The role of recombination and hybridisation in adaptive evolution

A thesis submitted to the School of Environmental Sciences of the University of East Anglia in partial fulfillment of the requirements for the degree of Doctor of Philosophy

By Ben J. Ward

Registration number: 1000135331

March 2017

Abstract

As one of the five evolutionary forces, recombination fulfills both a cleansing role, as well as a role in generating genetic diversity. Recombination cleanses by separating deleterious mutations from their genomic background, increasing the efficacy of purifying selection and curtailing the continuous accumulation of deleterious mutations. Recombination also plays a fundamental role in the repair of damaged DNA, and it can be a creative force, resulting in the formation of novel genotypes, haplotypes and alleles, thereby playing a key role in adaptive evolution. By uniting beneficial mutations that exist at different loci in separate lineages, meiotic recombination during sex accelerates adaptive evolution. Although recombination leaves a distinct signature or footprint in the genome of organisms, identifying this force can be difficult; subsequent recombination events tend to wipe out their past genomic footprints. This thesis presents the development of a novel software package called HybridCheck, for the detection of genomic regions affected by recombination in Next Generation Sequence data, and the rapid molecular dating of recombination events. Hybrid-Check was used to analyze recombination signal in different races of the plant pathogen *Albugo candida*, a generalist obligate biotroph that infects Brassica plants. I show that recombination facilitated occasional introgression and gene flow between host-specialized races. This may have accelerated the rate of adaptive evolution, and possibly broadened

the pathogen's host-range. Finally, the genome of the polar diatom *Fragilariopsis cylindrus* contains diverged alleles that are differentially expressed in different environmental conditions. The hypothesis that ancient asexuality explains how the diverged alleles evolved is challenged, but not rejected, based on evidence of recombination presented in this thesis. An alternative hypothesis is proposed: allelic divergence might have evolved despite the homogenizing effect of meiotic recombination as a result of very large effective population sizes and strong diversifying selection on *F. cylindrus* in the polar environment.

Acknowledgments

I would like to thank my friends and family for their support over the past 4 years. I would also like to thank Professor van Oosterhout for being both a friend and supervisor, for having confidence in my abilities, and for guiding my development into a young scientist, as well as comments and corrections on this manuscript. I thank Professor Kamoun for his guidance and advice about the direction of my PhD, as well as for comments and corrections on this manuscript. I thank Mark McMullan for every comment and discussion ever made over many informal conversations which helped me develop my own understanding. Finally, I thank Dr. Matt Clark and his Plant and Microbial Genomics Team at The Genome Analysis Center for welcoming me and providing the best environment I could ask for, in which to work and write up my thesis in my final year.

Contents

List of tables

List of figures

CHAPTER 1

General Introduction

This thesis presents work investigating the role that recombination plays in the adaptive evolution of two eukaryotic microorganisms, *Albugo candida* and *Fragilariopsis cylindrus*. Both of these organisms exist in environments that may be considered very dynamic.

In addition, methodological work was also conducted which implemented and tested software dedicated to making it easier to detect recombination in Next Generation Sequencing data. The software was also designed to help solve current methodological issues with distinguishing mosaic regions that are the result of hybridisation, and those that are the result of incomplete lineage sorting.

These works are presented in chapters 2, 3, and 4. Each has a more detailed and focused introduction to the concepts specific to them. It is the purpose of this chapter to provide an overview of the key concepts of population genetics that are relevant to this work and provide a wider context for the next three chapters.

In order to understand adaptive evolution, it is necessary to understand the five forces of population genetics and how they drive adaptive evolution. What follows is an overview of the five fundamental forces of evolutionary change. Afterwards, an overview of hybrid zones, and an overview of

current common Bioinformatics procedures and how they are used in population genetics analyses are presented.

1.1 The five forces of evolutionary change

1.1.1 Selection

Selection is the non-random, differential survival and reproduction of organisms as a result of their different phenotypes. A population contains many individuals, and these individuals vary in their genetic makeup; the population has genetic variance. This genetic variation, in combination with some environmental effects, is the cause of the phenotypic variation in a population (Ridley 2004). This phenotypic variation results in variation in survival, fecundity, and mating ability, and this ultimately determines whether an individual contributes any alleles to the next generation of that population: Individuals may be better or worse at surviving, or may not be chosen by the opposite sex to mate (Hedrick 2010). This can be expressed in terms of relative fitness. Relative fitness can be defined as the relative ability of different genotypes to pass on their alleles to future generations Charlesworth and Charlesworth 2010. Individuals with genotypes that have a higher relative fitness are expected to survive and pass their alleles on to the next generation, and so over several generations, those genotypes will increase in frequency in the population.

1.1.1.1 The basic diploid model

The basic diploid model of selection models how selection operates for a single diploid locus, with two alleles. The model assumes that there is random mating among individuals in a population, and that selection is operating identically for both sexes. In this model, selection occurs through differences in viability and it is constant through space and time i.e. it acts

on every individual in every generation, regardless of location. Generations are discrete and non-overlapping and no mutation is occurring. No gene flow or inbreeding occurs and the size of the population is infinite so there is no genetic drift (Charlesworth and Charlesworth 2010; Hedrick 2010). Despite these assumptions it is still a very useful model to explore and describe how selection operates.

Assume there are two alleles of a single locus, denoted as A_1, and A_2. With these two alleles, three possible diploid genotypes are possible. Two of them are heterozygous: A_1A_1, and A_2A_2, and the third, A_1A_2 is heterozygous. The relative fitnesses of A_1A_1, A_1A_2, and A_2A_2 are denoted as w_{11}, w_{12}, and w_{22} respectively (Wright 1937). The contribution of each genotype to the next generation can be calculated as the product of its relative fitness and its frequency prior to selection. The contributions of A_1A_1, A_1A_2, and A_2A_2 are $p_0^2w_{11}$, $2p_0q_0w_{12}$, and q_0^2w22, where p is defined as the frequency of A_1 and q is defined as the frequency of A_2 (Charlesworth and Charlesworth 2010; Hedrick 2010). Assuming Hardy-Weinberg allele proportions before selection, the mean fitness of the population is:

$$\bar{w} = p_0^2w_{11} + 2p_0q_0w_{12} + q_0^2w_{22} \tag{1.1}$$

The frequency of a genotype after selection can be calculated by dividing its contribution by the mean fitness, for example, for A_1A_1 this is $p_0^2w_{11}/\bar{w}$. The frequency of the alleles A_1 and A_2 after selection (p_1 and q_1) can be obtained by noting that the frequency of any of the two alleles is the sum of the frequency of the homozygous genotype and half the frequency of the heterozygous genotype (Charlesworth and Charlesworth 2010; Hedrick 2010).

$$p_1 = \frac{p(pw_{11} + qw_{12})}{\bar{w}} \tag{1.2a}$$

$$q_1 = \frac{q(pw_{12} + qw_{22})}{\bar{w}} \tag{1.2b}$$

The change in q over one round of selection can be defined as $\Delta q = q_1 - q_0$. Substituting q_1 and simplifying the formula gives equation 1.3 (Charlesworth and Charlesworth 2010; Hedrick 2010).

$$\Delta q = \frac{pq(w_{2\cdot} - w_{1\cdot})}{\bar{w}} \tag{1.3}$$

If p or q are 0, then there can be no change in frequencies of that allele, as it is not present in the population (Charlesworth and Charlesworth 2010; Hedrick 2010).

1.1.1.2 Different fitness relationships

The formulas and quantities just described can be used to explore the effects of selection for different fitness relationships. Different relative fitness values of w_{11}, w_{12}, and w_{22} can be generated for different fitness relationships through the combination of two other coefficients: s is the selection coefficient which measures the amount of selection against a homozygote, and h is the level of dominance (Charlesworth and Charlesworth 2010; Hedrick 2010). When h is multiplied by s, this measures the amount of selection against a heterozygote (Charlesworth and Charlesworth 2010; Hedrick 2010). These different fitness relationships are displayed in table 1.1.

