the remaining $k - 1$ loci $(B_2, C_2, \ldots)$.

2. The single introgression $2 \xrightarrow{B} 1$ results in an inviable genotype. On its own, this condition indicates that there is a DMI between the B_2 allele and one or

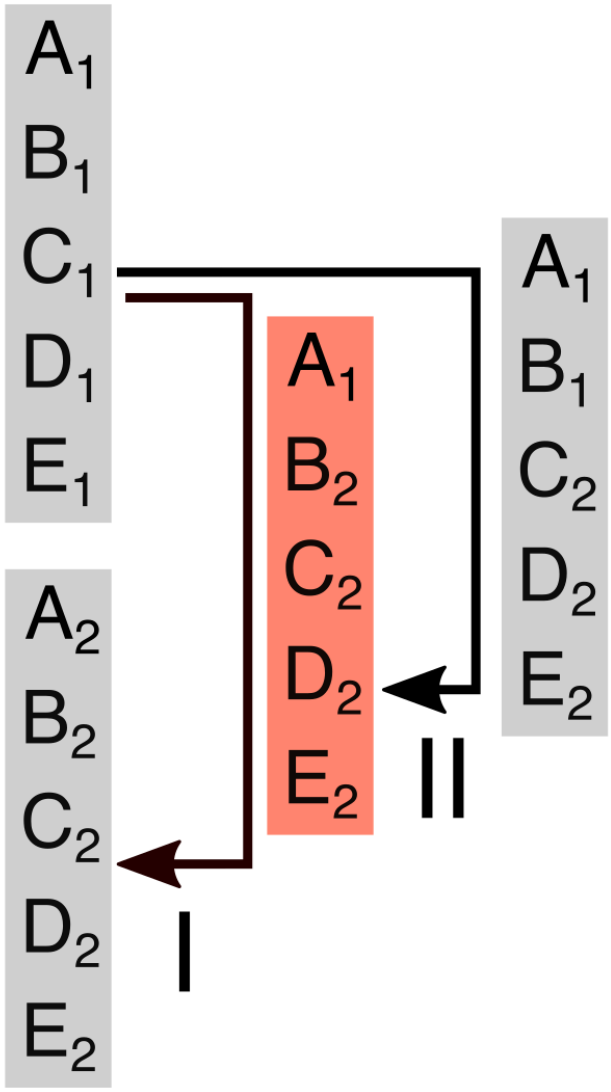

Figure 2.1: Detecting DMIs. To find simple DMIs, I use an introgression–rescue assay where I introgress one diverged allele between the two lineages (step I: $1 \xrightarrow{A} 2$), and if this substitution results in an inviable genotype (red), I try to rescue it with a second introgression (step II: $1 \xrightarrow{A,B} 2$). If the second introgression rescues viability, I conclude that the there is a DMI between the first introgressed allele (A_1) and the resident allele at the second locus (B_2). The additional criteria for establishing whether the DMI is simple or complex are explained on page 24

more alleles from genotype 1 at the remaining $k-1$ loci $(A_1, C_1, \dots)$. Taken together, conditions #1–2 are not sufficient to indicate that the A_1 and B_2 alleles participate in the same DMI.

3. The double introgressions $1 \xrightarrow{A,B} 2$ and $2 \xrightarrow{A,B} 1$ both result in viable genotypes (Figure 2.1, step II). In other words, a second introgression rescues viability. Taken together, conditions #1–3 indicate that the A_1 and B_2 alleles participate in the same DMI; the conditions do not, however, rule out the possibility that the DMI involves additional alleles from either genotype at the remaining $k-2$ loci $(C, D, \dots)$. In other words, the DMI might be simple or complex.

4. A_1 and B_2 are not both ancestral (Orr 1995). If conditions #1–3 are met but condition #4 is violated, then the DMI must involve a derived allele at an additional locus—i.e., the DMI is complex—because A_1 and B_2 were not incompatible in the ancestor.

5. If both A_1 and B_2 are derived alleles, this condition is ignored. If A_1 is an ancestral allele, then the B_2 substitution occurred after the A_2 substitution; if B_2 is an ancestral allele, then the A_1 substitution occurred after the B_1 substitution (Orr 1995). If conditions #1–4 are met but condition #5 is violated then the DMI is complex because A_1 and B_2 were not incompatible in the background in which the derived allele arose.

6. If the latest substitution at either the A or the B locus was the i-th substitution, and $i < k$, then conditions #1–3 are also met in the genotypes present

immediately after the i-th substitution. If conditions #1–5 are met but condition #6 is violated then the DMI is complex because its expression depends on the genetic background.

To count simple DMIs in my simulations, I introgress nucleotides between the two sequences at each of the k divergent sites, in both directions. Every time an introgression results in an inviable genotype (condition #1), I look for another introgression in the opposite direction that also results in an inviable genotype (condition #2). I then test both double introgressions involving these alleles to test for condition #3. If I find a pair of alleles satisfying conditions #1–3, I test for conditions #4–6 directly. I count simple DMIs after every substitution when $k \geqslant 2$.

Complex DMIs: Imagine that condition #1 for a simple DMI is met: a single introgression $1 \xrightarrow{A} 2$ results in an inviable genotype. As explained above, this is indicative of a DMI involving the A_1 allele. This DMI is complex if *any* of the following 4 conditions are met.

7. It satisfies conditions #2–3 for a simple DMI but violates one or more of conditions #4–6.

8. The double introgression $1 \xrightarrow{A,B} 2$ rescues viability, but the single introgression $2 \xrightarrow{B} 1$ results in a viable genotype (i.e., condition #2 is violated).

9. The double introgression $1 \xrightarrow{A,B} 2$ rescues viability, but the double introgression $2 \xrightarrow{A,B} 1$ results in an inviable genotype (i.e., condition #3 is violated).

10. There is no double introgression of the form $1 \xrightarrow{A,B} 2$ that rescues viability (i.e., condition #3 is violated).

A DMI is also complex if it satisfies the following condition:

11. The introgression of $1 < i < k$ alleles (e.g., $1 \xrightarrow{A,B,\ldots} 2$) results in an inviable genotype, but all the introgressions of each individual allele and of any combination of between 2 and $i - 1$ of the alleles result in a viable genotype. This condition indicates that the i alleles participate in a DMI of order $n \geqslant i + 1$.

The criteria described above (conditions #7–11) allow us to detect complex DMIs. However, counting them for highly diverged sequences (high k) is virtually impossible for two reasons. First, the number of high-order introgressions required is enormous. Second, as the conditions #1–3 for detecting simple DMIs highlight, establishing that alleles participate in the *same* DMI requires additional introgressions. For example, if alleles A_1 and B_1 from population 1 are incompatible with allele C_2 from population 2, then both the double introgression $1 \xrightarrow{A,B} 2$ and the single introgression $2 \xrightarrow{C} 1$ result in an inviable genotype. However, showing that the 3 alleles are involved in the same DMI of order $n = 3$ would require demonstrating that the triple introgressions $1 \xrightarrow{A,B,C} 2$ and $2 \xrightarrow{A,B,C} 1$ both result in viable genotypes. Thus, without conducting "rescue" introgressions, the introgressions in both directions will tend to overestimate the number of complex DMIs. To avoid this problem, I estimate the number of complex DMIs through

all single, double and triple introgressions in one direction only (e.g., from population 1 to population 2). For the single introgressions, I count complex DMIs using conditions #7–10 (these conditions require performing introgressions in both directions, but only DMIs detected from an introgression in one direction are counted). For double and triple introgressions, I use condition #11.

The resulting count of complex DMIs will still underestimate the true number for two reasons. First, if the introgressed alleles participate in more than one complex DMI, an introgression test can only detect a single DMI (this limitation does not apply to simple DMIs). Second, complex DMIs that can only be detected by introgressing four or more alleles will not be detected.

2.2.3 Proportion of single introgressions involved in a DMI

I use the single introgression data to calculate $\mathcal{P}_1$, the proportion of the $2k$ single introgressions at diverged sites (in both directions) that result in an inviable sequence (Welch 2004).

2.2.4 DMI network

The simple DMIs that might, potentially, affect a sequence can be computed exhaustively by measuring the fitness of all possible single and double mutants derived from the sequence. For every pair of sites, there are 9 combinations of double mutants. A potential simple DMI is defined as an inviable double mutant

between mutations that are individually neutral. I summarize the pattern of interactions between sites using an undirected network where the vertices are sites and the edges represent the existence of at least one potential simple DMI between them. The resulting network is an example of the networks of interactions described by Orr and Turelli (2001) and Livingstone *et al.* (2012).

I measure the degree of similarity between two DMI networks X and Y using the Jaccard index

$$J = \frac{|X \cap Y|}{|X \cup Y|} \quad , \tag{2.2}$$

where $|X \cap Y|$ is the number of edges shared between the two networks, $|X \cup Y|$ is the sum of $|X \cap Y|$ and the numbers of edges unique to X and to Y, and there is a one-to-one correspondence between the vertices of X and Y (i.e., between the sites in the corresponding sequences). J varies between 0 (the two networks have no edges in common) and 1 (the two networks are identical).

2.2.5 Reproductive isolation

The degree of RI between the sequences is defined as

$$\mathrm{RI} = 1 - \overline{w}_R \quad ,$$

where $\overline{w}_R$ is the mean fitness (Equation 1.3) of all possible 198 recombinants resulting from a single crossover between the two sequences.

2.2.6 "Holeyness" of the fitness landscape

For each simulation, I took the ancestor and each of the $k = 40$ genotypes generated during the course of evolution and measured the proportion of their single mutant neighbors (300 per sequence) that are inviable, excluding the 41 original sequences. This estimates the local holeyness of the fitness landscape traversed by the diverging lineages.

2.2.7 Direct simulation of the snowball model

I also simulate the accumulation of DMIs following the snowball model (Orr 1995). An ancestral genotype has multiple loci and is used to found two identical haploid lineages. The lineages are allowed to evolve by alternately accumulating neutral substitutions (Figure 2.2).

After the k-th substitution, simple DMIs are sampled at random with probability p from all pairs of alleles consisting of the latest derived allele paired with any of the $k - 1$ ancestral or derived alleles from the other population at loci that have previously undergone substitutions in either population. For example, when $k = 4$ the new possible simple DMIs are: D_2/A_1, D_2/B_0, and D_2/C_1 (Figure 2.2).

	Lineage 1	Lineage 2
$k = 0$	$\mathrm{A}_0\,\mathrm{B}_0\,\mathrm{C}_0\,\mathrm{D}_0\ldots$	$\mathrm{A}_0\,\mathrm{B}_0\,\mathrm{C}_0\,\mathrm{D}_0\ldots$
$k = 1$	$\mathbf{A_1}\,\mathrm{B}_0\,\mathrm{C}_0\,\mathrm{D}_0\ldots$	$\mathrm{A}_0\,\mathrm{B}_0\,\mathrm{C}_0\,\mathrm{D}_0\ldots$
$k = 2$	$\mathbf{A_1}\,\mathrm{B}_0\,\mathrm{C}_0\,\mathrm{D}_0\ldots$	$\mathrm{A}_0\,\mathbf{B_2}\,\mathrm{C}_0\,\mathrm{D}_0\ldots$
$k = 3$	$\mathbf{A_1}\,\mathrm{B}_0\,\mathbf{C_1}\,\mathrm{D}_0\ldots$	$\mathrm{A}_0\,\mathbf{B_2}\,\mathrm{C}_0\,\mathrm{D}_0\ldots$
$k = 4$	$\mathbf{A_1}\,\mathrm{B}_0\,\mathbf{C_1}\,\mathrm{D}_0\ldots$	$\mathrm{A}_0\,\mathbf{B_2}\,\mathrm{C}_0\,\mathbf{D_2}\ldots$

Figure 2.2: Sequence evolution in a direct simulation of the snowball model showing the first $k = 4$ substitutions. Only 4 loci are shown, denoted by A–D. Ancestral alleles are indicated by subscript 0. Derived alleles are shown in bold and indicated by subscripts 1 or 2 depending on the lineage.

2.3 Results

2.3.1 Simple DMIs do not snowball in the RNA-folding model

The snowball model predicts that the number of simple DMIs, I_k, should increase faster than linearly with the number of substitutions, k. I tested this prediction using 10^3 evolutionary simulations with the RNA-folding model. For each simulation, I fitted two models: the snowball model in Equation 1.2 and a linear model of the form

$$I_k = (k-1)b \quad , \tag{2.3}$$

where b is the slope. The $k-1$ term ensures that $I_1 = 0$, as in the snowball model. Both models have a single parameter that I estimated using the method of least squares. I compared the level of support for each model using Akaike's Information Criterion (AIC). If the difference in the AIC values (ΔAIC) was greater than a threshold, I concluded that there was stronger support for the model with the lower AIC. Setting the ΔAIC threshold at 2, 41.9% of RNA-folding simulations provided stronger support for the snowball model, 49.1% provided stronger support for the linear model, and 9.0% provided approximately equal support for both models (Figure 2.3). Increasing the ΔAIC threshold did not affect this result qualitatively (Figure 2.3). The average response in the number of DMIs in the RNA-folding simulations was approximately linear (Figure 2.5A), in agreement with the AIC analysis. To evaluate the extent to which the lack of support for the snowball model was caused by random noise in the simulations, I conducted

10^3 direct simulations of the snowball process over $k = 40$ substitutions assuming values of p estimated by fitting the snowball model in Equation 1.2 to the RNA-folding data (Figure 2.4A). As expected, these direct snowball simulations provided much stronger support for the snowball model than the RNA simulations (Figure 2.3). I conclude that simple DMIs do not snowball in at least some RNA-folding simulations.

2.3.2 The probability that a DMI appears is approximately constant in the RNA-folding model

What explains the lack of support for the snowball model in the RNA-folding simulations? One possibility is that p itself evolved, contrary to the assumption of the snowball model (Orr 1995).

If p declines with divergence according to the relationship

$$p_k = \frac{b}{k} \quad , \tag{2.4}$$

where b is a positive constant, and I substitute p by p_k in Equation 1.1, the linear model in Equation 2.3 is a solution to the resulting difference equation (assuming $I_1 = 0$). To test whether p changed as described by Equation 2.4, I measured it directly in each simulation as $p_k = \Delta I / k$, where ΔI is the number of new simple DMIs appearing as a result of the $(k+1)$-th substitution that involve the latest derived allele (see Equation 1.1). I found that, although p_k declined with k, the trend did not follow Equation 2.4. Indeed, when $k \gtrsim 10$, p_k was approximately constant (Figure 2.5B).

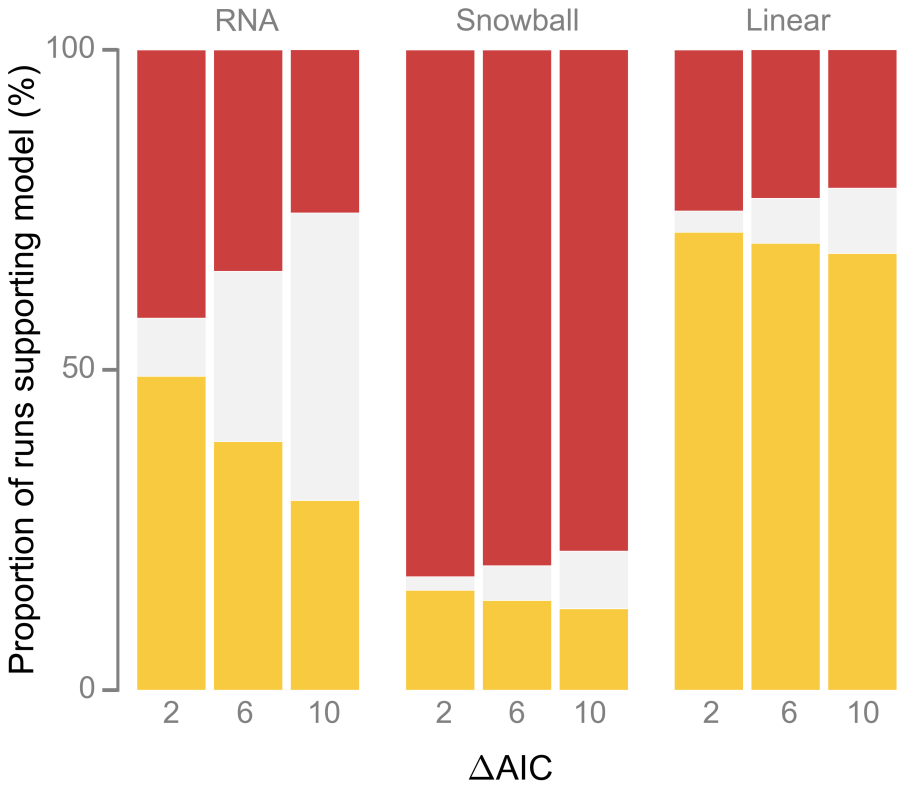

Figure 2.3: Simple DMIs do not snowball in the RNA-folding model. I fitted the snowball model (Equation 1.2) and a linear model (Equation 2.3) to each run from three kinds of simulations: simulations of the RNA-folding model ("RNA"), direct simulations of the snowball model ("Snowball") with values of p estimated by fitting the model in Equation 1.2 to each RNA-folding simulation (Figure 2.5A), and direct simulations of the snowball model ("Linear") with values of p_k from Equation 2.4 estimated by fitting the model in Equation 2.3 to each RNA-folding simulation (Figure 2.5B). Red segments show the proportions of runs providing stronger support for the snowball model; yellow segments show the proportions of runs providing stronger support for the linear model;gray segments show the proportions of runs providing approximately equal support for both models. Each bar is based on 10^3 stochastic simulations. The level of support for the two models was evaluated for three different ΔAIC thresholds.

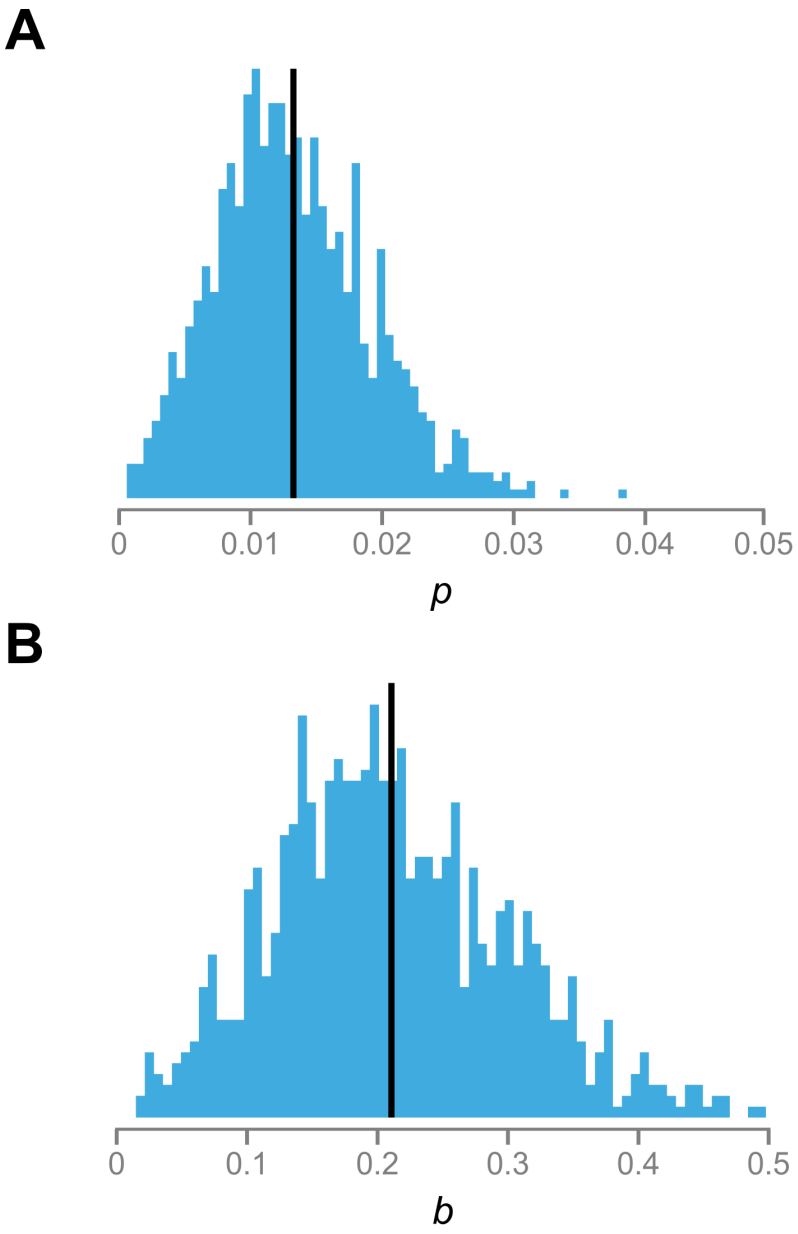

Figure 2.4: Distributions of the parameters of the snowball and linear models in the RNA-folding simulations. (A) Probability, p, that a simple DMI appears in the snowball model. (B) Rate of accumulation, b, of simple DMIs in the linear model. For each of the 10^3 stochastic RNA-folding simulations I estimated p and b by fitting the models in Equations 2 and 6, respectively, by the method of least squares. The solid black lines indicate the means of the distributions: $\bar{p} = 0.013$ and $\bar{b} = 0.211$.