A recessive lethal allele describes an allele which has a detrimental effect on the individual that is so severe it leads to death of the individual. Examples of alleles with such effects include those that cause Tay-Sachs disease in humans (Myerowitz 1997). Relative fitnesses for this situation

Table 1.1: Fitness values for different fitness relationships, adapted from Hedrick 2010.

Fitness Relationship	A_1A_1	A_1A_2	A_2A_2
Recessive lethal	1	1	0
Recessive detrimental	1	1	$1-s$
Additive detrimental	1	$1-(s/2)$	$1-s$
Purifying Selection	1	$1-hs$	$1-s$
Positive Selection	$1+s$	$1+hs$	1
Overdominance	$1-s_1$	1	$1-s_2$
Underdominance	$1+s_1$	1	$1+s_2$

are given in row one of table 1.1. Using these values in the formulas 1.1 and 1.3 it can be demonstrated that the mean fitness of a population reaches 1 when there is no A_2 allele in the population ($q = 0$). Furthermore, Δq is largest when q is large, and is smaller when q approaches 0 (Hedrick 2010). Therefore, when the frequency of a recessive lethal is high it is purged by selection very quickly from the population. The reason lethal recessive alleles are not purged as quickly when they are at low frequency is that they are present in heterozygotes, therefore the deleterious recessive alleles are not subject to differential selection (Hedrick 2010).

Some recessive alleles are not lethal, but they are detrimental to the fitness of an individual (Charlesworth and Willis 2009; Charlesworth and Charlesworth 2010). This type of fitness relationship is called a recessive deleterious relationship. Fitness values for this scenario are given in row 2 of table 1.1. The selection coefficient (s) reflects how detrimental allele A_2 is. If $s = 1$, then A_2 would be a recessive lethal allele and selection would act as previously described. Mean fitness is maximized when $q = 0$ and Δq is greatest when $q_0 = 2/3$, and lower for smaller values of q (Hedrick 2010). Again this is because A_2 mostly occurs in individuals with a heterozygote genotype for low q.

Heterozygous individuals may have phenotypes that are intermediate to those of the two homozygotes. If the phenotype of a heterozygote is exactly

halfway between that of the homozygotes this is referred to as additivity. Fitness values for additivity are shown on line 3 of table 1.1. In this scenario Δq is larger when both alleles are equally frequent in the population. Δq is greater at low value of q than in the previous scenarios. For low q, A_2 is mostly in heterozygotes, but the deleterious effects of A_2 are not masked in the heterozygotes when the fitness relationship is additive (Charlesworth and Charlesworth 2010; Hedrick 2010).

Alleles with additive and recessive effects have been discussed, but every possible level of dominance can be represented in the model with the h coefficient. Fitness relationships modeling different levels of dominance with h are shown on lines 4 and 5 of table 1.1. These fitness arrays describe purifying and positive selection. Purifying selection acts to reduce the frequency of a detrimental allele in a population (Hedrick 2010). In contrast, positive selection acts to increase the frequency of an alleles with effects that are beneficial in the current environment of a population. In reality, selection acts in both positive and purifying roles simultaneously. In both the models if $h = 0$ then the allele is recessive, if $h = 0.5$ it is additive, and if $h = 1$ it is dominant (Charlesworth and Charlesworth 2010). For positive selection, the fastest increase in p occurs when the allele is dominant. When the allele is additive, then p still increases quickly. However, it takes longer for p to increase when A_1 is recessive. At low frequencies, the beneficial A_1 allele typically occurs in heterozygotes, and as a recessive allele, selection does not act on it (Hedrick 2010).

In the scenarios previously described selection is a force acting to reduce genetic variation as an allele either increases or decreases in frequency in a population. However, circumstances can cause selection to maintain allelic diversity in the population. This is possible when the heterozygote individuals have a higher fitness than individuals of either of the two homozygote genotypes. The phenomenon is called overdominance.

The fitness values for overdominance are listed on row 6 of table 1.1. For selection to maintain both alleles in a population, Δq must be equal to 0 for some initial q_0 between 0 and 1 (Charlesworth and Charlesworth 2010). This is called the equilibrium frequency of q, and it is a function of both the selection coefficients for the two homozygotes. When q is below this equilibrium frequency, Δq is positive. When q is above the equilibrium frequency, Δq is negative. Thus, as q is perturbed away from this equilibrium Δq shifts such that q will return to this equilibrium (Hedrick 2010). Therefore, both alleles are maintained in the population at a certain ratio.

Warfarin resistance in Rats is an example of heterozygote advantage. Resistance was conferred to the rats by a dominant allele (R) at the VKORC1 locus. Individuals with one copy of R were resistant to Warfarin, but homozygous individuals had a much greater requirement for Vitamin K (Greaves et al. 1977). Heterozygote advantage has also been invoked to explain polymorphism at loci in the major histocompatibility complex (MHC) (Spurgin and Richardson 2010). Overdominance is also an explanation of hybrid vigour (heterosis) (Baranwal et al. 2012) and so this is of particular relevance to chapter 3, where the plausibility of of a generalist plant pathogen evolving through repeated hybridisation is discussed.

Underdominance describes the situation where heterozygous individuals have a lower fitness than homozygous individuals. Fitness values for this relationship are shown on the last line of table 1.1. As with overdominance, there is an equilibrium frequency of q for which $\Delta q = 0$. However, unlike overdominance, with underdominance, Δq is positive above the equilibrium point and negative below it (Hedrick 2010). Therefore the equilibrium is unstable, and allele frequencies move away from it, rather than towards it.

1.1.1.3 Selection and dynamic environments

The basic model of selection described effectively demonstrates the key concepts of when considering how selection acts. However there are extensions to the model, for example, the model has been extended to account for more than two alleles. Selection is the mechanism that causes adaptive evolution and directional selection and molecular evidence of past positive selection is abundant (Hoekstra and Coyne 2007). Most of the phenotypic characteristics we associate with species are thought to be the end result of selection, even if the adaptive function is not obvious.

However, the efficiency of selection can be reduced: Muller introduced the concept of Genetic Load. This is defined as the reduction in fitness from the maximum possible in a population (Davis and Columbia 2011). The principal factors causing genetic load are thought to be the presence of deleterious recessive mutations, maintained by a mutation-selection balance (see section 1.1.3), and the segregation of homozygotes when there is heterozygote advantage (Davis and Columbia 2011). Small isolated populations may suffer from genetic load because they can become fixed for detrimental alleles (see section 1.1.2).

Evidence of balancing-selection; selection that maintains polymorphism like overdominance, is not as common (Bubb et al. 2006), but there are scenarios in which selection does maintain polymorphism. Selection varying in time and space, frequency dependent selection, and host-pathogen evolution, are three such models that are particularly pertinent to the research presented in this thesis as they model selection operating in a dynamic and changing environments. A common aspect of these models is that they violate an assumption of the basic model: constant fitness (Charlesworth and Charlesworth 2010). If constant fitness is not assumed, it can be shown that selection may maintain polymorphism even in absence of heterozygote

advantage.

Relative fitnesses may depend on the frequency of of the different genotypes in the population. An allele may have a greater fitness when it is present in the population in low numbers and less fitness when it is present in larger numbers (Hedrick 2010; Charlesworth and Charlesworth 2010). This is called negative frequency dependent selection. Alternatively, an allele might increase in fitness as it increases in frequency (Hedrick 2010; Charlesworth and Charlesworth 2010).