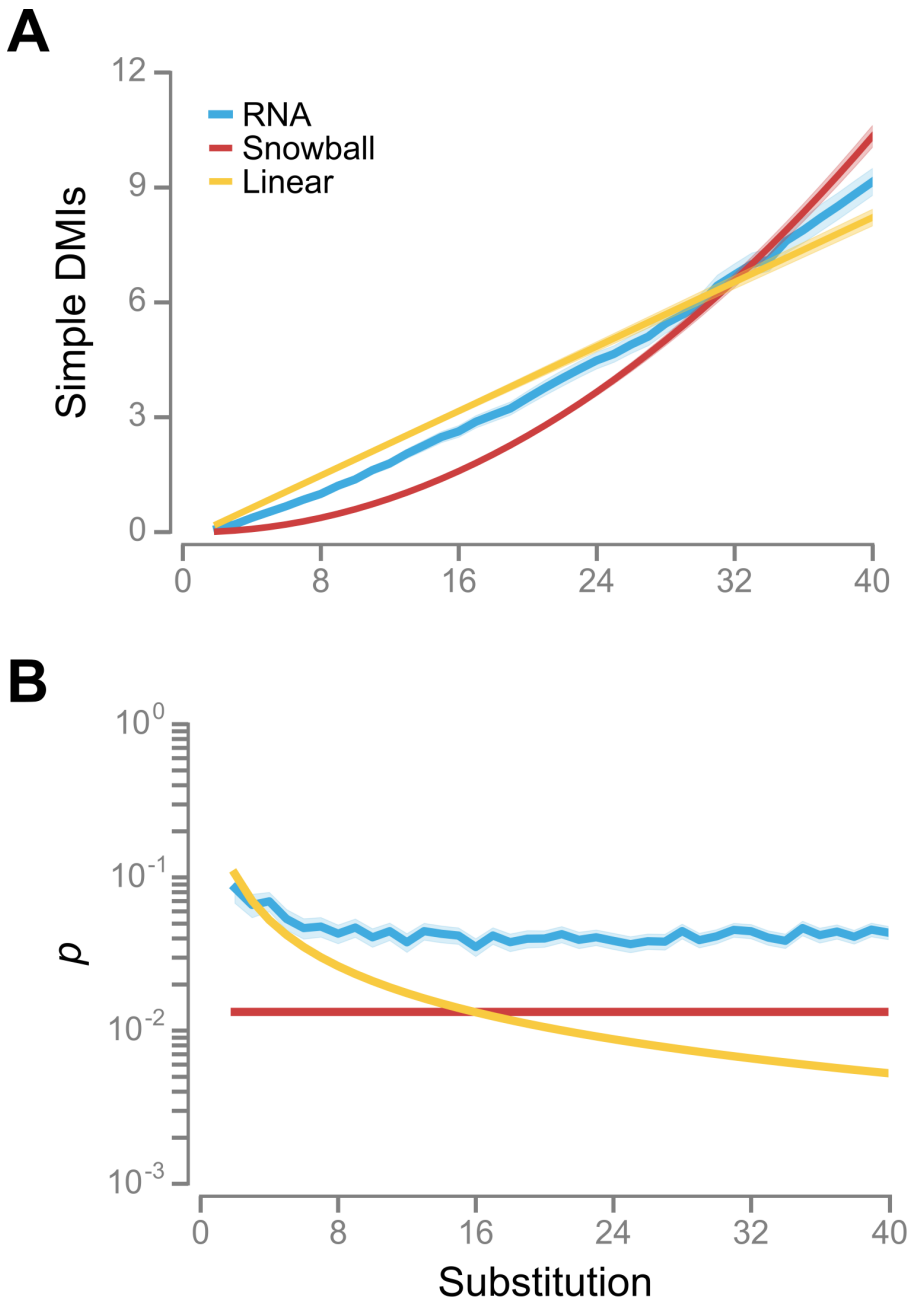
A
12
9
6
3
0
Simple DMIs
RNA
Snowball
Linear
0
8
16
24
32
40
B
10^{0}
10^{-1}
10^{-2}
10^{-3}
ρ
0
8
16
24
32
40
Substitution

Figure 2.5 ***(previous page)*****:** Simple DMIs do not snowball in the RNA-folding model. (A) Evolution of the number of simple DMIs, I_k, as two populations diverge by accumulating substitutions, k. Values are means of 10^3 runs of three different kinds of stochastic simulations: "RNA," simulations of the RNA-folding model (blue); "snowball," direct simulations of the snowball process with constant p estimated as explained in (B) (red); "linear," direct simulations of the snowball process with declining p estimated as explained in (B) (yellow). (B) Evolution of the probability, p_k, that there is a simple DMI between the latest derived allele after the $(k+1)$-th substitution and one of the k alleles at the loci that have previously undergone substitutions. The blue line ("RNA") shows the values of p_k estimated at each substitution directly from the RNA-folding simulations. The red line ("snowball") shows the values of p estimated by fitting the model in Equation 1.2 to each RNA-folding simulation (Figure 2.9A). The yellow line ("linear") shows the values of p_k from Equation 2.4 based on estimates of b obtained by fitting the model in Equation 2.3 to each RNA-folding simulation (Figure 2.9B). Values are means of 10^3 simulations. Shaded regions indicate 95% confidence intervals, CIs.

2.3.3 Simple DMIs do not persist indefinitely in the RNA-folding model

The previous analysis also revealed that fitting the snowball model to the RNA-folding data underestimated the true value of p by approximately 3-fold (Figure 2.5B). This discrepancy indicates that a more fundamental assumption of the snowball model may be violated in the RNA-folding model: that simple DMIs, once they have arisen, persist indefinitely. This assumption is implicit in the original description of the snowball model (Orr 1995) and, to my knowledge, has never been called into question.

To test this assumption, I estimated the DMI networks of sequences as they evolved in my RNA-folding model. Figure 2.6A shows an example of an RNA

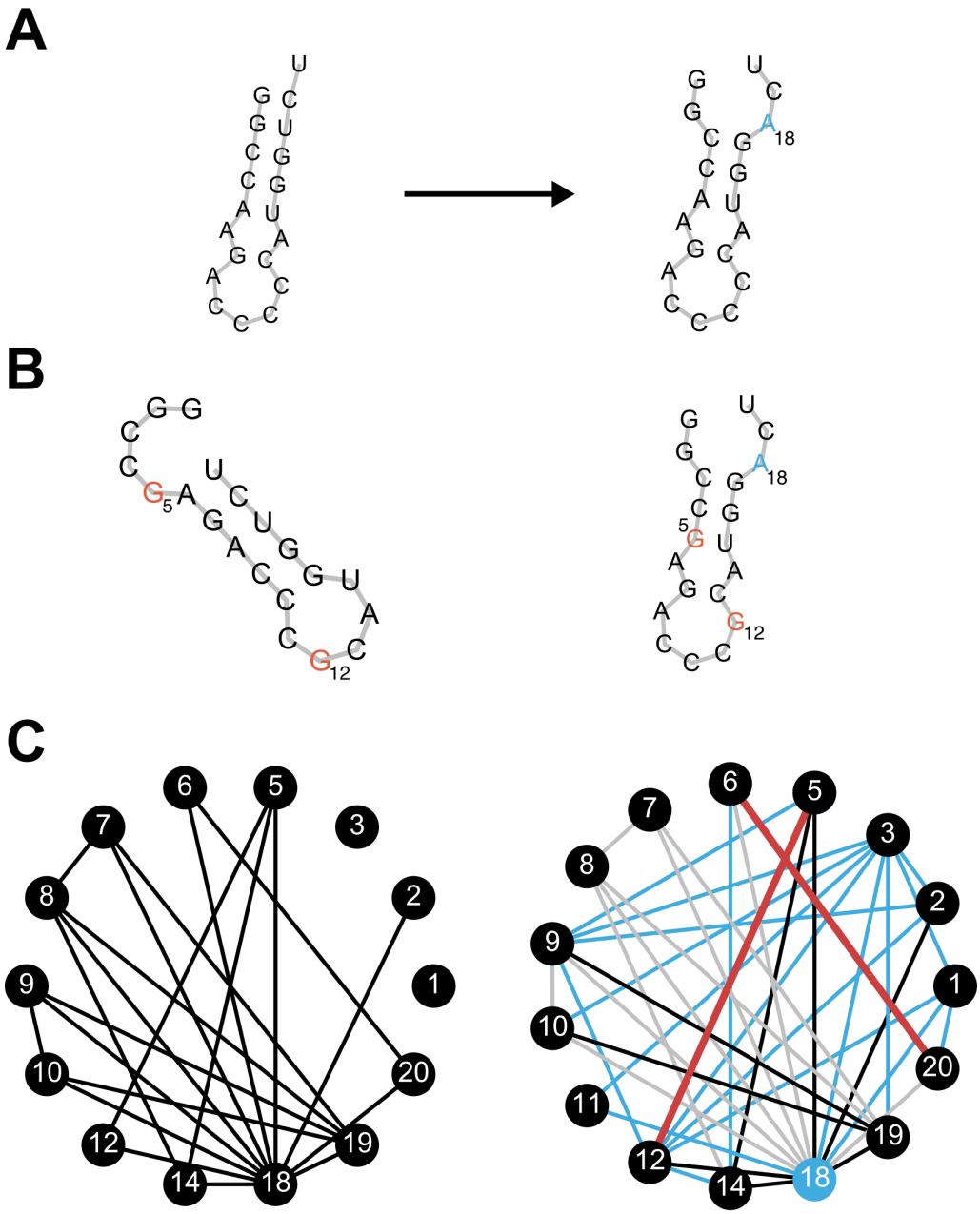
A
B
C
5
12
18

Figure 2.6 ***(previous page)*****:** A single substitution can dramatically rearrange the network of potential DMIs. (A) The 20-nucleotide-long RNA sequence on the left acquires a neutral U→A substitution at position 18 (blue). The holey fitness landscape is defined by $\alpha = 2$ (Equation 1.3). The secondary structure of the sequence on the left is the reference ($\delta_i = 0$ base pairs). The structure on the right is $\delta_i = 2$ base pairs away from the reference. (B) There is a potential simple DMI between positions 5 and 12 for the sequence on the left. A double mutant at those positions (5: A→G, 12: C→G, red) makes the structure inviable ($\delta_i = 11$ base pairs), even though the single mutations are neutral (not shown). However, a single substitution causes the potential simple DMI to disappear in the sequence on the right, although the single mutations remain neutral in the new background (not shown). In other words, the substitution causes the simple DMI to become complex. (C) DMI networks of the sequences in (A). Vertices correspond to positions in the sequences. An edge in the network on the left indicates that there is at least one potential simple DMI between the two sites (positions 4, 13 and 15–17 have no potential DMIs in either network and are not shown). Black edges in the network on the right are shared between the two networks. Blue edges exist only in the network on the right and indicate the appearance of new potential simple DMIs between sites caused by the substitution. Gray and red edges indicate losses of potential simple DMIs in the network on the right. Gray edges indicate losses due to the constituent alleles no longer being neutral in the new background. Red edges indicate losses caused by complexification; the DMI discussed in (B) is an example (5–12 edge). The Jaccard index (Equation 2.2) between the two networks is $J = 0.205$.

sequence evolving on a holey fitness landscape. Initially the sequence displays potential simple DMIs between 21 pairs of sites (Figure 2.6C). Figure 2.6B illustrates a potential simple DMI between positions 5 and 12. I refer to these simple DMIs as *potential* because if two diverging lineages each accumulate one of the substitutions underlying one of these DMIs, a simple DMI between the lineages will appear.

The snowball model assumes that the DMI network is static: as populations evolve they actualize potential DMIs (for an alternative, but equivalent, interpretation of DMI networks see Livingstone *et al.* 2012). However, DMI networks are not static in the RNA-folding model. After a single neutral substitution, 13 pairs of sites (62%) lost all potential simple DMIs, and potential DMIs appeared between 18 new pairs of sites (Figure 2.6C).

The "loss" of a potential DMI can occur in one of two ways. First, the substitution may cause the mutations involved in the simple DMIs to become deleterious so that they can no longer participate in potential simple DMIs. A loss of this kind means that a potential simple DMI is no longer accessible through independent substitution in two lineages because one of the substitutions cannot take place. Thus, such losses do not imply that DMIs cannot persist indefinitely. However, if there is a bias towards such losses of potential DMIs relative to gains of the same kind then p is expected to decline with divergence. The majority of losses in Figure 2.6C (gray lines) are of this kind.

The second kind of loss occurs when the substitution modifies the interaction between previously incompatible alleles (red lines in Figure 2.6C). In other

words, the simple DMIs become complex. The potential simple DMI between positions 5 and 12 shown in Figure 2.6B is lost in this way. This kind of loss—complexification—implies that some simple DMIs may not persist indefinitely.

The DMI networks corresponding to the evolving lineages in the RNA-folding simulations summarized in Figure 2.5 also change dramatically relative to the ancestor as a result of successive substitutions (Figure 2.7). This indicates that complexification may be occurring in these simulations as well. In the next section I explore the consequences of the complexification of simple DMIs for snowballing.

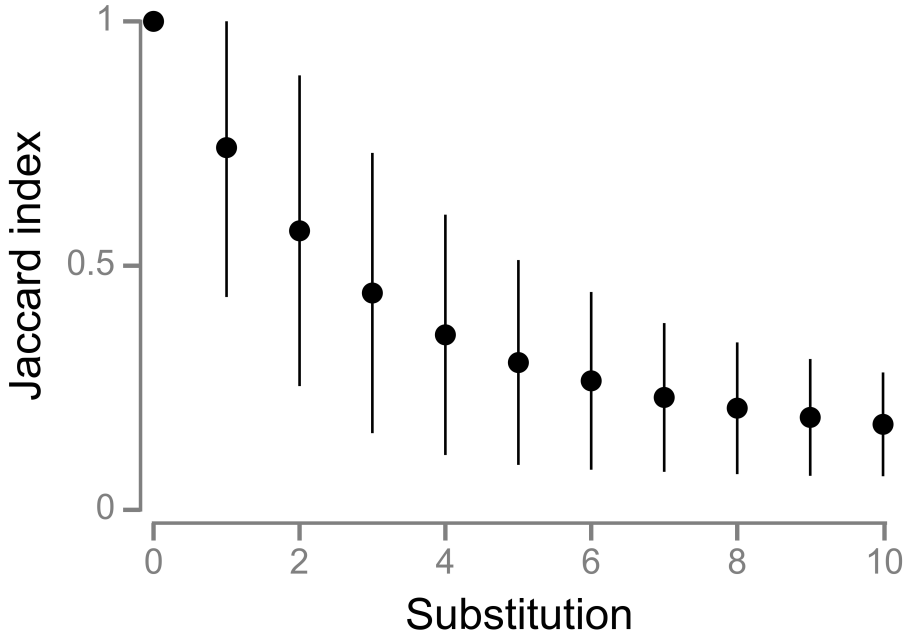

Figure 2.7: Networks of potential simple DMIs are not static in the RNA-folding model. Jaccard index (Equation 2.2) of the DMI networks of each descendant lineage after k substitutions compared to its ancestor. Values are means of 2×10^3 DMI networks (10^3 simulations, 2 lineages per simulation). Based on the simulations summarized in Figure 2.3. Error bars show ± 1 standard deviation.

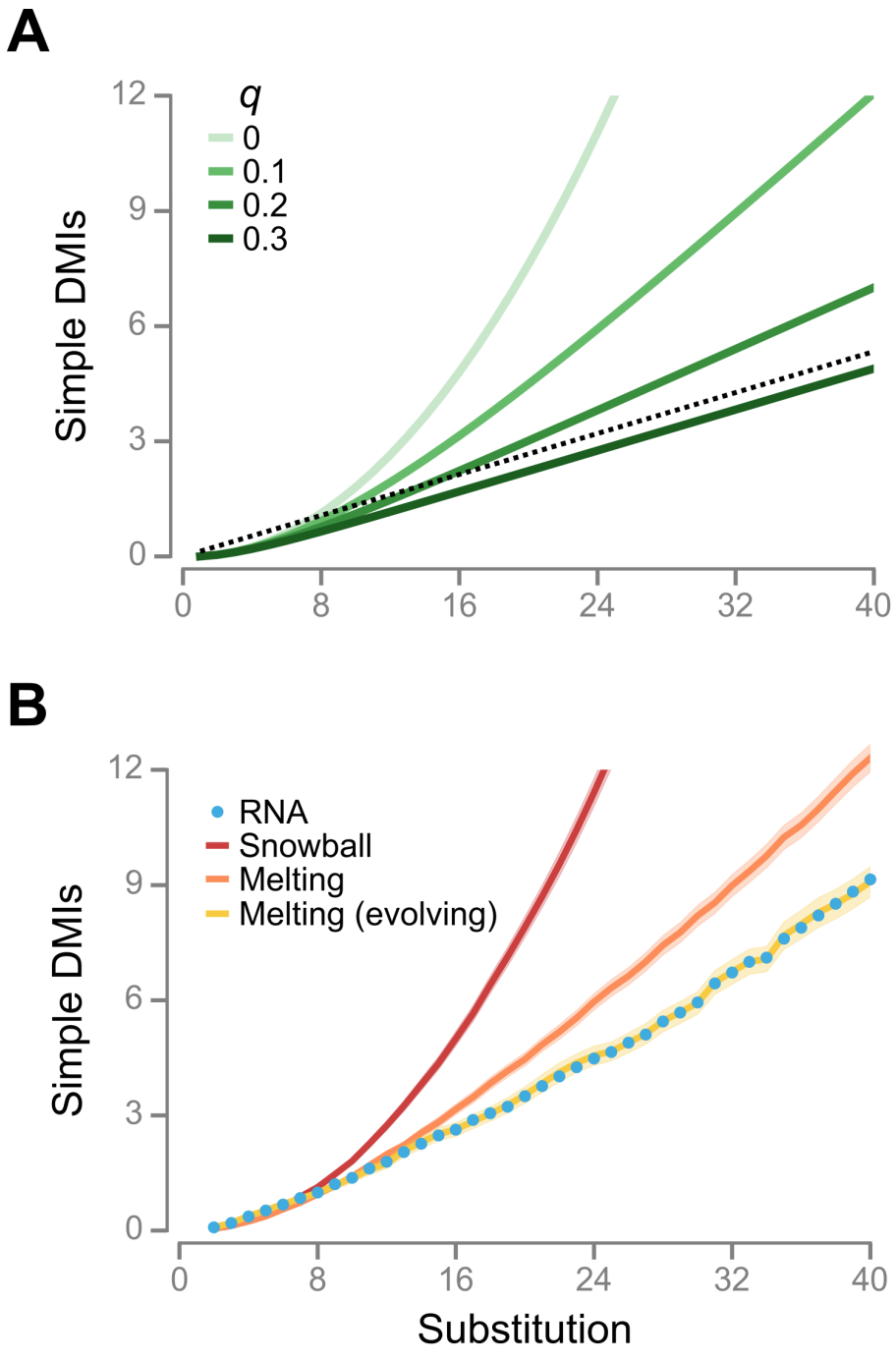
A
12
9
6
3
0
Simple DMIs
q
0
0.1
0.2
0.3
0
8
16
24
32
40
B
12
9
6
3
0
Simple DMIs
RNA
Snowball
Melting
Melting (evolving)
0
8
16
24
32
40
Substitution

Figure 2.8 *(previous page)*: The RNA-folding simulations agree with the the melting snowball model. (A) Evolution of the number of simple DMIs under the melting snowball model. Responses for $p = 0.04$ and different values of q. The dashed line shows a slope of p/q for $q = 0.3$. (B) Mean responses of 10^3 runs of four different kinds of stochastic simulations: "RNA," simulations of the RNA-folding model (blue circles, same data as in Figure 2.5A); "Snowball," direct simulations of the snowball with with constant values of p estimated directly from each RNA-folding simulation (Figure 2.9) (red); "Melting," direct simulations of the melting snowball model with constant values of p and q estimated directly from each RNA-folding simulation (Figure 2.9) (orange); "Melting (evolving)," direct simulations of the melting snowball model with evolving trajectories of p_k and q_k estimated directly from each RNA-folding simulation (yellow, dashed). Shaded regions indicate 95% CIs.

2.3.4 The RNA-folding simulations agree with the melting snowball model

I incorporate the dynamic nature of simple DMIs by extending the snowball model in Equation 1.1[1]

$$I_{k+1} = (1 - q)I_k + kp \quad , \tag{2.5}$$

where q is the probability that a simple DMI present after k substitutions becomes complex after the next substitution. Assuming $I_1 = 0$, the solution to Equation 2.5 is

$$I_k = \frac{p\left[(1-q)^k + kq - 1\right]}{q^2} \quad . \tag{2.6}$$

This prediction assumes that both p and q remain constant as populations diverge.

[1]The mathematical derivation of the melting snowball model was done by Ricardo B. R. Azevedo.

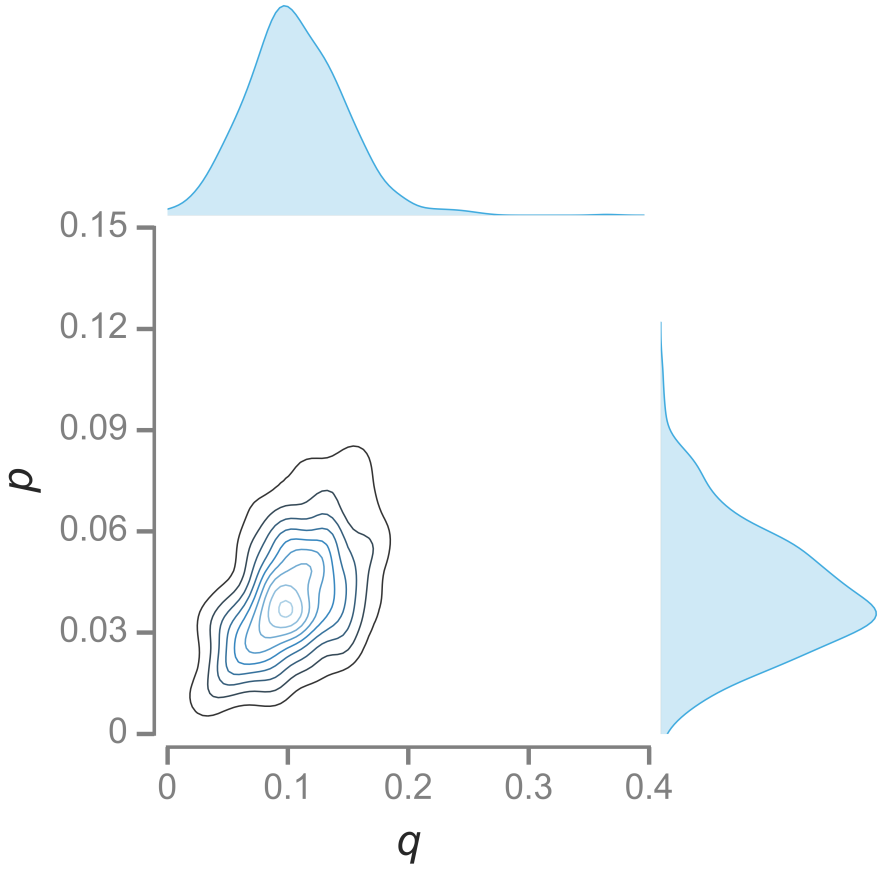

Figure 2.9: Distributions of the parameters of the melting snowball model in the RNA-folding simulations: p, the probability that a simple DMI arises, and q, the probability that a simple DMI becomes complex. One- and two-dimensional kernel density estimates based on 10^3 stochastic simulations. For each simulation I calculated p_k and q_k after every substitution (k). I then estimated an overall value of p and q as weighted averages. Values of p_k and q_k were weighted by $k(k-1)$ and I_k, respectively. The means of each distribution were $\bar{p} = 0.042$ and $\bar{q} = 0.107$.