Frequency dependent selection occurs where there are host-pathogen interactions. Pathogens have genes known as virulence factors and effector genes, which enable them to infect a host. New mutations in a host species that confer resistance to a pathogen will be at low frequencies but have a high selective advantage. As a result, the allele will start to spread in the host population. As the allele becomes more common, the pathogen will find fewer new hosts they can infect (Charlesworth 2006; Frank 1993; Seger and Antonovics 1988). Therefore, pathogen numbers decrease and the advantage gained by being resistant diminishes. Indeed, if there is a cost to maintaining the resistance it will even become detrimental. This process also happens with the pathogens. As hosts acquire resistance to a pathogen, pathogens with new mutations allowing them to infect previously resistant hosts will have a strong selective advantage. The now susceptible host genotype will decrease in frequency, as the pathogen increases in frequency. The selective advantage of the pathogen genotype is reduced and may even suffer a cost if it is less virulent than other pathogen genotypes at infecting other host genotypes (Charlesworth 2006; Frank 1993; Seger and Antonovics 1988). Parasite genotype frequencies may therefore become balanced in a population, resulting in highly polymorphic genes in pathogens, such as antigenic genes in malaria, and effector genes in pathogens like *Phytophthora infestans*(Morgan and Kamoun 2007; Policy

and Conway 2001). This type of process, typically assuming gene-for-gene interactions between host and pathogen, leads to cycles of allele frequency changes in both the host and pathogen (May and Anderson 1983). This may be of particular importance to haploid pathogens which by definition, will not have their polymorphism maintained by heterozygote advantage, and may be subject to clonal interference which restricts levels polymorphism and the speed of adaptation (Gerrish and Lenski 1998).

In addition to existing in balance, polymorphisms in a host or pathogen pathogen can become fixed due to their selective advantage, which can lead to a succession of fixation events in both host and pathogen as each is under selection pressure to counter adapt each others previous adaptations. This is called an evolutionary arms race, and can lead to long term variability and rapid evolution of DNA sequences such as effector genes in plant pathogens, and R genes in plants, and accelerated molecular evolution (see chapter 3) (Brown 2003; Charlesworth 2006; Morgan and Kamoun 2007; Paterson et al. 2010; Rose et al. 2004).

Selection may maintain variation when there is enough temporal variation in relative fitnesses of different genotypes. An allele with detrimental effects in one generation may confer an advantage in subsequent generations, should conditions change. This scenario is pertinent to chapter 4 as the environment of *Fragilariopsis cylindrus* is also temporally dynamic with seasonal changes such as freezing and thawing events. Models of temporally changing fitnesses have shown that polymorphism is only maintained by selection under very strict conditions: The geometric mean of fitness over n generations for both homozygotes must be smaller than that of the heterozygote (equation 1.4) (Haldane and Jayakar 1963).

$$\left(\prod_{i=1}^{n} w_{11\cdot i}\right)^{1/n} < 1 > \left(\prod_{i=1}^{n} w_{22\cdot i}\right)^{1/n} \tag{1.4}$$

This can be illustrated by considering two seasons, A_1 is advantageous in one season, and A_2 is advantageous in the other. Fitness values in season one then are $1+s$, 1, and $1-s$ for A_1A_1, A_1A_2, and A_2A_2 respectively. In the second season, this is reversed and A_1A_1 has fitness $1-s$ and A_2A_2 has fitness $1+s$. If the same number of generations is spent in each season, conditions for polymorphism are met, otherwise directional selection will result instead. Such expectations from theory have been validated in experimental evolution studies with bacteria, where serial transfer regimes were used to emulate the effects of temporal variation (Rainey et al. 2000). Therefore, it seems that there is little evidence polymorphism is maintained by selection where fitnesses vary in time, without heterozygote advantage or frequency dependent selection.

1.1.2 Genetic Drift and finite population sizes

Genetic drift is the chance changes in allele frequency that result from the random sampling of gametes from generation to generation in a finite population.

1.1.2.1 The effect of drift

Genetic drift has the same expected effect on all loci in a genome. In a large population, on average only a small change in allele frequencies will occur as a result of genetic drift. However, for smaller populations, genetic drift can cause larger fluctuations in allele frequencies and may even lead to the loss of fixation of alleles purely by chance alone (Hedrick 2010; Charlesworth and Charlesworth 2010). Simulations of genetic drift reveal that small population sizes can cause replicate populations to drift apart in allele frequency. The probability that an allele goes to fixation as a result of genetic drift in a finite population is proportional to its initial frequency, assuming differential selection is not occurring. $u(q) = q_0$ Over

replicate simulated populations, the mean allele frequency does not change as a result of drift, but the distribution of allele frequencies over replicate populations does (Hedrick 2010; Charlesworth and Charlesworth 2010). Therefore, drift is often examined by considering heterozygosity or the variance in allele frequencies of replicate populations.

Consider a Wright-Fisher model population with N (diploid) individuals and assume each contributes two haploid gametes to the next generation (Crow and Kimura 1970). For an offspring individual, the probability of drawing the same allele twice from the parents is $2N[1/(2N)]^2$. The probability that they are different is $1 - 1/(2N)$. Two alleles may also be identical by descent with probability:

$$f_{t+1} = \frac{1}{2N} + \left(1 - \frac{1}{2N}\right) f_t \tag{1.5}$$

This can be rewritten and the expected heterozygosity after t generations derived:

$$H_{t+1} = \left(1 - \frac{1}{2N}\right) H_t \tag{1.6a}$$

$$H_t = \left(1 - \frac{1}{2N}\right)^t H_0 \tag{1.6b}$$

This demonstrates that each generation, heterozygosity decreases at a rate that is an inverse function of the population size, and it is possible to calculate the expected heterozygosity after t generations (Hedrick 2010; Charlesworth and Charlesworth 2010). In addition, it is possible to relate observed, heterozygosity to the difference in expected heterozygosity and the variance in allele frequency. Taking account of this into the above equations and rearranging produces a formula for for the variation in allele

frequencies at time t. The formula shows that as the number of generations increases, the variance approaches a maximum value of $p0q0$. This Wright-Fisher model assumes parents produce many gametes and zygotes, and of those N are chosen to form the next generation. It is implicit that individuals are hermaphrodites and there is a small probability of self-fertilization. The mean time until fixation of an allele due to drift depends on initial frequencies of the allele and the initial frequency of the allele (Hedrick 2010; Charlesworth and Charlesworth 2010). As population size increases, the effect of drift becomes smaller as it takes more consecutive chance increases of an allele to fix it in the population. For any given population size, the lower the initial allele frequency is, the longer it is for that allele to become fixed by drift. With new neutral mutants, the expected time to fixation is four times the population size.

Explanations of drift often mention the population size N. However, in many situations the relevant value is the number of breeding individuals. This may be very different from the census population size. The concept of an effective population size makes it possible to consider an ideal population of size N in which all parents have an equal expectation of being a parent of any individual progeny. i.e the Wright-Fisher model. Effective population size can be measured by different methods: inbreeding, variance, and eigenvalue. When a population remains the same size these measures are similar, however they may differ when populations are growing or shrinking (Kimura and Crow 1963; Waples 2002). The effective population size can be influenced by the frequency of different sexes in a population, variance in reproduction, and varying numbers of individuals over several generations.