The original metaphor evokes a snowball rolling down a hillside, picking up snow (appearance of simple DMIs) as it rolls, causing it to increase in size. To stretch the metaphor, I call the new model the *melting snowball*: as the snowball rolls it also melts (complexification of simple DMIs), causing it to decrease in size. Neither metaphor should be taken too literally, though. For example, both metaphors give the mistaken impression that the accumulation of DMIs itself *causes* the emergence of new DMIs, which is not part of either model.

The snowball model is a special case of the melting snowball model when $q = 0$. When $q > 0$, the increase in the number of simple DMIs is given by

$$\Delta I = I_{k+1} - I_k = \frac{p}{q}\left[1 - (1-q)^k\right] \quad . \tag{2.7}$$

This equation has two consequences (Figure 2.8A). First, the increase in the number of simple DMIs eventually becomes linear with a slope of approximately p/q when k is sufficiently large. Second, if q is larger, the "linearization" of Equation 2.6 occurs for lower values of k.

To test whether the complexification of simple DMIs explains the results of the RNA-folding simulations I measured q directly in my simulations as $q_k = 1 - I'_k/I_k$, where I_k is the number of simple DMIs present after the k-th substitution, and I'_k is the number of simple DMIs present after the $(k+1)$-th substitution that do not involve the latest derived allele.

The melting snowball model predicts that simple DMIs will accumulate approximately linearly when q is large relative to p (Equation 2.7). The values of q were, on average, 3-fold higher than the values of p (Figure 2.9). Furthermore,

the q/p ratio was a good predictor of whether RNA-folding simulations supported the linear or the snowball model (Figure 2.3). When the ΔAIC threshold was set at 2, q/p was 3.36 ± 0.22 (mean and 95% confidence intervals, CIs) in runs that provided stronger support for the linear model, and 2.41 ± 0.12 in runs that provided stronger support for the snowball model (Wilcoxon rank sum test, $P < 10^{-6}$). Thus, the linear response in the number of simple DMIs in the RNA-folding simulations can be explained by the melting snowball model.

To evaluate the extent to which the melting snowball model can account for the lack of support for the snowball model in my RNA-folding simulations, I conducted 10^3 direct simulations of the melting snowball process over $k = 40$ substitutions assuming values of p and q estimated directly from the RNA-folding data (Figure 2.9). The support for the snowball and linear models provided by these direct melting snowball simulations was similar to that provided by the RNA-folding simulations (Figure 2.10). These results, in combination with those on the q/p ratio, indicate that the melting snowball model explains the RNA-folding results.

Figure 2.8B shows that the melting snowball model (orange) approximates the RNA-folding data better than the snowball model (red). However, the fit is far from perfect. The lack of fit is caused by the assumptions that both p and q are constant as populations diverge. Neither assumption was met by the RNA-folding data: p decreased and q increased with k, specially when $k \lesssim 10$ (Figures 2.5B and 2.11, respectively). When I allowed p and q to vary as they did in the RNA-folding simulations, direct simulations of the melting snowball process

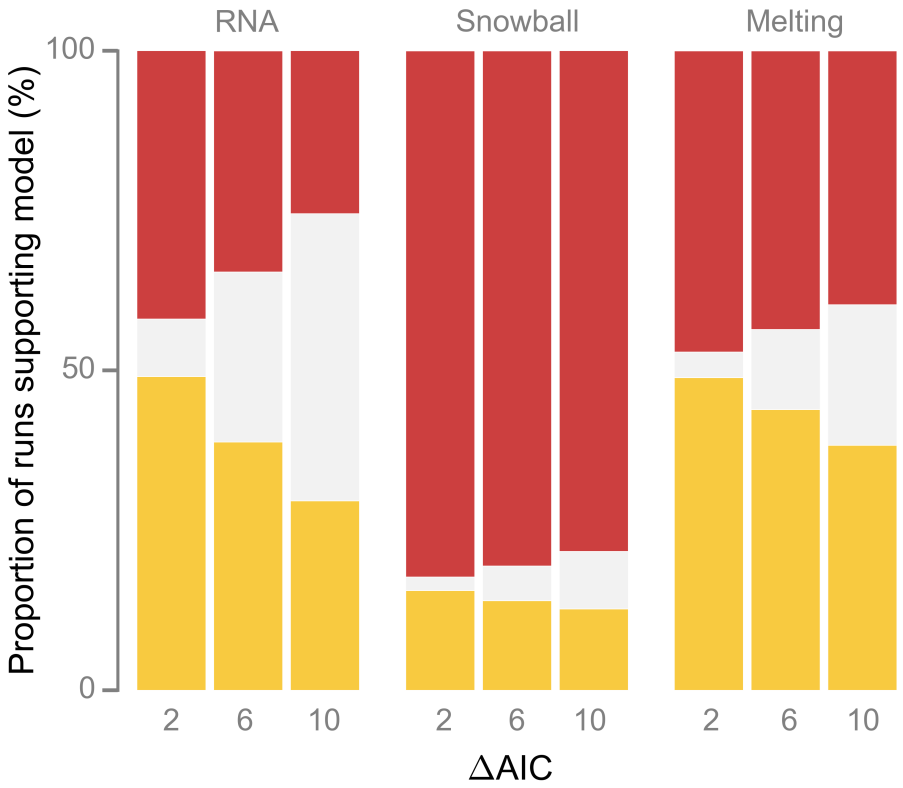

Figure 2.10: The RNA-folding model behaves as expected under the melting snowball model. I fitted the snowball model and a linear model to each run from three kinds of simulations: simulations of the RNA-folding model ("RNA"), direct simulations of the snowball model ("Snowball"), and direct simulations of the melting snowball model with values of p and q estimated directly from each RNA-folding simulation ("Melting") (Figure 2.9). Red segments show the proportions of runs providing stronger support for the snowball model; yellow segments show the proportions of runs providing stronger support for the linear model; gray segments show the proportions of runs providing approximately equal support for both models. Each proportion is based on 10^3 stochastic simulations. The level of support for the two models was evaluated for three different ΔAIC thresholds.

matched the RNA-folding data perfectly (Figure 2.8B). I conclude that the melting snowball model explains the results of the RNA-folding model, provided I relax the assumptions that p and q are constant.

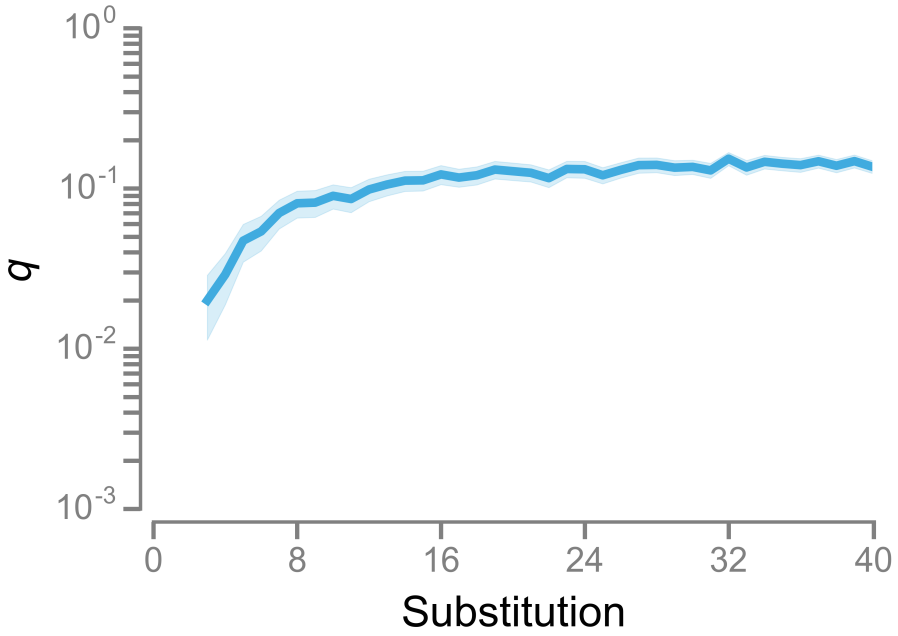

Figure 2.11: Evolution of the probability, q, that a simple DMI becomes complex. For each stochastic RNA-folding simulation we measured q_k, the probability that a simple DMI present after k substitutions will become complex after the next substitution. Values are means of 10^3 simulations at each k. Based on the simulations summarized in Figure 2.3. Shaded regions indicate 95% CIs.

2.3.5 Complex incompatibilities snowball in the RNA-folding model

So far I have focused exclusively on simple DMIs. The melting snowball model predicts that complex DMIs should exist if $q > 0$ because they will be generated continuously from simple DMIs. Furthermore, if q is high the number of DMIs

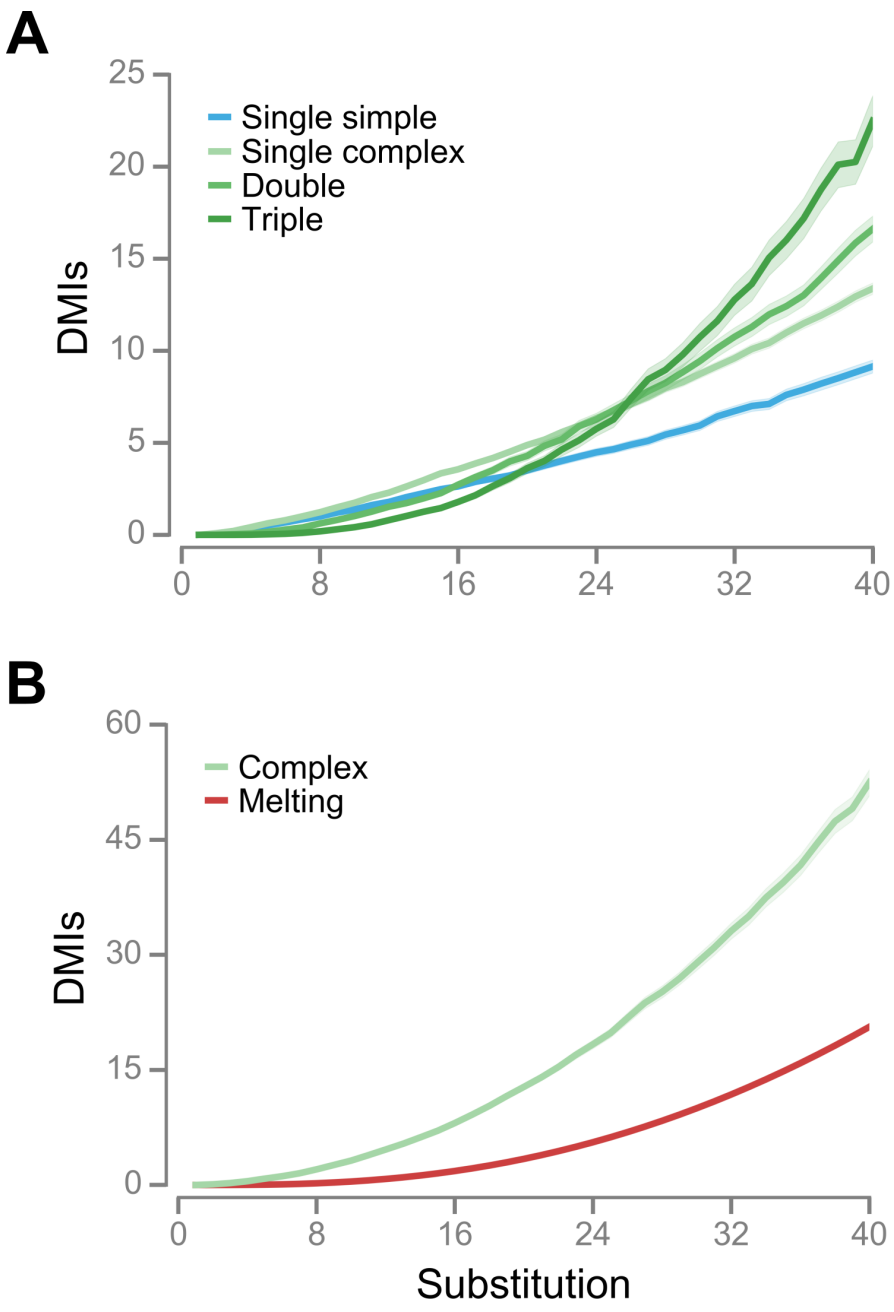

Figure 2.12: Complex DMIs snowball in the RNA-folding model. (A) DMIs inferred through single, double, and triple introgressions. (B) Total number of complex DMIs (green) compared to number predicted if all complex DMIs originate from the melting of simple DMIs ("Melting") and $p = 0.042$ and $q = 0.107$ (red). Values are means of 10^3 stochastic simulations. Shaded regions indicate 95% CIs.

should also be high. I tested this prediction in the RNA-folding model and found that complex DMIs accumulated in much higher numbers than simple ones: after $k = 40$ substitutions there were approximately 5-fold more complex DMIs than simple ones (Figure 2.12).

The snowball model predicts that the number of complex DMIs should snowball (Orr 1995; Welch 2004). Complex DMIs, unlike simple ones, did snowball (Figures 2.12 and 2.14). In addition, complex DMIs detected by introgressing more alleles accumulated faster (Figure 2.12B). Allowing multiple substitutions to occur per site during divergence did not change this pattern (Figure 2.13). These results indicate that higher-order DMIs accumulated faster than lower-order DMIs.

Did the complex DMIs originate from the "melting" of simple ones or did they appear *de novo*? If all complex DMIs arise through melting, then I would expect their number to increase according to the difference between Equations 1.2 and 2.6. Figure 2.12B shows that, although some complex DMIs likely arose from melting, many complex DMIs must have arisen *de novo*.

2.3.6 Reproductive isolation does not snowball in the RNA-folding model

Since most DMIs were complex and complex DMIs snowballed, RI would be expected to snowball in the RNA model. However, I found that RI showed a kind of inverse snowball—a "slowdown" with divergence. This pattern has been

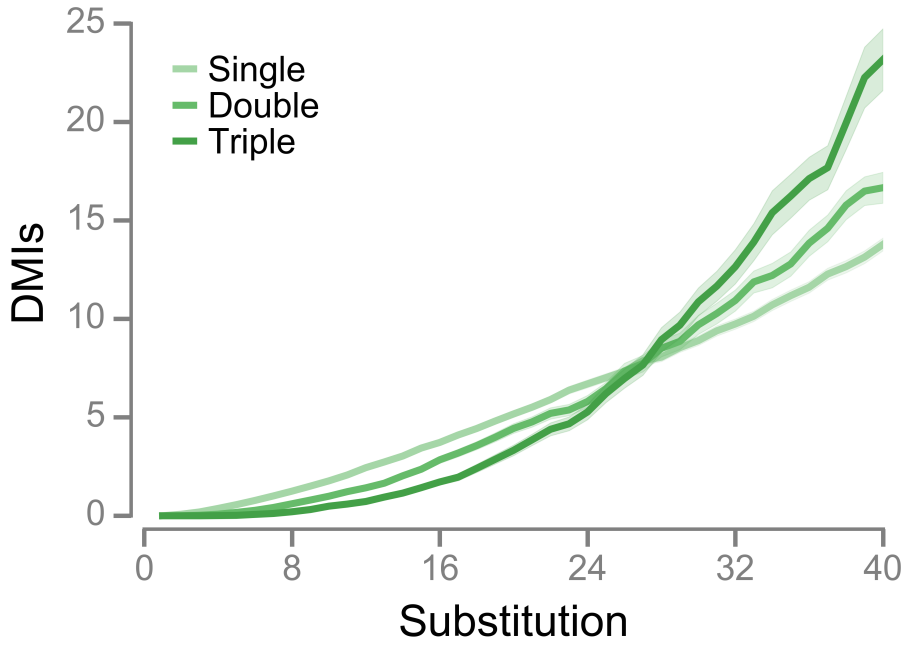

Figure 2.13: Allowing sites to undergo multiple substitutions does not affect the pattern of accumulation of DMIs inferred through single, double, and triple introgressions. The results are based on 10^3 RNA-folding simulations for $\alpha = 12$. Shaded regions indicate 95% CIs.

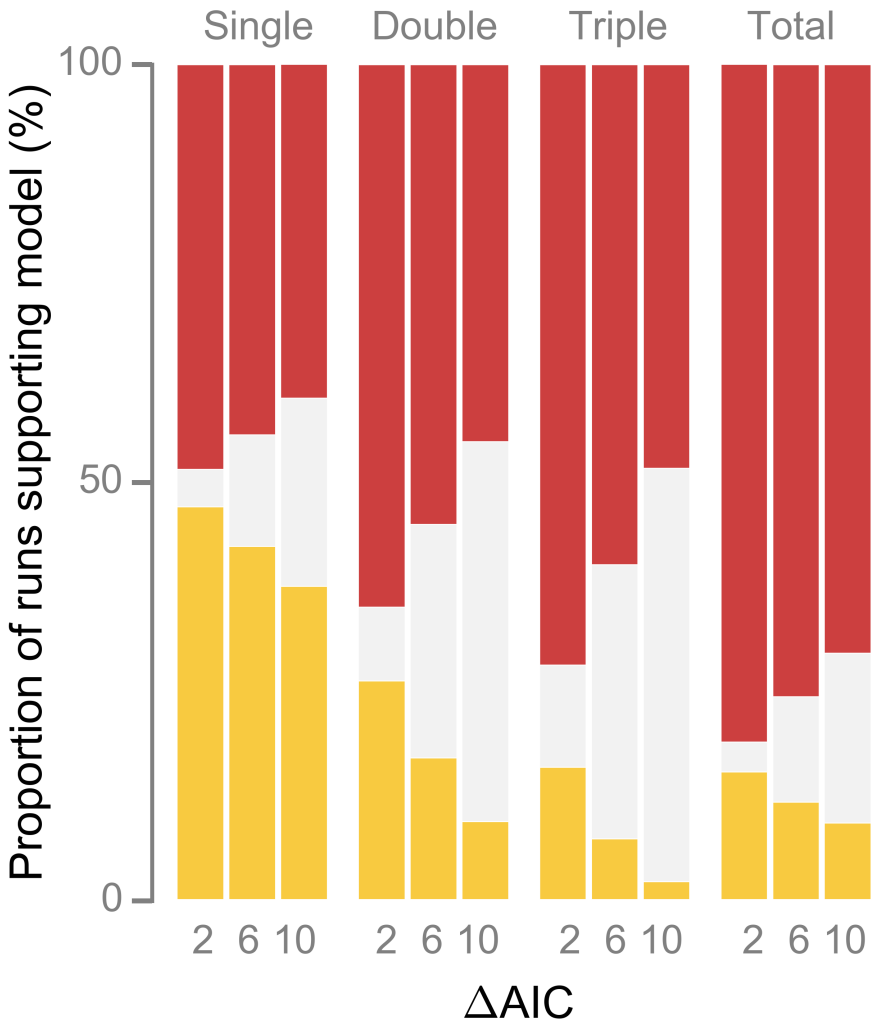

Figure 2.14: Complex DMIs snowball in the RNA-folding model. I fitted two models to the evolutionary responses in the numbers of complex DMIs found through single, double, and triple introgressions (Figure 2.12): a linear model ($I_k = bk$) and a snowball model ($I_k = bk^2$). Red segments show the proportions of runs providing stronger support for the snowball model; yellow segments show the proportions of runs providing stronger support for the linear model; gray segments show the proportions of runs providing approximately equal support for both models. Each proportion is based on 10^3 stochastic simulations. The level of support for the two models was evaluated for three different ΔAIC thresholds.

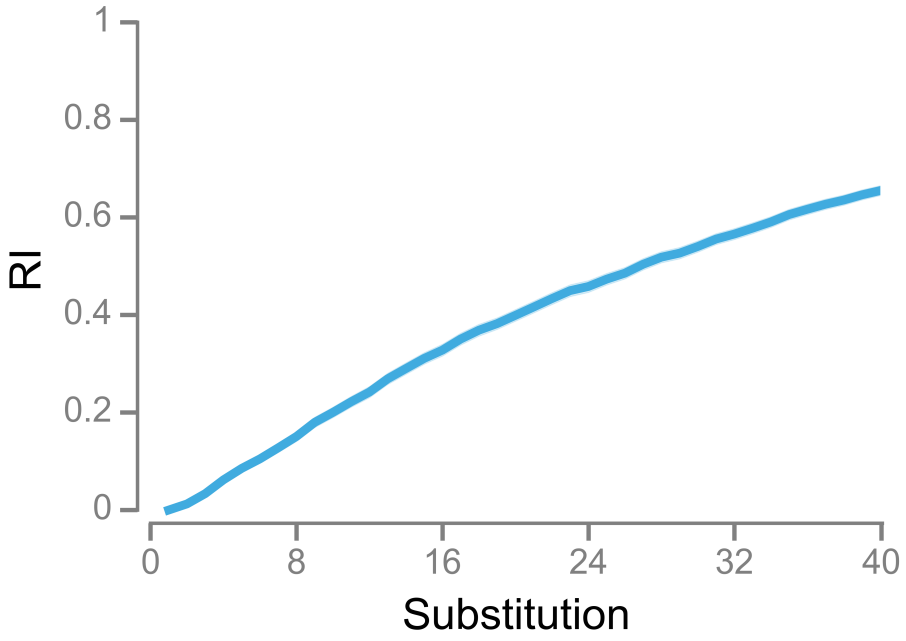

Figure 2.15: Reproductive isolation (RI) does not snowball in the RNA-folding model. Values are means of 10^3 stochastic simulations. Shaded regions indicate 95% CIs.

found in many organisms (e.g., Gourbière and Mallet 2010; Giraud and Gourbière 2012). This slowdown was caused by the fact that RI increased slower than linearly with the number of both simple and complex DMIs (Figure 2.16). Thus, DMIs did not act independently of each other on RI. One likely reason for this non-independence is that the total number of DMIs (simple and complex) among highly diverged sequences is high enough that a substantial fraction of individual sites must participate in multiple DMIs (Figure 2.12).

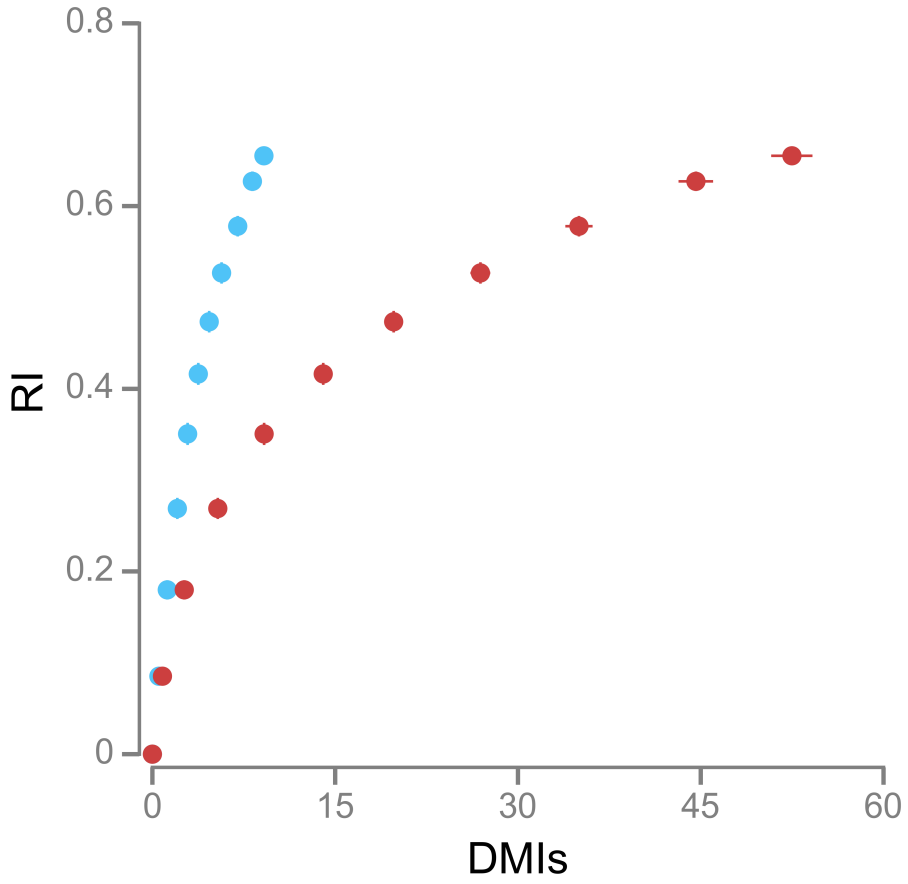

Figure 2.16: The numbers of simple (blue) and complex DMIs (red) are not linearly related to RI. Values are means of 10^3 simulations for $k = 0, 4, 8, \ldots, 40$. Error bars are 95% CIs.

2.3.7 The fitness landscape influences the parameters of the melting snowball model

Figure 2.9 shows two striking patterns about the parameters of the melting snowball model. First, p and q were strongly positively correlated with each other (Spearman's rank correlation coefficient: $\rho = 0.466$, $P < 10^{-6}$), indicating that the origination and complexification of simple DMIs are not independent. Second, the parameters varied extensively between simulations. What caused this variation? All simulations took place on the same sequence space, but with different fitness landscapes. Since all fitness landscapes were "holey" (Gavrilets 2004), it follows that the exact pattern of "holeyness" might have had an effect on the evolutionary dynamics. One component of the holeyness of a fitness landscape is the proportion of inviable single mutant neighbors of all the sequences generated during the course of evolution. This measure of the local holeyness of the fitness landscape was strongly positively correlated with both p and q ($\rho = 0.338$ and 0.210, respectively; both, $P < 10^{-6}$) (Figures 2.17A and 2.17C).

What determines holeyness? Fitness landscapes in my RNA-folding model have two determinants: the reference structure and the value of α (Equation 1.3). RNA secondary structures can differ in many ways, such as the number and size of base pair stacks, interior loops, and hairpin loops (Schuster *et al.* 1994). For a given reference structure, lower values of α are expected to specify fitness landscapes with more inviable sequences (i.e., holes) in them. To test whether these determinants of the fitness landscape influence holeyness, I ran 10^3 independent

evolutionary simulations at each of another four values of α. I found that holeyness was influenced by both determinants of the fitness landscape (Figure 2.18): it was positively correlated with the number of base pairs in the reference sequence ($\rho = 0.184$; $P < 10^{-6}$) and negatively correlated with α ($\rho = -0.583$; $P < 10^{-6}$).

Changing α did not affect the patterns of accumulation of simple and complex DMIs qualitatively (Figure 2.19). Interestingly, α was strongly positively correlated with both p and q (Figures 2.17B and 2.17D): the semi-partial rank correlation coefficient when the effect of holeyness was removed from α were $\rho = 0.282$ for p and $\rho = 0.301$ for q (both, $P < 10^{-6}$). This result is counterintuitive because α was negatively correlated with holeyness, which in turn was positively correlated with both p and q. I conclude that the parameters of the melting snowball model were influenced independently by both holeyness and α.

2.3.8 $\mathcal{P}_1$ and Welch's paradox

If all DMIs are simple and individual loci are at most involved in one DMI, then the proportion of the fixed differences between species where an allele from one species is deleterious in another species, $\mathcal{P}_1$, is expected to increase linearly with genetic distance (Equation 2.1; Welch 2004). This prediction is contradicted by the observation that $\mathcal{P}_1$ is approximately constant over large genetic distances (Kondrashov *et al.* 2002; Kulathinal *et al.* 2004)—a result I call Welch's paradox (Welch 2004). My results contradict both assumptions behind the prediction that $\mathcal{P}_1$ should increase linearly with genetic distance: most DMIs are complex, and

individual loci are involved in multiple DMIs. These effects are expected to act in opposite directions: the former would cause $\mathcal{P}_1$ to increase faster than linearly with k, whereas the latter would cause $\mathcal{P}_1$ to increase slower than linearly with k. In the RNA-folding simulations, $\mathcal{P}_1$ increased with divergence but did so slower than linearly (Figure 2.20), indicating that the lack of independence between DMIs dominates the evolution of $\mathcal{P}_1$. These results suggest a possible resolution for Welch's paradox: $\mathcal{P}_1$ can be constant even if DMIs snowball if individual loci participate in multiple DMIs. Alternative resolutions of Welch's paradox have been proposed (e.g., Fraïsse *et al.* 2016).

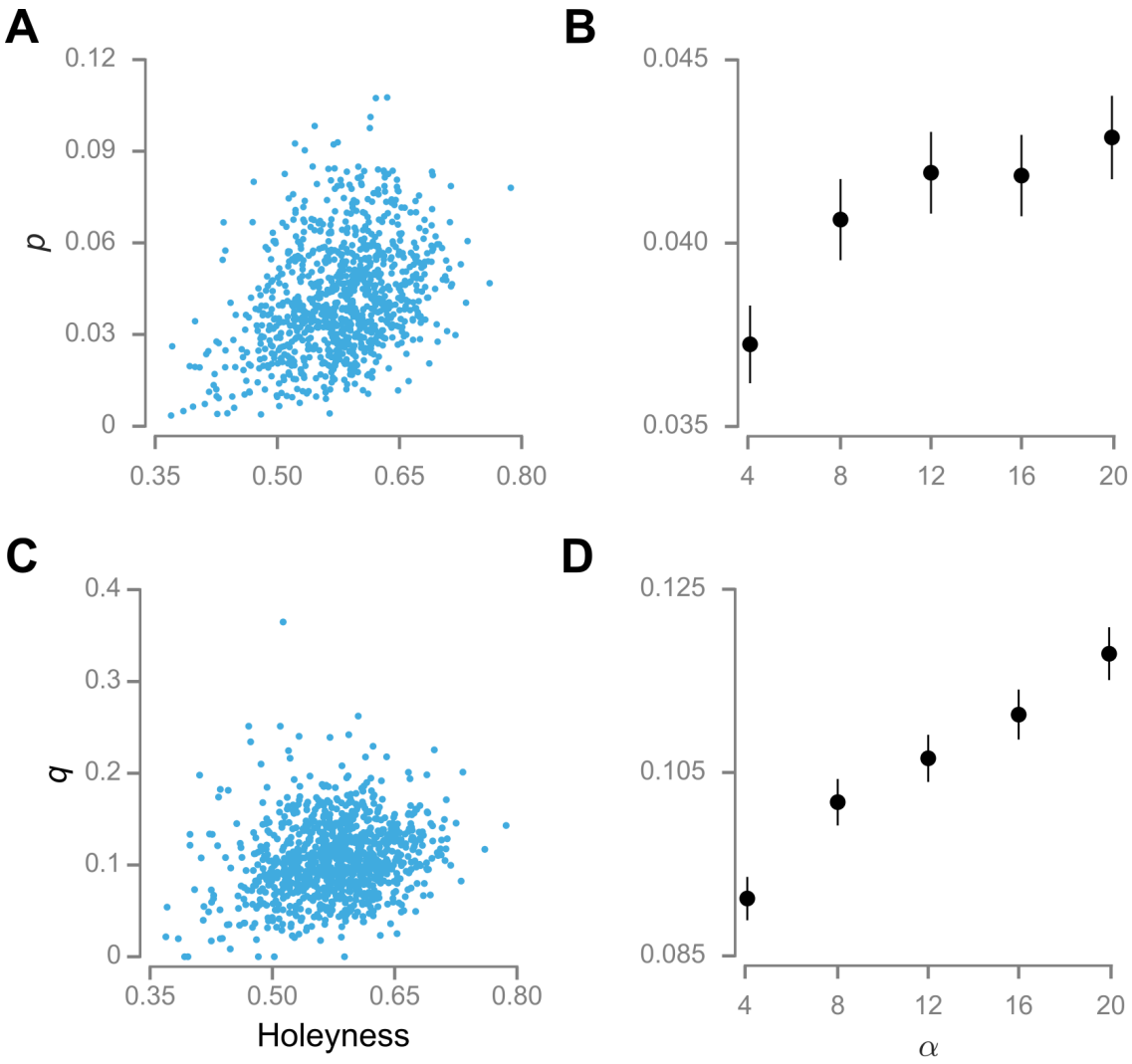

Figure 2.17: The fitness landscape influences the parameters of the melting snowball model. (A, C) Both parameters are positively related to the local holeyness of the fitness landscape. Values are individual estimates of p and q for each of 10^3 RNA-folding simulations for $\alpha = 12$. (B, D) Both parameters are positively related to α. Values are means of 10^3 RNA-folding simulations for each value of α. Error bars are 95% CIs.

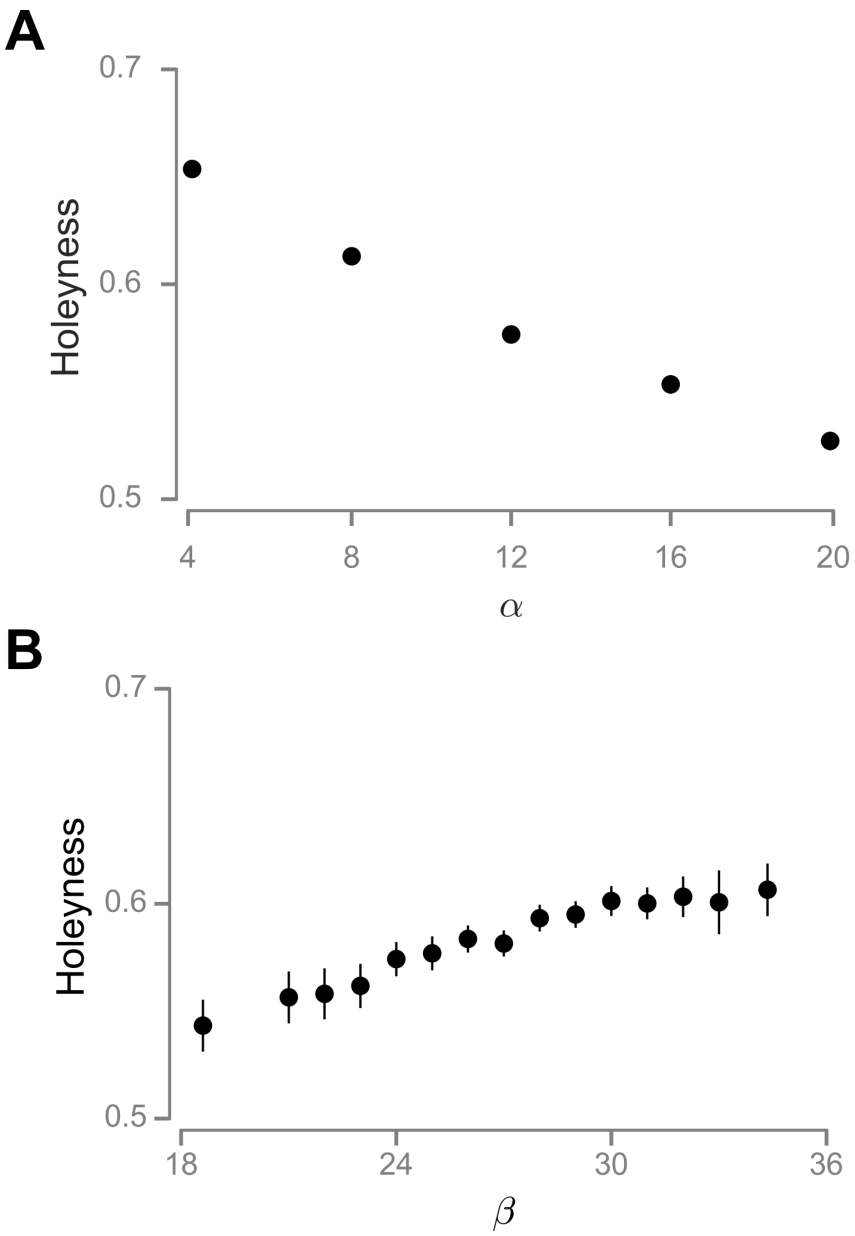

Figure 2.18: Holeyness decreases with the value of α (A) and increases with the number of base pairs, β, in the reference sequence (B). (A) Values are means of 10^3 RNA-folding simulations for each value of α. (B) The holeyness data from the 5×10^3 simulations used in (A) were grouped by individual values of β. I pooled estimates for $\beta \leqslant 20$ and for $\beta \geqslant 34$. The resulting β groups have sample sizes ranging from 120 to 581. Error bars are 95% CIs. The error bars in (A) are covered by the points.

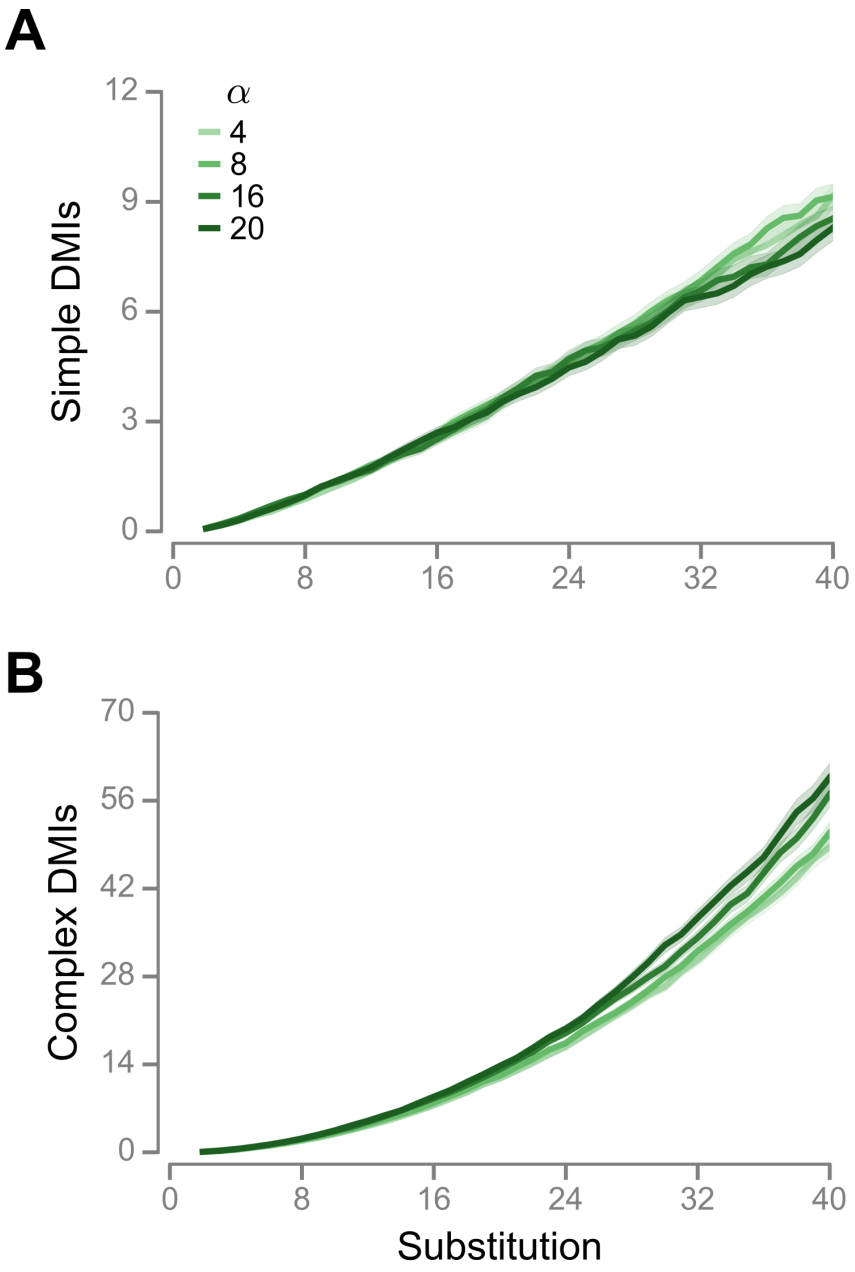

Figure 2.19: Simple DMIs accumulate more slowly (A) and complex DMIs accumulate faster (B) as α increases. The number of complex DMIs was calculated as in Figure 9B. Values are means of 10^3 RNA-folding simulations for each value of α. Shaded regions indicate 95% CIs.

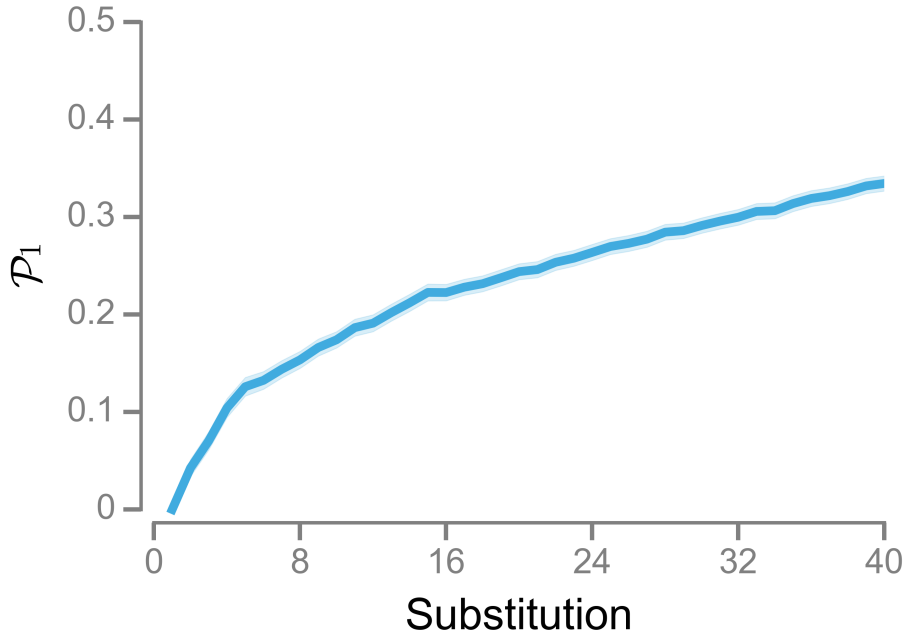

Figure 2.20: Evolution of the proportion of single introgressions involved in a DMI, $\mathcal{P}_1$, as populations diverge in the RNA-folding model. Values are means of 10^3 stochastic simulations. Shaded region indicates 95% CIs.

Chapter 3

How do populations affect the accumulation of incompatibilities?

> Perhaps a species is like a toy figure made of rubber which can be pulled in to all sorts of shapes without losing its cohesion. [...] To me it seems that all the available evidence indicates just the opposite.
>
> Mayr (1949)

3.1 Segregating DMIs

The genetics of speciation has been generally described as the process by which genetic factors, which are otherwise benign within the one species, become detrimental in the genetic background of some other species (Orr 2001; Turelli *et al.* 2001; Masly *et al.* 2006; Maheshwari and Barbash 2011). The "monomorphic" model, as described in section 2.2.1, operates in accordance with this definition, thus allowing each neutral substitution that arises to go fixation. This type of evolutionary regime, known as strong selection weak mutation (SSWM) is a valid approach to simulate evolution as a series of beneficial mutations going to fixation (Sniegowski and Gerrish 2010), but it should take $\approx 2N$ generations for a neutral mutation, akin to the ones arising in a holey landscape, to go to fixation (Kimura 1962). This issue will not be problematic if one is solely interested in studying the incompatibilities between the fixed alleles from two diverging lineages. However, recent studies have presented us with an inconvenient and yet intriguing reality: incompatibilities are segregating within species (Seidel *et al.* 2008; Corbett-Detig *et al.* 2013; Hou *et al.* 2014; Chae *et al.* 2014).

Let us focus on a pioneering study by Seidel *et al.* (2008), which demonstrates that incompatibilities can segregate within the genetic boundary of a single species. Seidel *et al.* (2008) use two strains of the nematode *Caenorhabditis elegans*, one from Hawaii and the other from Bristol, and created recombinants between the two. They show that crossing the F_1 males with the Hawaii hermaphrodites results in half of the embryos dying. The cause of the lethality

is the Bristol *peel-1* allele (paternal effect epistatic embryonic lethal - 1) that encodes a non-functional transmembrane protein, PEEL-1. PEEL-1 protein causes defects in muscle and epidermic tissues during embryogenesis and its production is suppressed in the Bristol strain by ZEEL-1 protein, which is encoded by *zeel-1* (zygotic epistatic embryonic lethal - 1) (Seidel *et al.* 2011). The Hawaii strain does not suffer from this predicament since it has a 19kb deficiency in place of the *peel-1* and *zeel-1* elements. Half of the sperm produced by F_1 males that carry *peel-1*Δ and *zeel-1*Δ (i.e., the deficient Hawaii versions of *peel-1* and *zeel-1* respectively) nevertheless inherit PEEL-1 protein from the F_1 males. The presence of this protein in the zygote without a functional *zeel-1* results in hybrid inviability (Figure 3.1).