Bottlenecks and founder events are two specific cases where a population changes size significantly, influencing the effective population size. A bottleneck describes a situation in which something occurs to drastically

reduce the number of individuals which survive in a population, or otherwise get to contribute to the next generation of the population. Typically, these are events such as natural disasters, overwintering, or epidemics. A founder event describes a situation in which a population is started from a low number of individuals, for example individuals being carried to a new island or location. In both cases, these events can cause large random changes in allele frequencies, resulting in lower heterozygosity and fewer alleles than the ancestral population. The changes in allele frequencies resulting from bottlenecks and founder events generate genetic distance between two populations, equation 1.7 gives the standard genetic distance (Nei 1987) after a bottleneck or founder event, where t is the the number of generations the event lasted (Chakraborty and Nei 1977).

$$D_t = -\frac{1}{2}\ln\left(\frac{1-H_0}{1-H_t}\right) \tag{1.7}$$

1.1.2.2 Drift and selection

In a finite population, when there is no differential selection at a locus, an allele may become fixed or lost as a result of genetic drift.

In a population of infinite size, by definition there is no genetic drift, and selectively favored alleles increase in frequency and asymptotically approach fixation. Detrimental alleles always reduce in frequency and approach loss. In finite populations however, because of the effects of genetic drift, alleles may not always be fixed when they are favorable, and detrimental alleles may be fixed despite their detriment. The probability of a favorable allele in a finite population is a function of the initial frequency of the allele, the extent to which selection favours that allele, and the size of the population. Kimura 1962 developed an equation that takes these factors to compute the probability of fixation of A_1 (Kimura and Ohta 1971). The probability of fixation of an allele is a function of its initial frequency, the level

of dominance, the effective population size, and its selective advantage. The probability of fixation of an allele increases with increasing initial allele frequency and with increasing Ns (the product of population size and selection coefficient). When $Ns << 1$, this indicates that $s << 1/N$ and that the selective advantage of an allele is very low. In this case, changes in allele frequency are determined by drift. When $Ns >> 1$, then s is higher than $1/N$ and changes in allele frequency depend more on selection than on drift. The effect where alleles with low selection coefficients (and hence only slightly deleterious effects), may act as if they were neutral in small populations was first identified by Wright 1931, and described in terms of molecular evolution by Ohta 1973, who called it the nearly neutral model.

In a neutral situation in a finite population, the loss of heterozygosity is $1 - 1/(2N)$. For any given balancing selection regime, the decay in heterozygosity can be defined as $H_{t+1} = (1 - d)H_t$, where d is the loss from unfixed allele frequency states and the gain for the absorbing states. With no selection, d is $1/(2N)$ i.e. the expression reduces to the neutral model of heterozygosity loss as a result of drift already described. The ratio of decay for a neutral locus over one undergoing selection is called a retardation factor (Robertson 1962). This factor is one when there is neutrality, but when d is less than 1, then selection can slow the rate of fixation, or when $d > 1$, then selection is increasing the rate of fixation. Even though selection may be balancing in an infinite population, in a finite population, less genetic variation may be retained than in a population with no selection. Populations with heterozygote advantage, and unequal homozygote fitness values genetic variation is eliminated faster than in populations with neutrality.

1.1.2.3 Impact of genetic drift

Genetic drift needs to be considered when studying plant pathogens and organisms in very dynamic environments, as those populations may experience periodic population expansions or contractions. Analysis of Q_{ST} values of eight traits, and F_{ST} values of eight neutral loci of the pathogenic fungus *Rhynchosporium commune* revealed that the majority of the traits analysed were evolving according to stabilizing selection, although a trait for growth at 22 degrees centigrade was subject to diversifying selection and local adaptation (Stefansson, McDonald, and Willi 2014). This was proposed to be due to the fact the pathogen exists in large rather homogeneous environments (i.e. homogeneous monoculture systems) where they mostly experience one host genotype, and therefore stabilizing selection plays a greater role than does drift or directional selection. Furthermore, the cycles of frequency dependent selection and maintenance of diversity previously described would only be expected to occur if there were some allelic diversity - rare advantageous alleles - in the host. Other plant pathogens have been significantly affected by changes in their population size. For example, the global pandemic of *Phytophthora infestans* was initiated by a single clone, which escaped to North America, and then to Europe, and then to the rest of the world (Goodwin, Cohen, and Fry 1994). Analyses of RFLP loci of the pathogen *Mycosphaerella graminicola* isolated from different locations, indicated that Mexican and Australian populations have low gene diversity (Zhan, Pettway, and McDonald 2003), consistent with founder events and genetic drift. Steele et al. 2001 found that in Australia, *Puccina striiformis* originates from a single founder event, the founding race identified corresponded to a race previously identified in Europe.

1.1.3 Mutation

Mutation is the alteration of the nucleotide sequence of the genome of an organism. Mutations may be caused by errors in the DNA replication process, the insertions of a transposable element, chromosome breakage, and errors in meiosis. Mutations may be be caused by chemicals or radiation, and these mutagens cause certain kinds of mutation, for example, ultraviolet light (Kozmin et al. 2005).

Many spontaneous mutations may have detrimental effects as they affect the normal functioning of a gene. However, many mutations have neutral or almost neutral effects, as they do not result in changes to proteins or otherwise change DNA only slightly (Grauer and Li 2000). A few mutations will confer beneficial effects and change proteins in a way that enhances the fitness of organism with the allele. Of course whether or not a mutant is beneficial, deleterious, or neutral also depends on the environment (Grauer and Li 2000).

Typically, the term mutation is often used to describe the smaller scale mutations which give rise to a new allele or sequence, larger alterations are often referred to as copy number variations, structural variations, or chromosomal abnormalities (Grauer and Li 2000; Hedrick 2010). A mutation may involve a change in one nucleotide base, or it may involve changes in several nucleotides. Short mutations where a few nucleotides are removed or inserted into the DNA sequence are called indels, which may cause a frame-shift mutation if the number of bases inserted or deleted is not a multiple of three. The change affects the grouping of nucleotides into codons, affecting the reading frame or possibly introducing a stop codon. Both base mutations and indels can cause a change in the protein produced transcription and translation of the gene (Grauer and Li 2000). Transposable elements are portions of DNA that can replicate themselves and move

location within the genome of an organism (Grauer and Li 2000; Wicker et al. 2007). 60% of the maize genome and 15% of the *Drosophila melanogaster* genome consists of transposable elements (Biémont and Vieira 2006). Transposable elements have been characterized as junk, neutral, and agents of mutation and adaptation. Their behavior ranges from that of an extreme parasite, to that of a mutualist depending on the transposable element, the organism, and the area of the genome affected by one (Grauer and Li 2000).

To understand genome evolution, mutation by gene duplication, deletion, and gene conversion are important. Many genes such as globins, histones, enzymes, and MHC genes are members of multigene families. Such families are composed of several homologous genes, with similar function, and are often situated close together on a chromosome i.e. they are closely linked (Hedrick 2010). Such multigene families are thought to have evolved through serial duplication of an ancestral gene. Duplicate genes may cause dosage effects, or they may diverge, resulting in new functionality (neofunctionalisation), or they may retain only a subset of their original functionality (subfunctionalisation). Further duplication and deletion of genes may occur through unequal crossing over or gene conversion (Grauer and Li 2000). Gene conversion is a process by which the nucleotide sequence of one allele or allele segment is replaced by a homologous sequence from another allele. Voordeckers et al. 2012 demonstrated how the MALS family of genes, which code for proteins specialised to act on disaccharides, were likely to have evolved through duplication of an ancestral gene. By reconstructing the ancestral genes, and testing their activity on different substrates, they found the ancestor was mostly active on maltose like substrates, but had some function on isomaltose like sugars. Duplication and mutation resulted in a series of enzymes specialised for different substrates. Many species of plant pathogens have genomes rich

in both repeats and transposable elements (Raffaele and Kamoun 2012; Kemen and Jones 2012) and it is therefore suspected they play a role in the evolution of effector repertoires and can influence the expression of effectors (Whisson et al. 2012).