Since it is known that outcrossing occurs between different *C. elegans* populations (Barrière and Félix 2005), one would expect the segregating DMI to be present in most of the strains. Seidel *et al.* (2008), using 62 different strains from 40 locations, show that this incompatibility is present at a global scale.

What are the implication of segregating DMIs? Do they, as Cutter (2011) suggests, contribute to the emergence of a burgeoning RI that would eventually result in incipient species as bona fide DMIs accumulate? Corbett-Detig *et al.* (2013) at the beginning of their paper on the prevalence of segregating incompatibilities in *Drosophila melanogaster*, suggest that the segregating incompatibilities provide an alternative to the Dobzhansky–Muller model of incompatibilities. But at this point, we simply do not know enough about incompatibilities segregating within populations to evaluate their relevance vis-à-vis speciation.

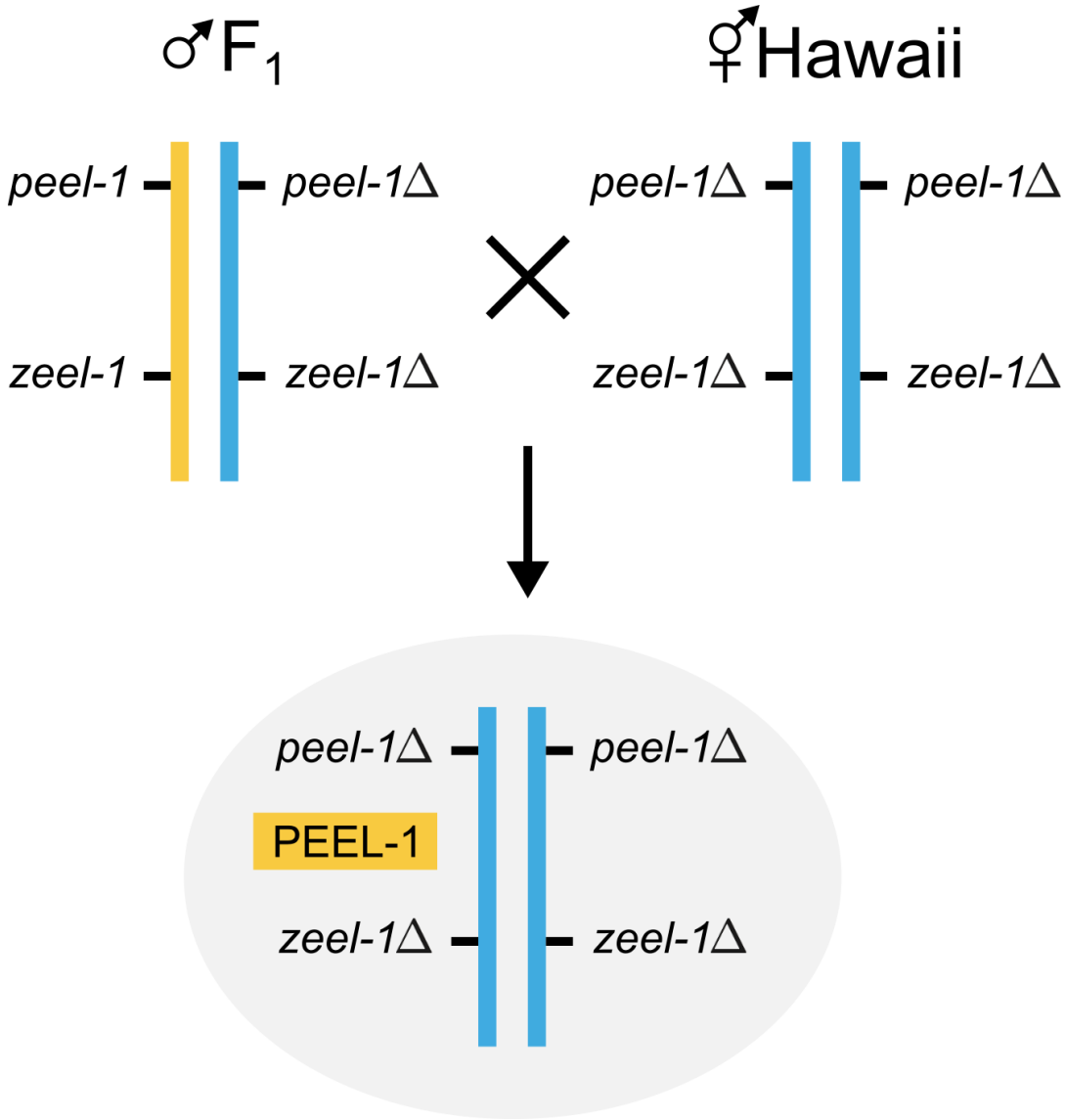

Figure 3.1: The segregating DMI in *C. elegans* occurs when ♂ F_1 which carries the Bristol chromosomes (yellow) and the Hawaii chromosome (blue) is crossed with ♀ Hawaii. The resulting zygote lacks active *zeel-1* to suppress the Bristol *peel-1* product (PEEL-1). ♀ F_1 × ♂ Hawaii does not result in hybrid lethality, presumably because PEEL-1 is dosage dependent and only sperm delivers enough PEEL-1 to cause incompatibility (Seidel *et al.* 2011).

The presence of segregating DMIs can also be influenced by recombination within a population and gene flow between populations. With recombination, one would expect to see a reduction in the number of segregating incompatibilities within population, since genotypes within a population recombine, they can potentially bring segregating incompatibilities together and selection will then purge these incompatibilities from the population.

The gene flow should, as Wang *et al.* (2015) point out, decrease the likelihood of incompatibilities arising between diverging populations. Kondrashov (2003) predicts that in a spatially structured population the number of DMIs should, under certain conditions, accumulate linearly. He argues that for complex DMIs to emerge multiple loci that can negatively interact with each other should arise and fix at the same time, an event that should be unlikely in the presence of gene flow.

I investigate these questions using an individual-based model.

3.2 Methods

3.2.1 The individual-based model

The initial step is identical to my monomorphic model, as described in section 2.2.1: I start from a random 100 nucleotide RNA sequence, henceforth referred to as the reference sequence. The fitness of any RNA sequence during simulation is

calculated relative to the reference sequence, according to Equation 1.3. The reference sequence undergoes 200 random neutral substitutions in succession. The resulting sequence is used as the ancestral sequence. The ancestral population consists of N individual ancestral sequences, where N is the population size (Figure 3.2). All the results presented in this section are based on 1000 simulations using the same reference sequences as in Table 2.1, $\alpha = 12$, and population size of $N = 1000$.

The ancestral sequence is used to found two identical haploid populations. At each generation, both populations recombine and mutate.

For a population of size N, I randomly sample two sets of $\frac{N}{2}$ sequences with replacement from the population and generate N recombinants (Figure 3.3). Two genotypes can undergo as many as $L - 1$ crossover events between each other with probability r per interval. r can vary from 0 (i.e., no recombination events) to 0.5 (i.e., free recombination between all loci). If no crossovers have taken place, the parental sequences are allowed to mutate, and then moved to the next generation.

I simulate mutation as a Bernoulli process where each site mutates according to the mutation rate per site per generation (u). All types of base-substitution mutations have equal probability. Insertions and deletions are not considered.

After recombination and mutation, I calculate the fitness of each sequence. The next generation is composed of viable genotypes after recombination and mutation.

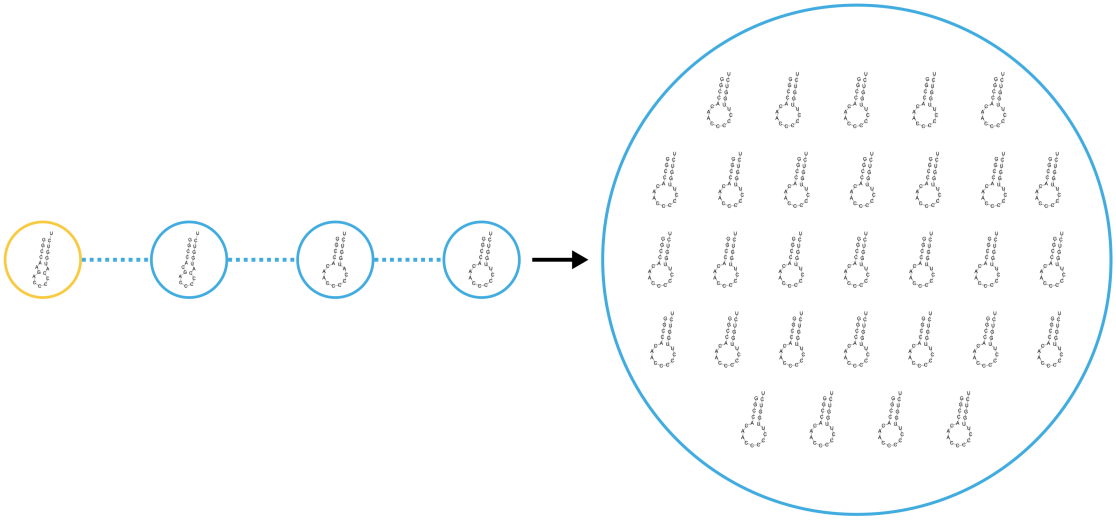

Figure 3.2: Generating the ancestral population in the individual-based simulation. A sequence is viable (yellow, blue circles) if its secondary structure both has more than $\alpha = 2$ base pairs and is at most $\alpha = 2$ base pairs away from the reference structure (yellow circle). Each population simulation starts with a burn-in period where a sequence with the reference structure undergoes 3 neutral substitutions (dashed blue lines). The resulting sequence is then used to create the ancestral population of size $N = 30$.

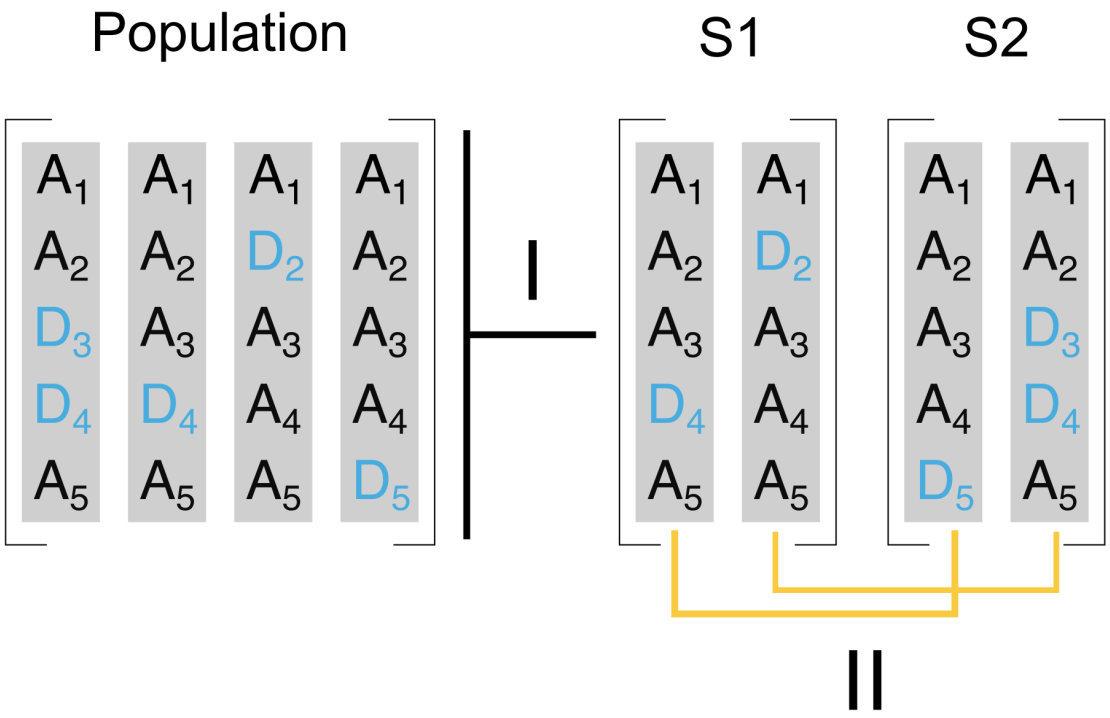

Figure 3.3: To recombine sequences, we randomly sample with replacement two sets of $N/2$ sequences from population of size N (I). Then we recombine sequences between two the random samples in order, i.e., the first sequence from $S1$ recombine with the first sequence from $S2$, the second sequence from $S1$ with the second from $S2$, and so on (II). Since each recombination event results in two recombinants, this approach results in N recombinants.

3.2.2 How to find DMIs

Finding DMIs at the population level requires certain modifications to my original approach, as outlined in section 2.2.2. In the individual-based model, populations can be polymorphic, mutations arise according to a mutation rate and a site can undergo multiple mutations, which makes distinguishing between ancestral and derived alleles methodologically impossible. For this reason, my approach to identify simple DMIs is not applicable to my population model.

A pure introgression approach, as used in 2.2.2 to identify complex DMIs, is more suited to the population level model. If the introgression of $1 < i < k$ alleles (e.g., $1 \xrightarrow{A,B,\ldots} 2$) results in an inviable genotype, but all the introgressions of each individual allele and of any combination of between 2 and $i-1$ of the alleles result in a viable genotype, then i alleles participate in a DMI of order $n \geqslant i+1$. I use single, double, and triple introgressions to estimate the number of DMIs between any two sequences. This introgression based approach can overestimate the number of DMIs, so I calculate the number of incompatibilities in one direction only, i.e., introgressing from population 1 to population 2. To limit my estimation of the number of incompatibilities to the more prevalent genotypes in the population, I only include the most common genotypes from each population in this analysis.

To find segregating DMIs within a population, I use the same introgression approach, but this time only between pairs of genotypes within the population that each has a frequency $\geqslant 0.05$. This arbitrary cutoff is reasonable given the

population size of 1000 since it excludes the rare genotypes form the assays.

3.2.3 Ancestral polymorphism

In order to start with a polymorphic ancestral population instead of a monomorphic one at the beginning of each simulation, I allow the population composed of solely ancestral sequences to evolve with mutation rate u and no recombination. After it reaches the desired level of polymorphism, calculated using Equation 3.1, I use the resulting population as the ancestral population for the simulation.

3.2.4 Gene flow

To simulate gene flow between the two populations, I allow symmetric migration between the two populations according the migration rate (m). At each generation, after mutation, recombination, selection with replacement, I allow n migrants to move between two populations, where n is drawn from a Poisson distribution with parameter m.

3.2.5 Gene diversity

For any locus, gene diversity is defined as:

$$H = 1 - \sum_{i=1}^{\alpha} {p_i}^2 \quad , \tag{3.1}$$

where p_i is the frequency of allele i and α is the number of alleles. The average H was calculated over all the loci.

3.2.6 Hamming distance

I calculate the Hamming distance between population 1 and population 2 by counting the number of sites at which the most prevalent genotype from 1 differs from its counterpart from 2.

3.2.7 Proportion of inviable sequences

After the parental genotypes have undergone recombination and mutation in order to create the next generation, I count the number of inviable sequences and divide that number by the population size N.

3.2.8 Reproductive isolation

I use two different measures of RI:

Maximum RI: To calculate the maximum level of RI, I generate all possible 198 recombinants resulting from a single crossover between the most common sequence in population 1 and its counterpart in population 2. I then calculate RI using Equation 2.2.5.

Recombination RI: At a given generation, we define recombination RI as:

$$\text{RI} = 1 - \frac{\overline{W}_H}{\overline{W}_S} \quad , \tag{3.2}$$

where $\overline{W}_S$ is the mean of the proportions of viable individuals within each population, and $\overline{W}_H$ is the proportion of N viable hybrids between the two populations. Hybrids were generated as offspring from pairs of made up of one viable individual from each population. Viability was assessed after recombination and mutation (section 3.2.1). Recombination RI is expected to be zero in asexual populations and increase with r.

3.3 Results

3.3.1 Incompatibilities snowball in the individual-based model

The accumulation of incompatibilities in the individual-based simulations with no recombination, i.e., "asexual" case, may seem linear at first, but this is due to the non-linear pattern of divergence (Figures 3.4A and 3.4B). The number of incompatibilities in the asexual individual-based simulations snowballs in a similar fashion to the monomorphic model with the same level of divergence, measured by the Hamming distance (Figure 3.4C). I allowed sites in the monomorphic model to undergo multiple substitutions, making it more comparable to the individual-based model. The accumulation of incompatibilities in the asexual individual-based simulations is robust to a wide range of population sizes, resulting in comparable levels of RI (Figure 3.4E)

3.3.2 Recombination slows down the snowballing of incompatibilities

The monomorphic model assumes that a population is at most composed of two genotypes differing by a single mutation. In such a population, recombination would have no effect on the evolutionary dynamics. The populations described in the previous section do not meet the assumptions of the monomorphic model; even the smallest populations are able to maintain considerable genetic variation (Figure 3.4D). To test the extent to which recombination would affect the results, I evolved populations of $N = 1000$ individuals with different levels of recombination.

The lowest recombination rate used in these simulations ($r = 0.001$) behaves similar to the asexual case, despite the fact that even at this low recombination level, approximately 10% of individuals produced every generation result from one or more crossover events among parental sequences.

Higher recombination rates ($r = 0.01$ and 0.1) lower the level of divergence between populations (Figure 3.5A). Interestingly, the reduction in sequence divergence at higher recombination rates not only reduces the total number of incompatibilities accumulated by the end of 1000 generations of evolution, it also produces fewer incompatibilities given the same level of divergence when compared to the asexual and low recombination cases (Figure 3.5C). Using AIC (see page 33) to compare the support for a linear or a snowball model shows that higher recombination rates decrease support for the snowball model. However,

it does not appear to increase support for the linear model. Rather, recombination increases the proportion of runs that provide equal support for the linear and snowball models. This result is consistent with a decrease in the statistical power to distinguish the two models due to the lower levels of divergence and smaller numbers of DMIs as recombination increases (Figure 3.6).

Higher recombination rates result in fewer DMIs and a decrease in the maximum RI (Figure 3.5E). However, this observation is misleading since in populations with low recombination only a few hybrids would actually be inviable. Higher recombination rates, while reducing the total number of DMIs, increase the level of recombination RI (Figure 3.5F). Recombination RI is a more realistic measure of RI since it reflects the actual proportion of inviable hybrids.

3.3.3 Recombination suppresses the emergence of segregating DMIs

DMIs do segregate within populations in the asexual individual-based simulations, but recombination rate suppress the emergence of segregating DMIs (Figure 3.5D). Why would recombination act against the emergence of segregating DMIs? Recombination can bring segregating DMIs together and result in selection against the incompatible combinations. To test whether recombination is

causing selection against segregating DMIs, I measured the proportion of inviable individuals produced each generation. Initially, populations are monomorphic so all mortality is caused by mutation. After 200 generations, however, populations are genetically diverse (Figure 3.5B), so mortality results from a mixture of mutation and recombination. Populations with higher r show increased mortality at that point despite having less genetic variation. This result is consistent with the hypothesis that recombination increases the strength of selection against segregating DMIs (Figure 3.7).

Why does recombination slow down the snowball? One possibility, following Cutter (2011), is that segregating DMIs within populations contribute to the accumulation of DMIs between populations. By selecting against segregating DMIs, recombination would suppress the accumulation of DMIs between populations. However, my results do not support this hypothesis. Populations evolving with a recombination rate of $r = 0.001$ accumulated DMIs at approximately the same rate as asexual populations, despite having a dramatically reduced number of segregating DMIs.

Another possibility is that recombination results in selection for mutationally robust genotypes, and genotypes that are more resistant to mutations are also more resistant to introgressions, which reduces the number of DMIs. The populations with the highest r have lower overall mortality at generation 1000 than those with the lowest r (Figure 3.7). This is remarkable because populations with the highest r experience two sources of inviability (mutation and recombination), whereas those with the lowest r experience mostly one source of inviability (no

more than $\sim$ 10% of inviable individuals have experienced recombination). Thus, these results indicate that the populations with the highest r are more mutationally robust than the populations with the lowest r, in agreement with the robustness hypothesis.

3.3.4 Ancestral polymorphism affects the accumulation of DMIs but not that of segregating DMIs

To test the robustness hypothesis more directly, I pre-evolve populations without sex so that they accumulate different numbers of segregating DMIs. I then used these polymorphic populations as the ancestors in individual-based simulation with high recombination ($r = 0.1$). If the robustness hypothesis is true one would expect the presence of segregating DMIs in the pre-evolved ancestral populations would result in stronger selection for mutationally robust genotypes, suppressing the accumulation of DMIs among populations in the process.

My results confirm my prediction (Figure 3.8C). Using AIC to compare support for a linear or a snowball model shows that an increase in ancestral polymorphism reduces the proportion of runs that fit the snowball model (Figure 3.9). However, there does not appear to be increased support for the linear model. Again, this is likely caused by a decrease in statistical power.

Given that the ancestral population undergoes mutation without recombination to reach the desired level of polymorphism, segregating DMIs can be found at the beginning of divergence, but these segregating DMIs are purged from the

population in the early stages of divergence (Figure 3.8C). This last piece of evidence suggests that recombination quickly removes segregating DMIs from the populations.

3.3.5 Gene flow slows down the snowballing of incompatibilities but does not eliminate them entirely

One limitation of my model is that DMIs involving neutral alleles are not expected to result in enduring RI in the face of gene flow (Bank *et al.* 2012). To test this prediction I ran individual-based simulations with high recombination with low ($m = 0.01$) and high ($m = 1$) migration rates.

Introducing high level of gene flow to individual-based simulation with high recombination results in lower levels of divergence (Figure 3.10A), which is expected consequence of gene flow. In addition, gene flow slows down the snowballing of incompatibilities, but it does not eliminate DMIs completely (Figure 3.10C). This is also consistent with the robustness hypothesis because gene flow can also generate selection for mutational robustness (Proulx and Phillips 2005).

Using AIC to compare support for a linear or a snowball model shows that high gene flow reduces the proportion of runs that fit the snowball model, but it does not appear to increase support for the linear model (Figure 3.11). Again, this result is likely caused by a decrease in statistical power.

High levels of gene flow did not result in the emergence of segregating DMIs

in my simulations (data not shown).