Mutations may occur anywhere across the genome stochastically, according to a mutation rate, however there are hotspots in the genome which experience mutations more often than other regions. Research into *E.coli* by Shee, Gibson, and Rosenberg 2012 has indicated such hotspots can be caused by double strand breaks in DNA which then lead to stress induced mutagenesis. In the plant pathogen *Neurospora crassa* duplicate sequences in DNA are detected and mutated during its sexual phase. The mechanism could cause linked duplicated genes to diverge further than unlinked ones (Cambareri, Singer, and Selker 1991).

It is often assumed that likelihood of mutation occurring is unaffected by selection, however there are exceptions. In microorganisms it is known that mutator phenotypes can arise (Barrick et al. 2009). These increase the number of mutations occurring in the population, and facilitate the adaptation of large asexual populations to new conditions, even when the frequency of the mutators is low. Such hyper-mutation can be genetically inherited, or can be transient. Clinical isolates of many pathogens such as *E. coli*, *Streptococci spp.*, and *Staphylococci spp.* have been found to contain high proportions of hypermutators (Jayaraman 2011). Localization of the hyper-mutation to contingency genes or specific regions of the genome limit the risk of accumulating too many detrimental mutations through hyper-mutation (Jayaraman 2011). In the case of an inheritable hyper-mutator allele, it may increase in frequency in a population through hitchhiking; it is physically linked to a selectively beneficial mutation it caused to occur (Giraud et al. 2001). Several models demonstrating how hypermutators persist and succeed exist (Taddei and Radman 1997; Tenaillon et al. 1999),

and Hyper-mutation is particularly beneficial strategy for microorganisms that are exposed to frequent and possibly unpredictable stresses (like pathogens) (Visser 2002; Tanaka, Bergstrom, and Levin 2003).

Mutation is an important evolutionary force that generates the variation the other forces act on. Several mechanisms in microbes and pathogens have been described through which such variation is generated, in addition to ways in which an organism might increase the rate at which this variation is generated during times of stress for for certain alleles. Next the effects mutation has on populations and how it exists in balance with previously described forces is presented.

1.1.3.1 Effect of mutations on populations

The effect of mutation on population allele frequencies can be evaluated by assuming a forward-backward model of mutation (Hedrick 2010). In this model, two types of allele are possible, a wild type allele (A_1) and a detrimental mutant (A_2). In addition, mutation is reversible and may change wild type alleles to the mutant alleles (forward mutation), and the mutant alleles may mutate back to the wild type (backward mutation). It is assumed forward mutations are more common than backward mutations. This is because forward mutations are mutations that resulting in gene malfunction. It is assumed only a limited number of possible mutations could compensate for such forward mutations and result in a backward mutation. Mutation from A_1 to A_2 occurs at a rate u, and mutation from A_2 to A_1 occurs at rate v. The change in frequency of A_2 due to only mutation is $\Delta q = up - vq$. This expression is linearly related to the allele frequency, but as u and v are small - mutation rates are typically low - mutation does not significantly affect the proportion of alleles in the population (Hedrick 2010). An equilibrium is achieved if the forward and backward mutation rates are equal, and if u is higher than v then it is expected that the frequency of detrimental alleles

519 would be higher than the wild type alleles (Hedrick 2010). However this
520 expectation is not realistic as it does not consider selection.

521 When mutations occur, they are the only copy in the entire population. All
522 the individuals in the population immediately after mutation are homozygous
523 for the wild type allele (A_1A_1), and the mutant is heterozygous (A_1A_2). This
524 one heterozygous individual must mate with a homozygous individual. The
525 new mutant may be lost, only homozygous wild type offspring may be the
526 outcome, or some offspring may be heterozygous with the new mutant
527 allele. If mating results in only one offspring, then there is a 50% chance it
528 is A_1A_1, and if A_1A_2 is the result, then there is still only one A_1A_2 individual
529 in the population. If mating results in two offspring, then the probability of
530 loosing A_2 is halved. So the frequency of A_2 in generations following the
531 mutation event depends on how many progeny are the result of mating,
532 and what type they are (Hedrick 2010).

533 The way in which purifying selection keeps detrimental alleles from
534 increasing in frequency has previously been described. The entire genome
535 is subject to the opposite effects of mutation and selection, and the joint
536 effects of mutation and selection is called the mutation-selection balance.
537 Assume that A_2 is deleterious and recessive, selection will act to reduce the
538 frequency of A_2 as previously described. Equation 1.8 rewrites 1.3 using
539 the fitness values for a recessive deleterious allele from table 1.1 (Hedrick
540 2010).

$$\Delta q_s = \frac{sq^2p}{1 - sq^2} \tag{1.8}$$

541 The increase in q due to mutation then is $\Delta q_{mu} = up$, and assuming
542 back mutation occurs at a low rate compared to u, as these forces have
543 opposite effects, there is a point where they are at equilibrium (equation 1.9)
544 and the total change in allele frequency is $\Delta q = \Delta q_{mu} + \Delta q_s = 0$ (Hedrick

2010).

$$up = \frac{sq^2p}{1 - sq^2} \tag{1.9}$$

If it is assumed that q^2 is small then equation 1.9 can be solved for the equilibrium genotype frequency ($q_e^2 = u/s$), and the equilibrium allele frequency ($q_e = \sqrt{u/s}$). This frequency is increased as a result of either higher mutation rate or lower selective disadvantage. If the deleterious mutant were not completely recessive, the level of dominance h can affect q_e. If h is much larger than 0 and q_e is small, then equilibrium allele frequency is approximately u/hs, and assuming p is almost 1, the frequency of the mutant phenotype at equilibrium is $2u/s$. As a general rule, as the level of dominance increases, the equilibrium allele frequency rapidly reduces (Hedrick 2010).

Mutations will contribute to the genetic load of a population, reducing its fitness from the maximum possible. For a deleterious recessive mutation the load is $L = sq^2$ and at equilibrium $u = sq^2$, load is roughly equal to the mutation rate. If the deleterious mutant is dominant, then load becomes $L = 2u$ which shows that depending on the level of dominance, the mutation load can be between the mutation rate and twice the mutation rate. If independence of fitness between loci is assumed, the fitness at locus i may be defined as $\bar{w}_i$, and the overall fitness of the population is defined ad $\bar{w} = \bar{w}_i^n$. The overall load is $L = 1 - \bar{w}$. Crow and Kimura 1970 gave a formula for approximating the total load caused by mutation:

$$L \approx C \sum u_i \tag{1.10}$$

Where C is a constant between 1 and 2 and u_i is the mutation rate of the locus i.

Joint consideration of mutation and drift forms the basis of the neutral

theory. The initial frequency of a new mutant A_1 in a population of A_2 alleles has an initial frequency of $p_0 = \frac{1}{2N}$. The two alleles are neutral respective to each other, thus the probability of this mutant being fixed in the population is equal to its initial frequency as described in section1.1.2, and the probability of losing the mutant from the population is $u(q) = 1 - \frac{1}{2N}$. Unless a population is very small, a new neutral mutation is likely to be lost from the population by drift alone (section 1.1.2). Loss of a mutant due to drift occurs more quickly than fixation. This is because the change in frequency necessary to lose a new mutant is much smaller than that necessary to fix the new mutant. Kimura and Ohta 1971 formulated the average time to fixation and loss of a new mutant due to drift alone:

$$T_1(p) = 4N_e \tag{1.11a}$$

$$T_0(p) = 2\left(\frac{N_e}{N}\right)\ln(2N) \tag{1.11b}$$

Assuming $N = N_e$ then the time to loss reduces to $2N/[\ln(2N)]$. As a result, polymorphism is often transient. Mutation acts to increase the number of alleles, whereas drift acts to reduce the number of alleles. The properties of this equilibrium for the infinite alleles model were explored by Kimura and Crow 1964 using the inbreeding coefficient. Recall that equation 1.5 gives the expected inbreeding coefficient. This may be modified by the probability both alleles did not mutate:

$$f_t = \left[\frac{1}{2N_e} + \left(1 - \frac{1}{2N_e}\right) f_{t-1}\right](1-u)^2 \tag{1.12}$$