3.3.6 Higher mutation rates result in more robust genotypes

The robustness hypothesis would predict that the higher the mutation rate, the more robust the resulting genotypes, and this would lead to fewer DMIs. To test this prediction, I ran asexual individual-based simulations at varying levels of mutation rates and then compared the number of incompatibilities at the same levels of divergence (measured via the Hamming distance). More robust populations accumulate fewer DMIs (Figure 3.12). This pattern disappears if I include the number of DMIs from double and triple introgressions (data not shown). This observation can be attributed to the fact that rarely is a sequence introduced to multiple mutations at once, and thus selection is mainly operating on robustness in the face of single introgressions.

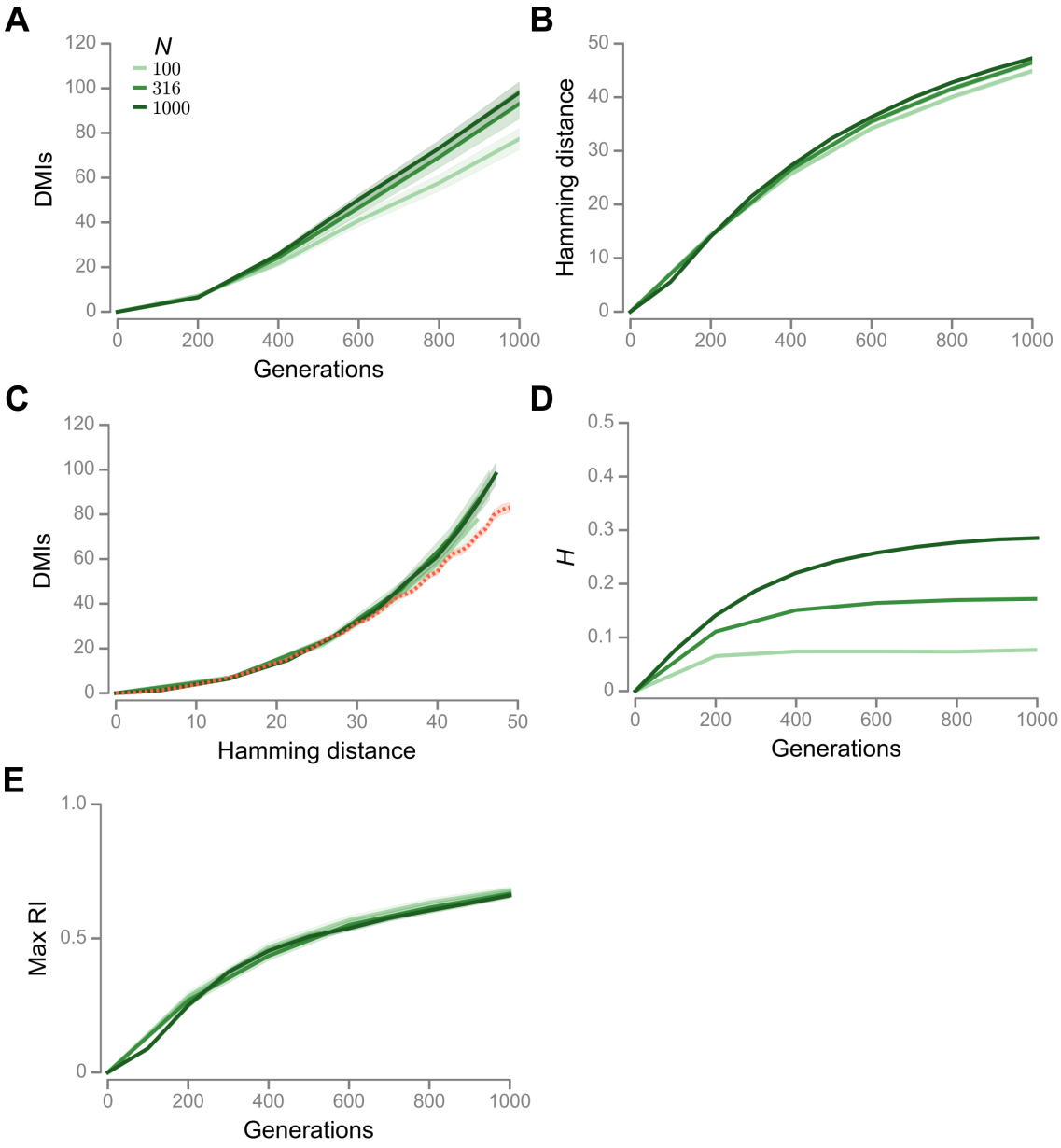
A
N
100
316
1000
DMIs
Generations
B
Hamming distance
C
DMIs
Hamming distance
D
H
Generations
E
Max RI
Generations

Figure 3.4 ***(previous page)*****:** Incompatibilities snowball in the asexual individual-based simulation. The number of incompatibilities between diverging populations after 1000 generations may seem linear (A) but this is caused by the non-linear pattern of divergence, as measured by the Hamming distance (B). Plotting the number of DMIs against the Hamming distance shows that incompatibilities snowball in the asexual individual-based simulations similar to the monomorphic model with multiple hits (red, dashed) and independent of the population size (C). Given the difference sizes, populations reach different levels of heterozygosity (D). The asexual individual-based simulations with different population sizes reach similar levels of maximum RI (E). Each trajectory is based on 1000 individual-based simulations each with $u = 10^{-3}$, and no recombination with $N = 100$, 316, and 1000, assayed every 200 generations. The monomorphic trajectory is based on 1000 simulations of the monomorphic model with multiple hits. Shaded regions indicate 95% CIs.

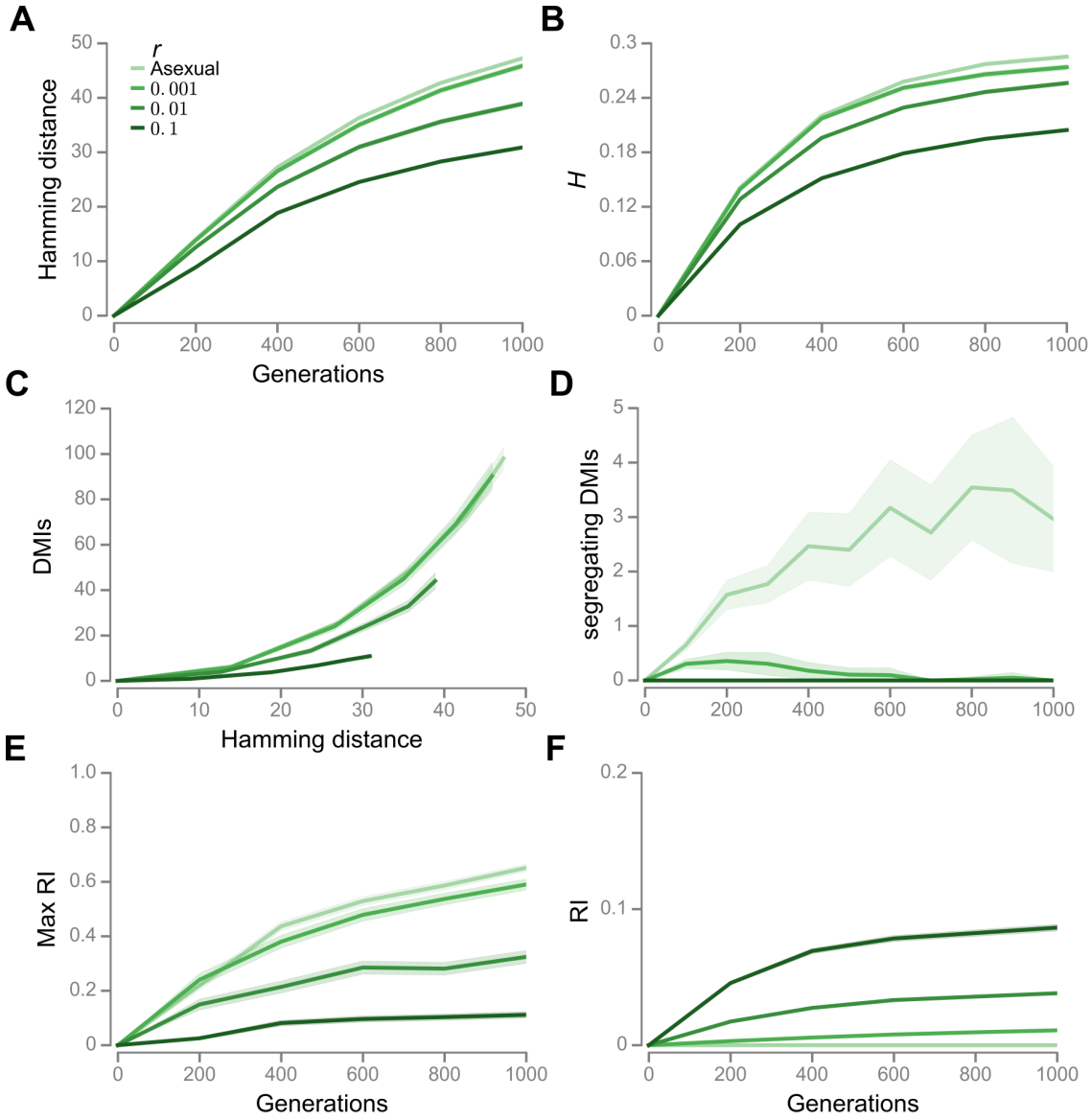
A
B
C
D
E
F
r
Asexual
0.001
0.01
0.1
Hamming distance
Generations
H
DMIs
Hamming distance
segregating DMIs
Max RI
Generations
RI
Generations

Figure 3.5 ***(previous page)*****:** Recombination can slow down the snowballing of DMIs in the individual-based model. The level differentiation between diverging lineages, captured by the Hamming distance (A), the average heterozygosity (B) are both affected by recombination. The accumulation of incompatibilities slows than as recombination rate is increased (C). Recombination suppresses the emergence of segregating DMIs (D). Interestingly while the maximum RI, calculated by generating all the single cross-over recombinants between the most prevalent sequences in each lineage, decreases as recombination rate increases (E). Recombination RI shows the opposite pattern (F). Each trajectory is based on 1000 individual-based simulations each with $N = 1000$, $u = 10^{-3}$, and the specified recombination rate. Runs were assayed every 200 generations. Shaded regions indicate 95% CIs.

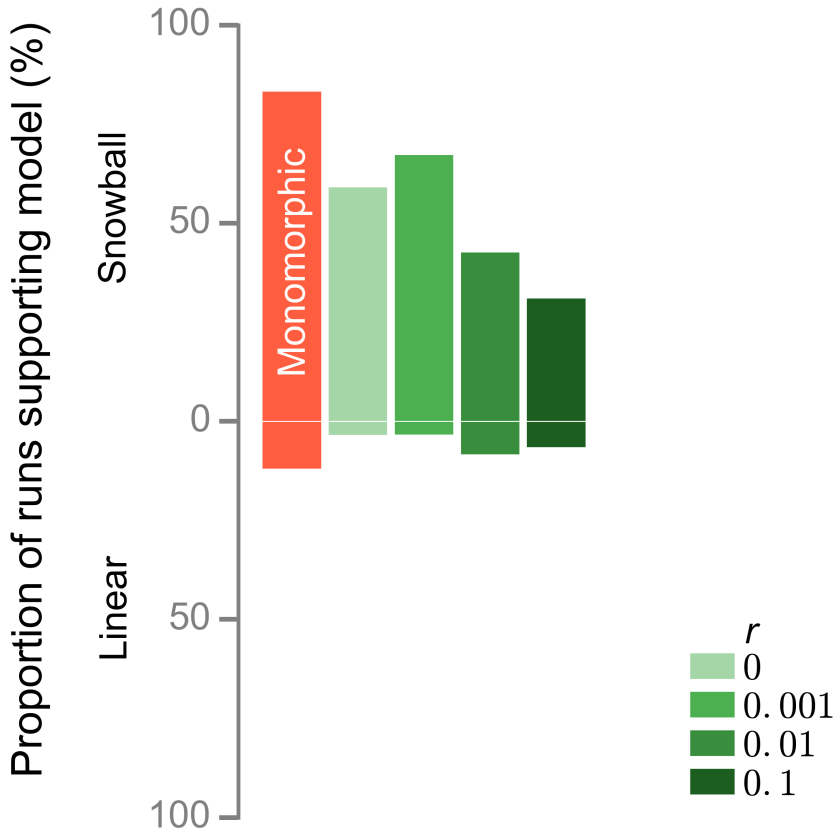

Figure 3.6: Recombination affects the snowballing of incompatibilities. I fit a linear model ($I_k = bk$) and a snowball model ($I_k = bk^2$) to each run of individual-based simulation. I use AIC to compare the support for each model. Positive bars show the proportions of runs providing stronger support for the snowball model; negative bars show the proportions of runs providing stronger support for the linear model; the proportions of runs providing approximately equal support for both models are not shown. Different colors indicate different recombination rates. Each trajectory is based on 1000 individual-based simulations as shown in Figure 3.5. The monomorphic runs are identical to the ones used in Figure 3.4C. Due to fewer data points from the individual-based simulations relative to the monomorphic model (i.e., individual-based simulations are assayed every 200 generations whereas the monomorphic simulations are assayed after every substitution that increase the Hamming distance) only the results for ΔAIC$= 2$ are shown.

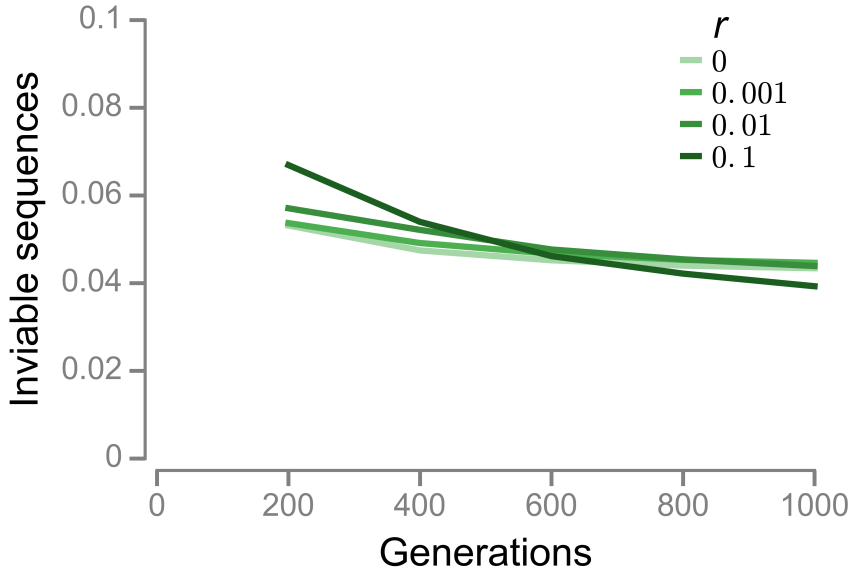

Figure 3.7: Recombination load results in an increase in the proportion of inviable offspring (sequences) at the beginning of divergence. Values are the proportions of inviable sequences generated after parental sequences undergo recombination and mutation. Each trajectory is based on 1000 individual-based simulations as shown in Figure 3.5. Shaded regions indicate 95% CIs.

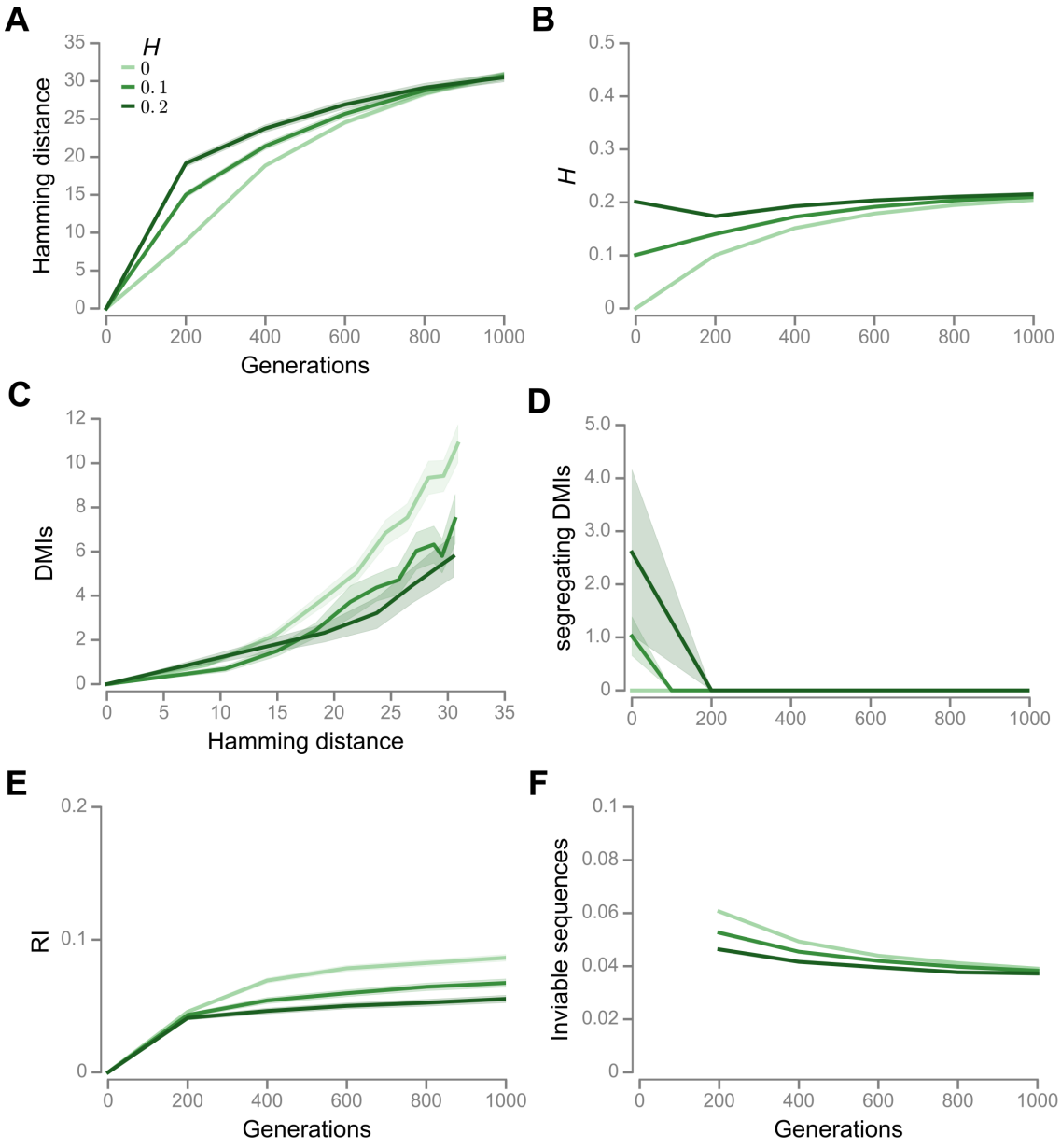
A
B
C
D
E
F
H
0
0.1
0.2
Hamming distance
Generations
H
DMIs
Hamming distance
segregating DMIs
RI
Generations
Inviable sequences
Generations

Figure 3.8 ***(previous page)*****:** Recombination load affects the way ancestral polymorphism influences the accumulation of DMIs within and between populations. The sequence divergence, measured by the Hamming distance (A) between evolving lineages and the level of heterozygosity converges irrespective of the ancestral polymorphism (B). The ancestral polymorphism affects the accumulation of DMIs (C). Given the initial polymorphism, populations start with segregating DMIs but they soon perish, as a result of recombination load (D). The levels of RI for different levels of ancestral polymorphism reflect the number of DMIs (E). The proportion of viable sequences for different levels of polymorphism and the monomorphic case (F). Each trajectory is based on 1000 individual-based simulations each with $N = 1000$, $u = 10^{-3}$, and $r = 0.1$. Runs were assayed every 200 generations. The desired ancestral polymorphism was achieved by allowing the ancestral population to evolve without mutation until it reached the desired level of polymorphism (see section 3.2.3). Shaded regions indicate 95% CIs.

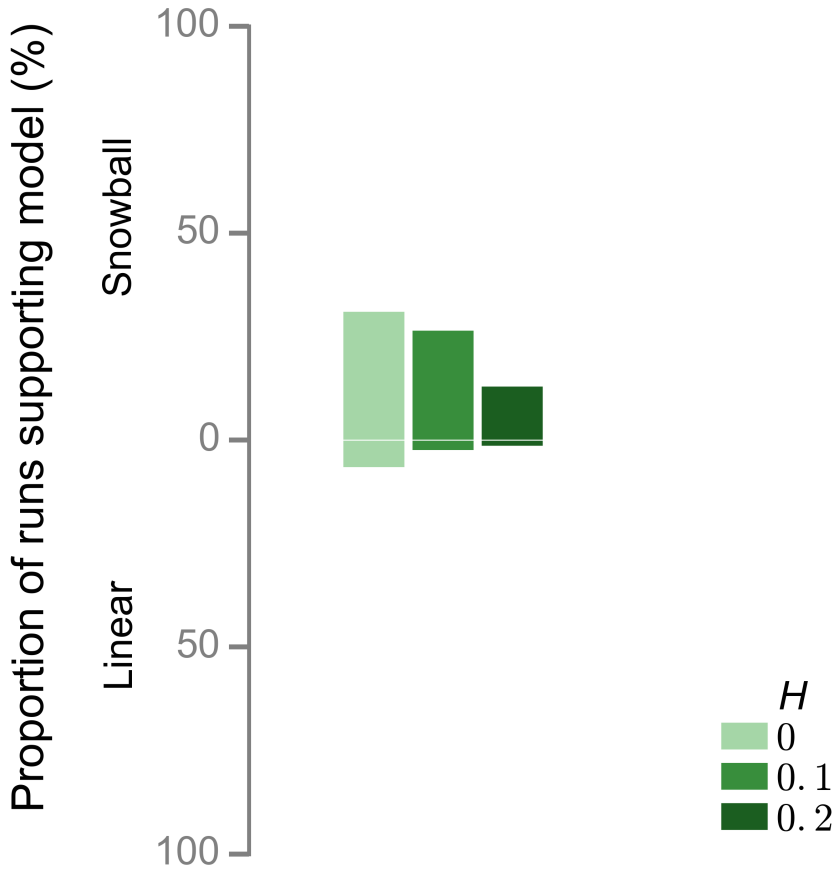

Figure 3.9: Ancestral polymorphism affects the snowballing of incompatibilities. I fit a linear model ($I_k = bk$) and a snowball model ($I_k = bk^2$) to each run of individual-based simulation. Positive bars show the proportions of runs providing stronger support for the snowball model; negative bars show the proportions of runs providing stronger support for the linear model; the proportions of runs providing approximately equal support for both models are not shown. Different colors indicate different levels of ancestral polymorphism. Each trajectory is based on 1000 individual-based simulations as shown in Figure 3.8. Only the results for ΔAIC$=2$ are shown.