Setting $f_0 = 1$ (heterozygosity $H_0 = 0$) and $u = 10^{-5}$ and examining the change in heterozygosity over many generations for various values of Ne it can be shown that it takes many generations, but eventually heterozygosity rises to approach an asymptotic value. Furthermore, the asymptotic level of heterozygosity is greater when N_e is greater. As a consequence, when

593 population size is small, the rise to the smaller asymptotic value occurs
594 more quickly as genetic drift has a greater impact on the genetic variation
595 change than does mutation (Kimura and Crow 1964; Hedrick 2010). If an
596 equilibrium between mutation adding variation and drift eliminating variation
597 from a population is assumed $f_t = f_{t-1} = f_e$, formula 1.12 reduces to:

$$f_e \approx \frac{1}{4N_e u + 1} \tag{1.13}$$

598 Because $H = 1 - f$, equilibrium heterozygosity for the infinite allele
599 neutral model can be obtained, where $\Theta = 4N_e u$:

$$H_e = \frac{\Theta}{\Theta + 1} \tag{1.14}$$

600 This equilibrium is different to equilibrium previously described, as the
601 allele frequencies are constantly changing, but the distribution of alleles
602 remains mostly constant. The above equation demonstrates that when
603 $\Theta \approx 1$, then $H_e \approx 0.5$. When $\Theta \gg 1$ then mutation primarily affects
604 heterozygosity rather than drift and so H_e is quite high. The opposite is
605 true, when $\Theta \ll 1$ then drift is the major determinant of heterozygosity and
606 H_e is low (Kimura and Crow 1964; Hedrick 2010).

607 To examine the effect of a population bottleneck, assume a population
608 starts at mutation-drift equilibrium. The population goes through a bot-
609 tleneck and grows large once again (Nei 2005). The expected genetic
610 variation after the bottleneck depends on heterozygosity prior to the bottle-
611 neck, the size of the bottleneck, and the rate of increase after the bottleneck
612 (Nei, Maruyama, and Chakraborty 1975). The size of the bottleneck has a
613 large effect on the number of alleles in a population, but average heterozy-
614 gosity is mostly affected by the rate of growth after the bottleneck. This is
615 because whilst heterozygosity is reduced by the decrease in population size,
616 when growth of the population after the bottleneck is slow, heterozygosity

is lost each generation until it is large enough. Faster population growth rates allow populations to rebound as loss of heterozygosity only occurs during the first few generations following the bottleneck (Nei, Maruyama, and Chakraborty 1975).

Mutations can have selective effects. When s is less than $1/(2N)$ genetic drift is the stronger factor affecting allele frequency than selection and the mutant behaves neutrally, and deleterious mutants may become fixed as if they were neutral in small populations (Kimura 1983; Lynch and Gabriel 1990; Lande 1994). Over time, fitness declines which can lead to further reductions in population size, and hence mutations of increasingly detrimental effect behave as if they are neutral, and are more likely to be fixed. Such a feedback is called mutation meltdown, and in theory could make small populations go extinct, (Lynch, Conery, and Burger 1995).

1.1.4 Population structure and gene flow

Populations may be split into subpopulations due to geographical, ecological, or behavioral factors. When a population is divided or there is more than one population, the amount of genetic exchange, or gene flow, between the subpopulations may differ between the different populations or subpopulation. When gene flow is high between two populations or subpopulations, they are highly connected genetically and the amount of genetic variation between them is homogenized. Conversely, when the amount of gene flow is low between populations or subpopulations, then genetic drift, selection, and mutation in the populations and subpopulations may lead to genetic differentiation (Charlesworth and Charlesworth 2010; Hedrick 2010).

Some types of movement of individuals like migrations will not actually result in gene flow, especially if the individual is only transiently passing through a population and does not breed with members of the population

(Hedrick 2010). Gene flow may be distinguished from simple migration as movement between groups that results in genetic exchange (Endler 1977).

When considering population subdivision it is often assumed that the subpopulations are always present. Another view assumes they can die out, but they are repopulated from neighboring subpopulations, this is termed a metapopulation (Hanski 1998), and the dynamics of extinction and re-population make metapopulations differ from the basic concept of a subdivided population. What follows is a basic description of how gene flow effects populations using a simple genetic model, before the joint effects of gene flow and drift, and gene flow and selection are considered.

The continent-island model models a situation in which a large continent population is connected to a smaller island population (Charlesworth and Charlesworth 2010). The smaller island population receives migrants from a larger continent population. The larger continent population is assumed to be large enough to render the effect of genetic drift negligible compared to the effect of gene flow. Gene flow is assumed to have negligible effect on the source population. In this model, the proportion of migrants moving to the island is m, and the proportion of residents in the island population is $1m$. The proportion of A_2 in the migrants coming from the continent is q_m and the frequency of A_2 on the island before the gene flow is q_0 (Hedrick 2010).

Frequency of A_2 on the island after gene flow is calculated as:

$$q_1 = (1 - m)q_0 + mq_m \tag{1.15}$$

Formula 1.15 can be reduced to $q_0 - m(q_0 - q_m)$.

The change in frequency of q is then defined as:

$$\Delta q = q_1 - q_0 \tag{1.16}$$

Formula 1.16 reduces to $-m(q_0 - q_m)$.

q_m and m are assumed to be constant (Hedrick 2010). From these equations it is clear that $m = 0$ then there is not migration from the continent to the island and so there is no change in allele frequency. If $q_0 < q_m$ then the frequency of q increases on the island. If $q_0 > q_m$ the frequency decreases. This indicates that there is a stable equilibrium freq of A_2 at $q_m = q_0$.

A general formula to calculate the frequency of A_2 for any generation t has been derived as:

$$q_t = (1 - m)^t q_0 + [1 - (1 - m)^t] q_m \tag{1.17}$$

In this formula, as t increases the first term approaches 0, and the second term approaches q_m (Hedrick 2010). Therefore eventually the frequency of A_2 in the island population converges to the frequency of A_2 in the continent population. This is because gene flow is unidirectional, and therefore eventually all in the island population are descended from migrants. Thus, the allele frequencies approach that of the continent i.e. the source of the migrants (Charlesworth and Charlesworth 2010). In this model, allele frequency changes at a maximum rate initially, and as the equilibrium is approached, it decreases.

A more general model assumes gene flow can occur among all parts of a structured population. The model assumes there is k different subpopulations, and that the proportion of individuals migrating from a subpopulation i to another subpopulation j is m_{ij} (Hedrick 2010). The values of m_{ij} then can form a matrix called a backward migration matrix (Bodmer and Cavalli-Sforza 1968). In this matrix, the proportion of residents (i.e. not migrants) in each subpopulation i are given by the diagonal values of the matrix (i.e. m_{ii}). Each row of the matrix sums to 1, because it describes the proportion

of migrants coming into a population i from the other j populations. For this model, the amount of allele A_2 in any subpopulation i after gene flow is:

$$q_i' = \sum_{j=1}^{k} m_{ij} q_j \tag{1.18}$$

To process of allele frequency change over time can be described with matrix notation, where M is the migration matrix, and Q_t is the vector of allele frequencies in each population at generation t:

$$Q_{t+1} = MQ_t \tag{1.19}$$

The above can be generalized for any t

$$Q_t = M^t Q_0 \tag{1.20}$$

(Hedrick 2010)

In this model, as with the continent-island model previously described, after a period of time, allele frequencies in the subpopulations converge and approach an asymptotic value. This value can be calculated with equation 1.18 using a migration matrix raised to a power of t large enough that all elements have reached their asymptotic values. This demonstrates the homogenizing effect gene flow has on populations when it is sustained for a period of time (Charlesworth and Charlesworth 2010; Hedrick 2010).