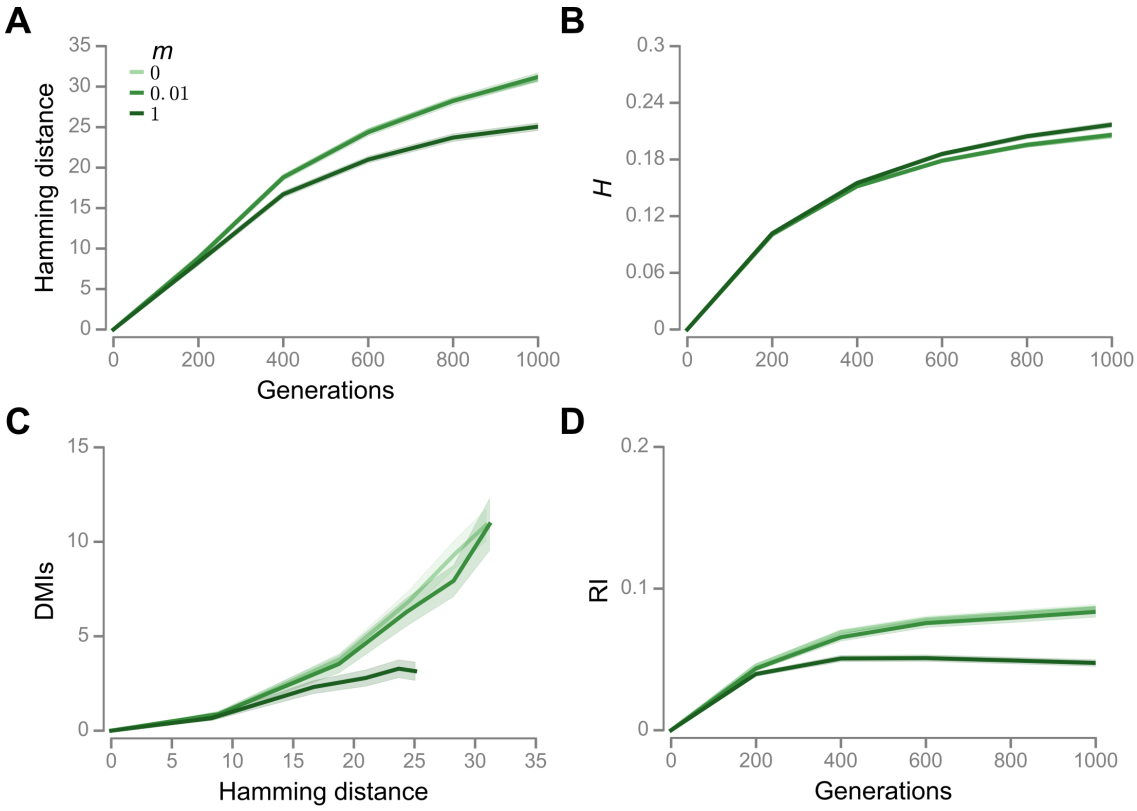

Figure 3.10: Gene flow can slow down the snowballing of incompatibilities but it does not eliminate them. The sequence divergence, measured by the Hamming distance (A) decreases as gene flow increases. Gene flow positively influences heterozygosity within population (B). High level of migration slows down the accumulation of DMIs but does not eliminate them entirely (C). The remaining DMIs provide a lower level of RI compared to low migration and no migration simulations (D). Each trajectory is based on 1000 individual-based simulations each with $N = 1000$, $u = 10^{-3}$, and $r = 0.1$. Runs were assayed every 200 generations. Symmetric migration is modeled by allowing a random set of individuals to moved between the two populations at each generation according to a Poisson distribution with parameter m. Shaded regions indicate 95% CIs.

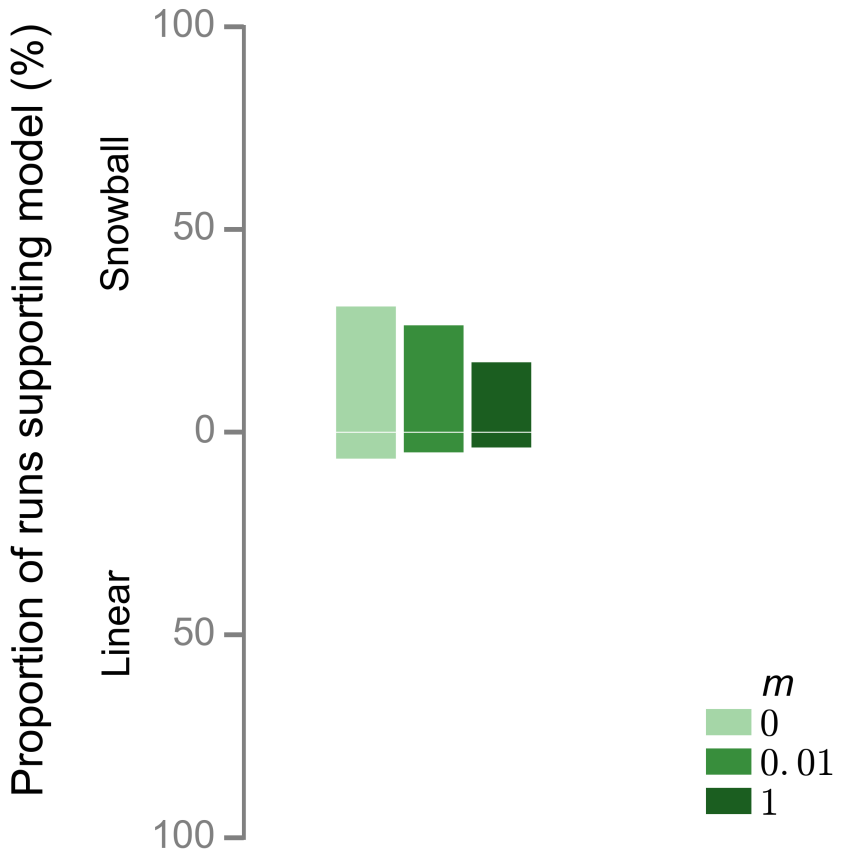

Figure 3.11: Gene flow affects the snowballing of incompatibilities. I fit a linear model ($I_k = bk$) and a snowball model ($I_k = bk^2$) to each run of individual-based simulation. Positive bars show the proportions of runs providing stronger support for the snowball model; negative bars show the proportions of runs providing stronger support for the linear model; the proportions of runs providing approximately equal support for both models are not shown. Different colors indicate different migration rates. Each trajectory is based on 1000 individual-based simulations as shown in Figure 3.10. Only the results for ΔAIC$= 2$ are shown.

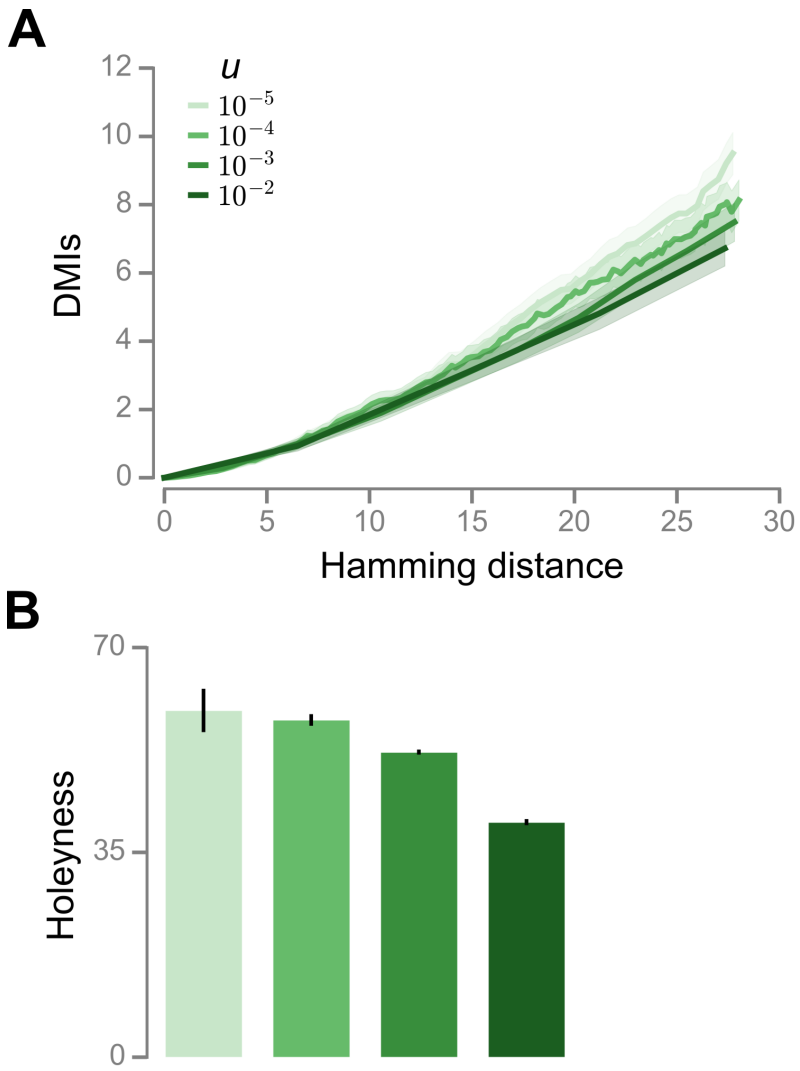

Figure 3.12: The robustness hypothesis predicts that higher mutation rates should result in more robust genotypes, i.e., less susceptible to introgressions. The accumulation of DMIs assayed via single introgressions shows this very pattern, where the simulations with lower mutation rates accumulate more DMIs compared to those with higher mutation rates (A). For each simulation, I measured the proportion of inviable sequences generated during the course of its evolution. Dividing the mean proportion of inviable sequences for each run by the mutation rate gives us the proportion of lethal mutations, i.e., holeyness. Higher mutation rates reduce holeyness, which means that they evolve more robust genotypes (B). Each trajectory is based on 1000 asexual individual-based simulations with $N = 100$. Runs were assayed every 50 generations. Shaded regions and bars indicate 95% CIs.

Chapter 4

Insights from studying speciation in a RNA world

> In triumph he returns to us, and brings us back this prize:
> To know what things can come about, and what cannot arise, and what law limits the power of each, with deep-set boundary stone.
>
> Lucretius, *De Rerum Natura*

4.1 What the RNA-folding model teaches us about DMIs

4.1.1 The RNA-folding model supported the central prediction of the snowball model

I tested both predictions and assumptions of the snowball model using a computational model of RNA folding. My results provide mixed support for the snowball model (Table 4.1).

Simple DMIs accumulated linearly, contrary to one of the main quantitative predictions of the snowball model (Orr 1995) (Figures 2.3 and 2.5A).

To elucidate why the snowball appeared to be "missing" from the RNA-folding simulations I tested two assumptions of the snowball model. First, that simple DMIs arise with constant probability, p. Although I did detect a decline in p (Figure 2.5B), it was not sufficient to account for the approximately linear pattern of accumulation of simple DMIs. Second, I tested the assumption that simple DMIs, once they have arisen, persist indefinitely. I found that this assumption was violated in the RNA-folding model. Instead, simple DMIs had a tendency to become more complex as further substitutions took place. I conclude that the snowball was "melting" for simple DMIs, not missing.

I proposed an extended snowball model incorporating the complexification of simple DMIs—the melting snowball. The RNA-folding simulations agree with

this model. In contrast to simple DMIs, the number of complex DMIs did snowball, in agreement with the prediction of the snowball model. In conclusion, the RNA-folding model supported the central prediction of the snowball model that the number of DMIs snowballs, but challenged some of its underlying assumptions. Despite the snowballing of DMIs, RI did not snowball because DMIs did not act independently of each other on RI.

Table 4.1: The monomorphic model provides mixed support for the snowball model

Test	Confirmed?
Prediction	
Simple DMIs snowball	No
Complex DMIs snowball	Yes
RI snowballs	No
Assumption	
Constant p with divergence	Yes, roughly
Simple DMIs persist indefinitely	No
Linear relationship between number of DMIs and RI	No

These results indicate that RI is a poor indicator for the number of DMIs in my model. Thus, the pattern of change in RI with divergence is unsuitable to test the snowball model (Johnson 2006; Gourbière and Mallet 2010; Presgraves 2010a).

An earlier test of the snowball model using a computational model of gene networks also found no evidence for a snowball effect in RI, and concluded that some assumptions of the snowball model were not met (Palmer and Feldman

2009). However, the extent to which the complexification of DMIs influenced their results is unclear because they did not attempt to investigate the dynamics of the DMIs underlying RI.

In one direct empirical test of the snowball effect, DMIs affecting pollen sterility were found to accumulate linearly, whereas DMIs affecting seed sterility were found to accumulate faster than linearly (Moyle and Nakazato 2010). My results suggest a possible explanation for the discrepancy: faster complexification (i.e., higher q) of pollen sterility DMIs. Sherman *et al.* (2014) found evidence of greater complexity of the DMIs involved in pollen sterility.

4.1.2 A possible resolution for Welch's paradox

If all DMIs are simple and individual loci are at most involved in one DMI, then the proportion of the fixed differences between species where an allele from one species is deleterious in another species, $\mathcal{P}_1$, is expected to increase linearly with genetic distance (Equation 2.1; Welch 2004). This prediction is contradicted by the observation that $\mathcal{P}_1$ is approximately constant over large genetic distances (Kondrashov *et al.* 2002; Kulathinal *et al.* 2004)—a result I call Welch's paradox (Welch 2004). My results contradict both assumptions behind the prediction that $\mathcal{P}_1$ should increase linearly with genetic distance: most DMIs are complex, and individual loci are involved in multiple DMIs. These effects are expected to act in opposite directions: the former would cause $\mathcal{P}_1$ to increase faster than linearly with k, whereas the latter would cause $\mathcal{P}_1$ to increase slower than linearly with k.

In the RNA-folding simulations, $\mathcal{P}_1$ increased with divergence but did so slower than linearly (Figure S11), indicating that the lack of independence between DMIs dominates the evolution of $\mathcal{P}_1$. These results suggest a possible resolution for Welch's paradox: $\mathcal{P}_1$ can be constant even if DMIs snowball if individual loci participate in multiple DMIs. Alternative resolutions of Welch's paradox have been proposed (e.g., Fraïsse *et al.* 2016).

4.1.3 Complex DMIs are more abundant in the RNA-folding model

I found that complex DMIs are more abundant than simple DMIs in the RNA-folding model. Complex DMIs have been discovered in many introgression studies (reviewed in Fraïsse *et al.* 2014). For example, Orr and Irving (2001) investigated the sterility of male F1 hybrids between the USA and Bogota subspecies of *D. pseudoobscura* and found that it is caused by an DMI between loci in both chromosomes 2 and 3 of USA and loci in at least three different regions of the X chromosome of Bogota—a DMI of order $n \geqslant 5$. More generally, high-order epistasis appears to be common (Weinreich *et al.* 2013; Kondrashov and Kondrashov 2015; Taylor and Ehrenreich 2015). However, the relative prevalence of simple and complex DMIs in nature is unclear because complex DMIs are more difficult to detect.

Two explanations for the abundance of complex DMIs have been proposed. First, that more complex DMIs evolve more easily than simpler DMIs because

they allow a greater proportion of the possible evolutionary paths between the common ancestor and the evolved genotypes containing the DMI (Cabot *et al.* 1994; Orr 1995). Fraïsse *et al.* (2014) tested this mechanism using simulations and concluded that it is unlikely to be effective. Second, that the number of combinations of n loci increases with n (Orr 1995). This explanation is difficult to evaluate in the absence of more information on the probability of origination of complex DMIs. my results indicate that that probability could be higher than previously thought because complex DMIs are continuously generated from simple DMIs.

4.1.4 How simple incompatibilities become complex

Perhaps the central insight from my study is that simple DMIs have a tendency to become complex. At first glance this claim might seem absurd. Surely a DMI cannot be simple one moment and complex the next. The solution to this puzzle rests, I believe, on the difference between a DMI having a certain order n and my ability to infer that it has order n through genetic crosses. Consider the evolving sequences depicted in Figure 2.2. Now, imagine that there is a complex DMI of order $n = 3$ between the alleles A_1, B_2, and C_0, and that there are no simple DMIs between pairs of the three alleles (i.e., A_1/B_2, A_1/C_0, and B_2/C_0). For simplicity, I also assume that none of the other alleles at the A, B and C loci are involved in DMIs. The existence of a DMI is defined in the strict sense that any conceivable genotype containing all alleles involved in the DMI is inviable (conversely, the absence of a DMI indicates that at least one of the genotypes containing all alleles involved in the DMI are viable). Despite the $A_1/B_2/C_0$ DMI being complex,

after two substitutions ($k = 2$), my introgression and rescue tests would detect a nonexistent simple DMI between alleles A_1 and B_2. Only after the third substitution ($k = 3$) would the true complex DMI be inferred. In the language I have been using so far, the simple DMI would appear to *become* more complex.

The snowball model (Orr 1995) assumes that it is possible to tell whether a DMI is simple or not. However, a strict definition of "DMI of order n" cannot be applied in practice because the number of genotypes that would have to be tested is astronomically large *and* would have to include mutations that have not even occurred yet. my protocol for inferring a simple DMI is, as far as I know, the most exhaustive ever devised (the data summarized in Figure 2.12A required the construction of approximately 6×10^4 introgression genotypes for each individual simulation), but it cannot infer simple DMIs in the strict sense. Simple DMIs in the strict sense may not even exist. The idea of complexification of DMIs is a natural consequence of using a more practical, broad-sense definition of simple DMI.

The extent to which the RNA-folding model is representative of other types of epistatic interactions is unclear. One possible criticism is that I used very short sequences and that these are likely to experience unusually strong epistatic interactions. Orr and Turelli (2001) estimated $p \approx 10^{-7}$ in *Drosophila* a much lower value than found in my simulations. However, an evolution experiment in *Saccharomyces cerevisiae* detected a simple DMI between two lineages that had only accumulated 6 unique mutations each ($k = 12$) (Anderson *et al.* 2010). This indicates a value of $p = 0.015$, within the range of what I observed in the RNA-folding

model (Figure 2.9).

I found that my results were robust to a broad range of holey fitness landscapes defined in the RNA-folding model.

However, the holey landscape model makes two strong assumptions about the fitness landscape: all viable genotypes had the same fitness, and all low fitness genotypes were completely inviable. Neither assumption is met universally: many alleles involved in DMIs appear to have experienced positive selection during their evolutionary history (Presgraves 2010b; Rieseberg and Blackman 2010; Maheshwari and Barbash 2011), and some DMIs are only mildly deleterious rather than lethal (Presgraves 2003; Schumer *et al.* 2014). Other fitness landscapes can be implemented readily within the RNA-folding model (e.g., Cowperthwaite *et al.* 2005; Draghi *et al.* 2011). The extent to which relaxing the assumptions of the holey landscape model will affect my results is a question for future research.

My study has identified one determinant of the origination and complexification of DMIs: the holeyness of the fitness landscape. In a holey fitness landscape, our measure of holeyness is inversely related to the mutational robustness of the genotypes assayed (van Nimwegen *et al.* 1999; Ancel and Fontana 2000). In my model (as in Orr's) "populations" are assumed to contain a single genotype; periodically, a mutant genotype arises and either goes to fixation or disappears. In such a model, mutational robustness is not expected to evolve (van Nimwegen *et al.* 1999). Individual-based simulations would allow us to investigate the intriguing possibility that factors that influence the evolution of mutational robustness (e.g., mutation rate, recombination rate: Wilke *et al.* 2001; Gardner and

Kalinka 2006; Azevedo *et al.* 2006) may influence the accumulation of DMIs (discussed below).

4.2 How populations shape the accumulation of incompatibilities

4.2.1 Selection for robustness can affect the accumulation of DMIs within and between populations

Mutational robustness can be defined as the ability of a phenotype to be viable in the face of mutations (Gardner and Kalinka 2006). Using digital organisms, Misevic *et al.* (2006) show that sexual populations become more insensitive to mutation, i.e., they are more robust, than asexual populations. Gardner and Kalinka (2006) also predict that increasing recombination rate results in an increased robustness. The link between robustness and recombination stems from the fact that recombination can result in selection for "mixability", i.e., selection for mutations that can perform well in a variety of genetic backgrounds (Livnat *et al.* 2008; Azevedo *et al.* 2006). Lohaus *et al.* (2010) show that, at least in artificial gene networks, recombination in can result in selection for mixable genotypes. This selection for mixability should, by definition, inhibit the development of incompatibilities between genotypes.

The relation between recombination and selection for mutational robustness

can explain why recombination slows down the accumulation of incompatibilities in the individual-based simulations (Figures 3.5C and 3.6). The increased recombination load (Figure 3.7) in the early stages of divergence does increase as recombination rate increases. This selection for mixability can also explain the suppression of the segregating DMIs (Figure 3.5D). This explanation is also consistent with the observation that the disappearance of segregating DMIs in the individual-based simulations with polymorphic ancestors corresponds with high recombination load at the early stages of evolution (Figures 3.8D and 3.8F). It should be noted the proportion of inviable genotypes is lower in polymorphic ancestral populations since they had evolved asexually to reach the desired level of polymorphism (see section 3.2.3), and consequently had been under selection, unlike the monomorphic ancestral population. Consistent with this explanation, the presence of segregating DMIs in the ancestral population and the resulting selection for mixability through recombination decreases the number of DMIs between diverging populations as well (Figures 3.8C).

In addition, the fact that asexual individual-based simulations with lower mutations rates accumulate more DMIs when compared to simulations with higher mutation rates further supports the veracity of the robustness hypothesis (Figure 3.12).

So why do segregating DMIs persist in natural populations in the face of recombination? In the case of the *peel-1* and *zeel-1* elements in *Caenorhabditis elegans*, Seidel *et al.* (2008) suggests that this segregating DMI is maintained via balancing selection. In other examples, such as segregating DMIs *Drosophila* (Corbett-Detig

et al. 2013), it is more likely that segregating DMIs exist between geographically isolated populations with low levels of gene flow.