1.1.4.1 Gene flow - drift balance

Gene flow acts to homogenize populations as described above. However populations are finite in size and so genetic drift will cause differences between the populations through the random fixation and loss of alleles. The joint effects of gene flow and drift can be examined using a simple model of replicate island populations (Wright 1940). Each island has N

individuals and receives a proportion of migrants each generation m, from a continent population.

When the gene flow between islands, and the population size of the islands are large the allele frequencies on the islands behave as previously described: they will converge to the frequencies of the continent. However if population sizes are small, and the amount of gene flow is low, then the allele frequencies of the islands may differ from each other (Hedrick 2010). So genetic drift causes allele frequencies in subpopulations to drift apart, whilst gene flow acts to homogenise the allele frequencies: Take N to be equal to N_e, the probability two alleles coalesce in generation $t-1$ is $1/(2N)$ and the probability that they do not is $1-1/(2N)$ (Hedrick 2010). The expected homozygosity in generation t can be given as:

$$f_t = \frac{1}{2N} + \left(1 - \frac{1}{2N}\right) f_{t-1} \tag{1.21}$$

This expression can be modified by the probability that both alleles are not migrants:

$$f_t = \left[\frac{1}{2N} + \left(1 - \frac{1}{2N}\right) f_{t-1}\right](1-m)^2 \tag{1.22}$$

Assuming there is an equilibrium between gene flow homogenizing variation, and drift generating variation, then $f = f_t = f_{t-1}$ and $f = F_{ST}$, then

$$F_{ST} = \frac{(1-m)^2}{2N - (2N-1)(1-m)^2} \tag{1.23}$$

(Hedrick 2010)

F_{ST} is the fixation index, a measure of genetic differentiation over subpopulations. When $m = 0$ then $F_{ST} = 1$, and when $m = 1$, then $F_{ST} = 0$. In other words when levels of gene flow are high, the genetic differentiation over subpopulations is low. Ignoring the powers of two, and reducing the

formula, F_{ST} can be approximated:

$$F_{ST} \approx \frac{1}{4Nm + 1} \tag{1.24}$$

Assuming k subpopulations, the differentiation between populations can be given as

$$G_{ST} = \frac{1}{4Nm\left(\frac{k}{k-1}\right)^2 + 1} \tag{1.25}$$

(Slatkin 1995). In both equations, Nm means the absolute number of migrants entering a population every generation.

F_{ST} for any generation t has been derived when $m = 0$

$$F_{ST(t)} = 1 - e^{t/2N} \tag{1.26}$$

(Wright 1943).

The above expression is 0, when subpopulations are not very separated in early generations, and reaches a maximum of 1 as subpopulations are separated by drift. The smaller the population size, the faster the subpopulations diverge due to drift. The increase in F_{ST} is fastest for the first $2N$ generations, after which time it approaches the maximum of 1.

Iterating over formula 1.22 allows examination of the rate of approach to equilibrium for different values of N and m. When population size is large and the amount of gene flow is large, then approach to equilibrium is fast, but when populations are large and gene flow is small, then the approach to equilibrium is slow (Hedrick 2010).

Population subdivision also affects the N_e of populations. For the island model:

$$N_e = \frac{kN}{1 - F_{ST}} \tag{1.27}$$

If F_{ST} is low, then $N_e \approx kN$, but if gene flow is low then N_e might be larger than kN (Wright 1943).

Wright 1940 gave an explicit method of estimating allele frequencies incorporating the effects of gene flow and drift for the island model. Assuming the frequency of A_2 in migrants (q_m) is constant, when observing a large number of islands, their average allele frequency will be q_m, but depending on drift and gene flow, the distribution over the islands will vary. The shape of the distribution depends on $4Nmq_m$ and $4Nm(1-q_m)$. With large amounts of gene flow and large population sizes, the allele frequencies over the islands will not depart far from the mean (Hedrick 2010). However, with lower $4Nmq_m$ and $4Nm(1-q_m)$, and if $q_m = 0.5$, then the distribution takes on a U shape: Drift plays a greater role in determining allele frequencies as alleles enter the islands by gene flow, and islands become temporarily fixed for either A_2, or instead for A_1.

Other models add an extra consideration by assuming different populations occupy positions in space, and that gene flow is restricted to certain routes or directions. For example, the stepping stone model arranges populations in a one dimensional structure, and restricts gene flow to occurring only between populations that are adjacent in that one dimensional space (Hedrick 2010). The effective population size of such a linearly divided population can be approximated as $N_e \approx kN$ (Maruyama 1970). If populations are distributed across a landscape according to available habitat, then there may be distance-dependent gene flow between the populations. In such case, expected patterns of genetic variation may be similar to the stepping stone models (Wright 1943). It has been suggested that the amount of genetic divergence as estimated with Nm or $F_{ST}/(1-F_{ST})$ should change as an inverse linear function of geographic distance (Nm), or as a linear function of geographic distance ($F_{ST}/(1-F_{ST})$) (Rousset 1997).

In metapopulations (Levins 1969), the dynamics of recolonization and

extinction greatly influence N_e, the genetic variation present in the metapopulation, and the distribution of genetic variation over the subpopulations (Slatkin 1977; Hedrick and Gilpin 1997; Whitlock and Barton 1997; Nunney 1999). Many parameters can influence the rate at which genetic variation is lost, for example, the source of individuals recolonizing a previously extinct path might be from a single path, or a group of individuals from all other non-extinct patches. A metapopulation with 20 patches, an infinite population size in each patch, and no gene flow except during recolonization, will have an effective size of 150 when recolonization of a patch is from a single female from another patch. This low N_e is due to the low number of founders in each recolonization (Hedrick and Gilpin 1997; Hedrick 2010).

1.1.4.2 Gene flow - selection balance

Gene flow and selection are often both important forces driving allele frequencies in a population. Both forces are diverse in their effects on allele frequencies and so the interaction of the two forces can lead to complex results (Lenormand 2002). Therefore, only a simple scenarios of selection and gene flow is introduced here.

Consider again the continent-island model, if the change in allele frequency due selection is Δq_s, and the change in allele frequency due to gene flow is Δq_m, then the change in allele frequency due to the joint effect of the two forces is $\Delta q = \Delta q_m + \Delta q_s$(Hedrick 2010). Assuming the fitness values of A_1A_1, A_1A_2, and A_2A_2 are 1, $1-s$, and $1-2s$ respectively, then Δq can be expressed as $\Delta q = sq^2 - (m+s)q + mq_m$ (Li 1976). When $\Delta q = 0$, there is equilibrium, and the equilibrium frequency is found by solving the quadratic equation.

$$q_e = \frac{1}{2s}\{(m+s) \pm [(m+s)^2 - 4msq_m]^{1/2}\} \tag{1.28}$$

A_1 is favored if s is positive, otherwise A_2 is favored (Hedrick 2010). There are three main scenarios to consider, one where gene flow is much less, greater than, or equal to the absolute value of selection ($|s|$). As m increases with respect to $|s|$, genetic differentiation does not occur. This is intuitive, as gene flow has a homogenizing effect as previously described, and with increasing m, its effects become more influential than the effects of selection, and the island's equilibrium frequency approaches that of the migrants coming from the continent (Li 1976).

Generally, the equilibrium frequency of an island depends on the selective advantage, the level of dominance on the island, and the amount of gene flow. With high amounts of gene flow, even a favorable variant can be lost from an island, no matter its level of dominance. This is called patch disappearance (Haldane 1948). Thus gene flow is a force which limits selection and local adaptation (Lenormand 2002).