Given the negative relation between number of DMIs and the recombination rate, it is plausible that at the genomic level, where the recombination rate is not homogenous (Myers *et al.* 2005), suppression of recombination rate in regions of the genome can make them more likely to be involved in an incompatibility. Although such reasoning has been suggested for recombination between populations (Nosil and Feder 2012), to my knowledge, this mechanism linking the suppression of recombination to the emergence of incompatibilities has not been proposed before.

The effect of recombination on robustness and, consequently, on the accumulation of incompatibilities means that one should be cautious when dealing with a theoretical/computation model that does not take recombination into account. In the absence of recombination, an asexual model would result in an overestimation of the number of incompatibilities and high level of RI (Figures 3.5A and 3.5E). In the presence of recombination, selection for mixability would reduce the number of DMIs accumulated over divergence, a fact that is absent from an asexual theoretical/computation model. The higher levels of RI observed in an asexual model may also be misleading since in populations with low recombination only a few hybrids would actually experience low fitness.

4.2.2 Ancestral polymorphism acts against the snowballing of incompatibilities

Cutter (2011) suggests that neglecting the ancestral polymorphism can be problematic for tests of the snowball model. Excluding the ancestral polymorphism results in inferring a longer divergence time between lineages, which in turn can imply faster than linear accumulation of incompatibilities even when incompatibilities accumulate linearly. My model shows a similar pattern, where more ancestral polymorphism results in fewer runs that support the snowball model (Figure 3.9). But the reason for the more linear accumulation of DMIs cannot be attributed to the segregating DMIs since in the sexual simulations, segregating DMIs within polymorphic ancestral populations are suppressed at the beginning of divergence (Figure 3.8D). I suggest that the combined effects of selection for phenotypes that are mutationally robust, i.e., sequences that are more likely to be viable after mutating, and selection for mixability through recombination on segregating DMIs results in linearization of the snowballing of DMIs.

It should be noted that while my results shows that an increase in ancestral polymorphism reduces the proportion of runs that fit the snowball model, but it does not appear to increase support for the linear model, which may be caused by a decrease in statistical power rather than a change in the behavior from snowball to linear. Longer runs and more data points are needed to test if the ancestral polymorphism does indeed linearizes the accumulation of incompatibilities.

4.2.3 DMIs can persist in spite of gene flow

My results provides some evidence in favor of the prediction made by Kondrashov (2003) that in a the presence of gene flow the number of DMIs can accumulate linearly (Figure 3.11). Given that many of the runs do not fit the snowball model or the linear model, longer runs and more data points are needed to test if migration does lead to a linear accumulation of DMIs.

Bank *et al.* (2012) suggest that maintenance of neutral DMIs is impossible with gene flow, but while the number of DMIs in my simulations decrease at high level of gene flow, they still retain DMIs and, consequently, RI (Figures 3.10C and 3.10D).

4.3 Philosophical obstacles to a complete understanding of DMIs

This central idea of this thesis revolves around a simple and elegant model, proposed by Orr (1995). The snowball model was an attempt to reduce the evolution of genetic incompatibilities into its most basic components. The snowball model can be described as an intuition pump (Dennett 2013), an elegant way to think about an extremely complicated matter. Instead of the reductionist approach inherent in the snowball model, the work presented here was an attempt to complexify the snowball model, transfiguring it into a model where RNAs fold and evolve in a holey landscape and later going even further and construct

an individual-based simulation on top of it. I believe this slight level complexification, allows to understand the implicit implications of the snowball model. The individual-based simulation presents us with more intrigue, demonstrating an unexpected effect of recombination on the accumulation of DMIs. I hope that the project underlying this thesis would allow for a better understanding of the genetics of speciation.

Bibliography

Allen, W. and R. Short, 1997 Interspecific and extraspecific pregnancies in equids: Anything goes. Journal of Heredity **88**: 384–392.

Ancel, L. W. and W. Fontana, 2000 Plasticity, evolvability, and modularity in RNA. Journal of Experimental Zoology Part B Molecular and Developmental Evolution **288**: 242–283.

Anderson, J. B., J. Funt, D. A. Thompson, S. Prabhu, A. Socha, C. Sirjusingh, J. R. Dettman, L. Parreiras, D. S. Guttman, A. Regev, and L. M. Kohn, 2010 Determinants of divergent adaptation and Dobzhansky-Muller interaction in experimental yeast populations. Current Biology **20**: 1383 – 1388.

Azevedo, R. B. R., R. Lohaus, S. Srinivasan, K. K. Dang, and C. L. Burch, 2006 Sexual reproduction selects for robustness and negative epistasis in artificial gene networks. Nature **440**: 87–90.

Bank, C., R. Bürger, and J. Hermisson, 2012 The limits to parapatric speciation: Dobzhansky–Muller incompatibilities in a continent–island model. Genetics **191**: 845–863.

Barrière, A. and M.-A. Félix, 2005 High local genetic diversity and low outcrossing rate in *Caenorhabditis elegans* natural populations. Current Biology **15**: 1176 – 1184.

Barton, N. H., D. E. Briggs, J. A. Eisen, D. B. Goldstein, and N. H. Patel, 2007 *Evolution*. Cold Spring Harbor Laboratory Press.

Bernet, G. P. and S. F. Elena, 2015 Distribution of mutational fitness effects and of epistasis in the 5'untranslated region of a plant RNA virus. BMC Evolutionary Biology **15**: 274.

Bindewald, E., T. Kluth, and B. A. Shapiro, 2010 Cylofold: secondary structure prediction including pseudoknots. Nucleic Acids Research **38**: W368–W372.

Bolnick, D. I. and T. J. Near, 2005 Tempo of hybrid inviability in centrarchid fishes (Teleostei: Centrarchidae). Evolution **59**: 1754–1767.

Brown, J. D. and R. J. O'Neill, 2010 Chromosomes, conflict, and epigenetics: Chromosomal speciation revisited. Annual Review of Genomics and Human Genetics **11**: 291–316, PMID: 20438362.

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.

Chae, E., K. Bomblies, S.-T. Kim, D. Karelina, M. Zaidem, S. Ossowski, C. Martín-Pizarro, R. A. Laitinen, B. A. Rowan, H. Tenenboim, S. Lechner, M. Demar,

A. Habring-Müller, C. Lanz, G. Rätsch, and D. Weigel, 2014 Species-wide genetic incompatibility analysis identifies immune genes as hot spots of deleterious epistasis. Cell **159**: 1341 – 1351.

Cock, A. and D. R. Forsdyke, 2008 *Treasure Your Exceptions: The Science and Life of William Bateson.*. Springer, New York.

Corbett-Detig, R. B., J. Zhou, A. G. Clark, D. L. Hartl, and J. F. Ayroles, 2013 Genetic incompatibilities are widespread within species. Nature **504**: 135–137.

Cowperthwaite, M. C., J. J. Bull, and L. A. Meyers, 2005 Distributions of beneficial fitness effects in RNA. Genetics **170**: 1449–57.

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

Coyne, J. A. and H. A. Orr, 1997 "Patterns of speciation in *Drosophila*" revisited. Evolution **51**: 295–303.

Coyne, J. A. and H. A. Orr, 2004 *Speciation*. W.H. Freeman, New York.

Cracraft, J., 1987 Species concepts and the ontology of evolution. Biology and Philosophy **2**: 329–346.

Cutter, A. D., 2011 The polymorphic prelude to Bateson-Dobzhansky-Muller incompatibilities. Trends in Ecology & Evolution **27**: 209–218.

Darwin, C., 1859 *On the Origin of Species by Means of Natural Selection*. J. Murray, London.

Dennett, D. C., 2013 *Intuition Pumps And Other Tools for Thinking*. W. W. Norton, New York.

Diderot, D. and J. l. R. d'Alembert, 1751 *Encyclopédie, ou dictionnaire raisonné des sciences, des arts et des métiers, etc.*. University of Chicago: ARTFL Encyclopédie Project.

Do, C. B., D. A. Woods, and S. Batzoglou, 2006 CONTRAfold: RNA secondary structure prediction without physics-based models. Bioinformatics **22**: e90–e98.

Dobzhansky, T., 1937 *Genetics and the Origin of Species*. Columbia University Press, New York.

Draghi, J. A., T. L. Parsons, and J. B. Plotkin, 2011 Epistasis increases the rate of conditionally neutral substitution in an adapting population. Genetics **187**: 1139–52.

Draghi, J. A., T. L. Parsons, G. P. Wagner, and J. B. Plotkin, 2010 Mutational robustness can facilitate adaptation. Nature **463**: 353–355.

Edmands, S., 2002 Does parental divergence predict reproductive compatibility? Trends in Ecology & Evolution **17**: 520–527.

Fitzpatrick, B. M., 2002 Molecular correlates of reproductive isolation. Evolution **56**: 191–198.

Forsdyke, D. R., 2011 The B in 'BDM.' William Bateson did not advocate a genic speciation theory. Heredity **106**: 202.

Fraïsse, C., J. A. D. Elderfield, and J. J. Welch, 2014 The genetics of speciation: Are complex incompatibilities easier to evolve? Journal of Evolutionary Biology **27**: 688–699.

Fraïsse, C., P. A. Gunnarsson, D. Roze, N. Bierne, and J. J. Welch, 2016 The genetics of speciation: Insights from Fisher's geometric model. Evolution **70**: 1450–1464.

Gardner, A. and A. T. Kalinka, 2006 Recombination and the evolution of mutational robustness. Journal of Theoretical Biology **241**: 707 – 715.

Gavrilets, S., 2004 *Fitness Landscapes and the Origin of Species*. Princeton University Press, Princeton.

Giraud, T. and S. Gourbière, 2012 The tempo and modes of evolution of reproductive isolation in fungi. Heredity **109**: 204–214.

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.

Greig, D., M. Travisano, E. J. Louis, and R. H. Borts, 2003 A role for the mismatch repair system during incipient speciation in saccharomyces. Journal of Evolutionary Biology **16**: 429–437.

Hamada, M., H. Kiryu, K. Sato, T. Mituyama, and K. Asai, 2009 Prediction of RNA secondary structure using generalized centroid estimators. Bioinformatics **25**: 465–473.

Hou, J., A. Friedrich, J. de Montigny, and J. Schacherer, 2014 Chromosomal rearrangements as a major mechanism in the onset of reproductive isolation in *Saccharomyces cerevisiae*. Current Biology **24**: 1153 – 1159.

Huxley, T. H., 1863 *On our knowledge of the causes of the phenomena of organic nature: six lectures to working men*. Hardwicke, London.

Johnson, N. A., 2006 The evolution of reproductive isolating barriers. In *Evolutionary Genetics: Concepts and Case Studies*, edited by C. W. Fox and J. B. Wolf, pp. 374–398, Oxford Univ. Press, Oxford, U.K.

Johnson, N. A., 2009 One hundred years after bateson: a pair of incompatible genes underlying hybrid sterility between yeast species. Heredity **103**: 360–361.

Jordan, K., 1905 Der Gegensatz zwischen geographischer und nichtgeographischer Variation. Zeitschrift für Wissenschaftliche Zoologie **83**: 151–210.

Josefsson, C., B. Dilkes, and L. Comai, 2006 Parent-dependent loss of gene silencing during interspecies hybridization. Current Biology **16**: 1322 – 1328.

Kaneshiro, K. Y., 1980 Sexual isolation, speciation and the direction of evolution. Evolution **34**: 437–444.

Kimura, M., 1962 On the probability of fixation of mutant genes in a population. Genetics **47**: 713–719.

Kondrashov, A. S., 2003 Accumulation of Dobzhansky-Muler incompatibilities within a spatially structured population. Evolution **57**: 151–153.

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

Kondrashov, D. A. and F. A. Kondrashov, 2015 Topological features of rugged fitness landscapes in sequence space. Trends in Genetics **31**: 24–33.

Kulathinal, R. J., B. R. Bettencourt, and D. L. Hartl, 2004 Compensated deleterious mutations in insect genomes. Science **306**: 1553–1554.

Larcombe, M. J., B. Holland, D. a. Steane, R. C. Jones, D. Nicolle, R. E. Vaillancourt, and B. M. Potts, 2015 Patterns of reproductive isolation in *Eucalyptus*—a phylogenetic perspective. Molecular Biology and Evolution **32**: 1833–1846.

Lewontin, R. C., 1974 *The Genetic Basis of Evolutionary Change*. Biological Series, Columbia University Press, New York.

Lijtmaer, D. A., B. Mahler, and P. L. Tubaro, 2003 Hybridization and postzygotic isolation patterns in pigeons and doves. Evolution **57**: 1411–1418.

Livingstone, K., P. Olofsson, G. Cochran, A. Dagilis, K. MacPherson, and K. A. Seitz Jr., 2012 A stochastic model for the development of BatesonDobzhansky-Muller incompatibilities that incorporates protein interaction networks. Mathematical Biosciences **238**: 49 – 53.

Livnat, A., C. Papadimitriou, J. Dushoff, and M. W. Feldman, 2008 A mixability theory for the role of sex in evolution. Proceedings of the National Academy of Sciences of the United States of America **105**: 19803–19808.

Lohaus, R., C. L. Burch, and R. B. R. Azevedo, 2010 Genetic architecture and the evolution of sex. Journal of Heredity **101**: S142–S157.

Lorenz, R., S. H. Bernhart, C. Höner Zu Siederdissen, H. Tafer, C. Flamm, P. F. Stadler, and I. L. Hofacker, 2011 ViennaRNA Package 2.0. Algorithms for Molecular Biology **6**: 26.

Maheshwari, S. and D. A. Barbash, 2011 The genetics of hybrid incompatibilities. Annual Review of Genetics **45**: 331–355.

Mallet, J., 2004 Perspectives poulton, wallace and jordan: How discoveries in papilio butterflies led to a new species concept 100 years ago. Systematics and Biodiversity **1**: 441–452.

Masly, J. P., C. D. Jones, M. A. F. Noor, J. Locke, and H. A. Orr, 2006 Gene transposition as a cause of hybrid sterility in *Drosophila*. Science **313**: 1448–1450.

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 **1518**: 1518–1521.

Matute, D. R. and J. Gavin-Smyth, 2014 Fine mapping of dominant X-linked incompatibility alleles in *Drosophila* hybrids. PLoS Genetics **10**.

Mayr, E., 1942 *Systematics and the Origin of Species*. Columbia University Press, New York.

Mayr, E., 1949 Speciation and selection. Proceedings of the American Philosophical Society **93**: 514–519.

Mendelson, T. C., B. D. Inouye, and M. D. Rausher, 2004 Quantifying patterns in the evolution of reproductive isolation. Evolution **58**: 1424–1433.

Michalak, P., 2008 Epigenetic, transposon and small rna determinants of hybrid dysfunctions. Heredity **102**: 45–50.

Misevic, D., C. Ofria, and R. E. Lenski, 2006 Sexual reproduction reshapes the genetic architecture of digital organisms. Proceedings of the Royal Society of London B: Biological Sciences **273**: 457–464.

Moyle, L. C. and T. Nakazato, 2010 Hybrid incompatibility "snowballs" between *Solanum* species. Science **329**: 1521–1523.

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

Myers, S., L. Bottolo, C. Freeman, G. McVean, and P. Donnelly, 2005 A fine-scale map of recombination rates and hotspots across the human genome. Science **310**: 321–324.

Nosil, P. and J. L. Feder, 2012 Genomic divergence during speciation: causes and consequences. Philosophical Transactions of the Royal Society B: Biological Sciences **367**: 332–342.

O'Neill, R. J. W., M. J. O'Neill, and J. A. M. Graves, 1998 Undermethylation associated with retroelement activation and chromosome remodelling in an interspecific mammalian hybrid. Nature **393**: 68–72.

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

Orr, H. A., 1996 Dobzhansky, Bateson, and the genetics of speciation. Genetics **144**: 1331–1335.

Orr, H. A., 2001 The genetics of species differences. Trends in Ecology & Evolution **16**: 343 – 350.

Orr, H. A. and S. Irving, 2001 Complex epistasis and the genetic basis of hybrid sterility in the *Drosophila pseudoobscura* Bogota-USA hybridization. Genetics **158**: 1089–1100.

Orr, H. A. and M. Turelli, 2001 The evolution of postzygotic isolation: accumulating Dobzhansky-Muller incompatibilities. Evolution **55**: 1085–1094.

Palmer, M. E. and M. W. Feldman, 2009 Dynamics of hybrid incompatibility in gene networks in a constant environment. Evolution **63**: 418–431.

Paulson, D. R., 1974 Reproductive isolation in damselflies. Systematic Zoology **23**: 40–49.

Poulton, E. B., 1908 What is a species? In *Essays on evolution 1889-1907*, chapter 2, pp. 46–94, Claredon Press, Oxford.

Presgraves, D. C., 2002 Patterns of postzygotic isolation in Lepidoptera. Evolution **56**: 1168–1183.

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

Presgraves, D. C., 2010a Speciation genetics: search for the missing snowball. Current Biology **20**: R1073–4.

Presgraves, D. C., 2010b The molecular evolutionary basis of species formation. Nature Reviews Genetics **11**: 175–180.

Proulx, S. R. and P. C. Phillips, 2005 The opportunity for canalization and the evolution of genetic networks. The American Naturalist **165**: 147–162, PMID: 15729647.

Ptacek, M. B. and S. J. Hankison, 2009 The pattern and process of speciation. In *Evolution: the first four bilion years*, edited by M. Ruse and J. Travis, The Belknap press.

Reznick, D. N., 2010 *The Origin then and now : an interpretive guide to the Origin of species*. Princeton University Press, Princeton.

Rieseberg, L. H. and B. K. Blackman, 2010 Speciation genes in plants. Annals of Botany **106**: 439–455.

Sasa, M. M., P. T. Chippindale, and N. A. Johnson, 1998 Patterns of postzygotic isolation in frogs. Evolution **52**: 1811–1820.

Sato, K., Y. Kato, M. Hamada, T. Akutsu, and K. Asai, 2011 IPknot: fast and accurate prediction of RNA secondary structures with pseudoknots using integer programming. Bioinformatics **27**: 185–193.

Satyaki, P. R. V., T. N. Cuykendall, K. H.-C. Wei, N. J. Brideau, H. Kwak, S. Aruna, P. M. Ferree, S. Ji, and D. A. Barbash, 2014 The *Hmr* and *Lhr* hybrid incompatibility genes suppress a broad range of heterochromatic repeats. PLoS Genetics **10**: e1004240.

Schumer, M., R. Cui, D. L. Powell, R. Dresner, G. G. Rosenthal, and P. Andolfatto, 2014 High-resolution mapping reveals hundreds of genetic incompatibilities in hybridizing fish species. eLife **3**: e02535.

Schuster, P., W. Fontana, P. F. Stadler, and I. L. Hofacker, 1994 From sequences to shapes and back: a case study in RNA secondary structures. Proceedings of the Royal Society B: Biological Sciences **255**: 279–284.

Seidel, H. S., M. Ailion, J. Li, A. van Oudenaarden, M. V. Rockman, and L. Kruglyak, 2011 A novel sperm-delivered toxin causes late-stage embryo lethality and transmission ratio distortion in *C. elegans*. PLoS Biology **9**: 1–21.

Seidel, H. S., M. V. Rockman, and L. Kruglyak, 2008 Widespread genetic incompatibility in *C. Elegans* maintained by balancing selection. Science **319**: 589–594.

Sherman, N. A., A. Victorine, R. J. Wang, and L. C. Moyle, 2014 Interspecific tests of allelism reveal the evolutionary timing and pattern of accumulation of reproductive isolation mutations. PLoS Genetics **10**: e1004623.

Sherratt, T. N. and D. M. Wilkinson, 2009 *Big Questions in Ecology and Evolution*. Oxford University Press, New York.

Sloan, P. R., 2008 Originating species: Darwin on the species problem. In *The Cambridge Companion to the 'Origin of Species'*, edited by M. Ruse and R. J. Richards, chapter 5, pp. 67–86, Cambridge University Press, New York.

Sniegowski, P. D. and P. J. Gerrish, 2010 Beneficial mutations and the dynamics of adaptation in asexual populations. Philosophical Transactions of the Royal Society of London B: Biological Sciences **365**: 1255–1263.

Sokal, R. R. and T. J. Crovello, 1970 The biological species concept: A critical evaluation. The American Naturalist **104**: 127–153.

Swenson, M. S., J. Anderson, A. Ash, P. Gaurav, Z. Sükösd, D. A. Bader, S. C. Harvey, and C. E. Heitsch, 2012 GTfold: Enabling parallel RNA secondary structure prediction on multi-core desktops. BMC Research Notes **5**: 1–6.

Taylor, M. B. and I. M. Ehrenreich, 2015 Higher-order genetic interactions and their contribution to complex traits. Trends in Genetics **31**: 34–40.

Templeton, A. R., 1980 The theory of speciation via the founder principle. Genetics **94**: 1011–1038.

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

van Nimwegen, E., J. P. Crutchfield, and M. Huynen, 1999 Neutral evolution of mutational robustness. Proceedings of the National Academy of Sciences of the United States of America **96**: 9716–9720.

van Valen, L., 1976 Ecological species, multispecies, and oaks. Taxon **25**: 233–239.

Wagner, A., 2008 Robustness and evolvability: a paradox resolved. Proceedings of the Royal Society B: Biological Sciences **275**: 91–100.

Wang, H., E. D. McArthur, S. C. Sanderson, J. H. Graham, and D. C. Freeman, 1997 Narrow hybrid zone between two subspecies of big sagebrush (artemisia tridentata: Asteraceae). iv. reciprocal transplant experiments. Evolution **51**: 95–102.

Wang, R. J., M. A. White, and B. A. Payseur, 2015 The pace of hybrid incompatibility evolution in house mice. Genetics **201**: 229–242.

Weinreich, D. M., Y. Lan, C. S. Wylie, and R. B. Heckendorn, 2013 Should evolutionary geneticists worry about higher-order epistasis? Current Opinion in Genetics & Development **23**: 700–707.

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

Wiley, E. O., 1978 The evolutionary species concept reconsidered. Systematic Zoology **27**: 17–26.

Wilke, C. O., J. L. Wang, C. Ofria, R. E. Lenski, and C. Adami, 2001 Evolution of digital organisms at high mutation rates leads to survival of the flattest. Nature **412**: 331–333.

Zuker, M. and P. Stiegler, 1981 Optimal computer folding of large RNA sequences using thermodynamics and auxiliary information. Nucleic Acids Research **9**: 133–148.