1.1.4.3 Importance of gene flow

Gene flow and genetic structuring significantly influence plant pathogen and marine plankton populations. Gene flow is the force which introduces new virulence alleles into a new agricultural field, far from the source of original mutation. Plant pathogen populations are often made up of one or a few clonal lineages which differentiate themselves from other populations (in chapter 3 these are called 'races') (Koenig, Ploetz, and Kistler 1997). In such populations, it may help instead to think of genotype flow rather than gene flow because of the high degree of linkage. Genotype flow refers to the movement of entire genotypes between distinct populations. Since many plant pathogens have an asexual stage and a sexual stage, both genotype flow and gene flow can occur. An existing example of gene flow between plant pathogen populations is provided by Zhan, Pettway, and McDonald 2003, who demonstrated that *Mycosphaerella graminicola*

populations shared RFLP alleles, but no two populations had completely identical fingerprints, indicating that gene flow, but not genotype flow, was occurring. An example of genotype flow is the global movement of a single clone of *Phytophthora infestans*, out of Mexico in the 1840's as previously described. Only one mating type escaped and spread globally, and as the organism has two mating types, sexuality was not possible until the other mating type escaped in the 1970's (Goodwin et al. 1992; Goodwin, Cohen, and Fry 1994; Goodwin et al. 1995).

There is substantial evidence of genetic structuring in marine plankton populations despite the high dispersal capacity of those organisms that might usually lead one to expect high levels of gene flow (Sildever et al. 2016). Oceanographic features like currents and eddies will create habitat heterogeneity which in turn leads to genetic population structuring (White et al. 2010; Sanford and Kelly 2011; Casabianca et al. 2012), as do chemical and biotic properties of the oceans such as pH levels, temperature, salinity, and the presence or absence of predators and parasites (Cousyn et al. 2001; Decaestecker et al. 2007; Weisse et al. 2007; Yampolsky, Schaer, and Ebert 2014; Defaveri and Meril 2014). All these factors may cause local adaptation resulting in population structuring.

1.1.5 Recombination and linkage

In the theory introduced so far, it has been assumed that alleles at a locus under consideration are transmitted independently of any alleles at any other loci. This is called independent assortment (Hedrick 2010). It was also assumed that the fitnesses of genotypes at any given locus were independent of the fitnesses of other genotypes at other loci. However, this simplification is not valid in the majority of cases. The transmission of genetic variants does not occur independently of other genetic variants. This is because of linkage between genetic variants; variants are distributed

across DNA molecules, and two variants situated on the same molecule are said to be physically linked. The non-random association of alleles is called linkage disequilibrium (LD) (Lewontin and Kojima 1960). The amount of LD is generally an inverse function of the rate of recombination. Where recombination is the rearrangement of genetic material, especially by crossing over in chromosomes or by the artificial joining of segments of DNA from different organisms

If one assumes a large randomly mating population has two alleles at one locus A (A_1, A_2), and two alleles at a second locus B (B_1, B_2), then four gametes or haplotypes are possible: A_1B_1, A_1B_2, A_2B_1, and A_2B_2. The frequencies of these four haplotypes are denoted as x_{11}, x_{12}, x_{21}, and x_{22}. The frequencies of each allele are $p_1 = x_{11} + x_{12}$, $p_2 = x_{21} + x_{22}$, $q_1 = x_{11} + x_{21}$, and $q_2 = x_{12} + x_{22}$ for A_1, A_2, B_1, and b_2, respectively (Lewontin and Kojima 1960).

Assuming random association between alleles in gametes, then the frequency of each gamete is equal to the product of the frequencies of the alleles it is made of. In other words $x_{11} = p_1q_1, x_{12} = p_1q_2, x_{21} = p_2q_1, x_{22} = p_2q_2$. However, when this assumption does not hold and there is nonrandom association between alleles, the frequencies must be written as a function of these expected frequencies, with some deviation D from the expectation. Therefore, $x_{11} = p_1q_1 + D, x_{12} = p_1q_2 - D, x_{21} = p_2q_1 - D, x_{22} = p_2q_2 + D$. D is the LD parameter and it is a measure of the deviation from random association between alleles at different loci, $D = x_{11} - p_1q_1$ (Lewontin and Kojima 1960). In other words it is the observed frequency of a gamete, minus the expected frequency of the gamete. By substituting values p_1 and q_1, D may be written as:

$$D = x_{11}x_{22} - x_{12}x_{21} \tag{1.29}$$

The gametes can be categorized as coupling or repulsion gametes. Coupling gametes are those with alleles of the same subscript, and repulsion gametes are those with alleles with different subscripts. D then is the product of the frequencies of the two coupling gametes, minus the product of the frequencies of the repulsion gametes (Hedrick 2010).

From these four gametes, 10 genotypes are possible. The genotypes and their expected proportions are listed in Table 1.2. These derivations make sense given that A_1B_1/A_1B_1 genotypes only produce A_1B_1 gametes, and that A_1B_1/A_1B_2 genotypes produce $1/2A_1B_1$ and $1/2A_1B_2$ gametes. Double heterozygotes produce gametes different from the parental gametes due to recombination, e.g. A_1B_2 and A_2B_1 gametes can be produced by recombination of A_1B_1/A_2B_2 individuals. The recombination rate is denoted as c in Table 1.2. c ranges from 0 where there is no recombination between loci A and B, to 0.5 or independent assortment. The frequency of each gamete in the next generation can be calculated the summing each of columns 3 to 6 in Table 1.2, the simplified way of working out such sums are given on the bottom line of the table, where D_0 is the initial amount of LD (Hedrick 2010).

The amount of D after one generation then is $D_1 = x'_{11}x'_{22} - x'_{12}x'_{21}$. After substitution and simplification this becomes $D_1 = (1-c)D_0$, which is recursive and so can become

$$D_t = (1-c)^t D_0 \tag{1.30}$$

with D_t meaning the amount of LD after t generations (Hedrick 2010).

With this formula we see that when there is no linkage ($c = 0.5$) most disequilibrium is lost within a few generations, and with lower recombination rate, linkage is tighter as recombination does not break up associations between alleles as frequently, and so LD does not decay as fast.

Table 1.2: Expected frequencies for different gametes in a two-allele, two-locus system, adapted from Hedrick 2010.

		Gametes of offspring			
Genotypes	Frequencies	A_1B_1	A_1B_2	A_2B_1	A_2B_2
A_1B_1/A_1B_1	x_{11}^2	x_{11}^2	–	–	–
A_1B_1/A_1B_2	$2x_{11}x_{12}$	$x_{11}x_{12}$	$x_{11}x_{12}$	–	–
A_1B_2/A_1B_2	x_{12}^2	–	x_{12}^2	–	–
A_1B_1/A_2B_1	$2x_{11}x_{21}$	$x_{11}x_{21}$	–	$x_{11}x_{21}$	–
A_1B_1/A_2B_2	$2x_{11}x_{22}$	$(1-c)x_{11}x_{22}$	$cx_{11}x_{22}$	$cx_{11}x_{22}$	$(1-c)x_{11}x_{22}$
A_1B_2/A_2B_1	$2x_{12}x_{21}$	$cx_{12}x_{21}$	$(1-c)x_{12}x_{21}$	$(1-c)x_{12}x_{21}$	$cx_{12}x_{21}$
A_1B_2/A_2B_2	$2x_{12}x_{22}$	–	$x_{12}x_{22}$	–	$x_{12}x_{22}$
A_2B_1/A_2B_1	x_{21}^2	–	–	x_{21}^2	–
A_2B_1/A_2B_2	$2x_{21}x_{22}$	–	–	$x_{21}x_{22}$	$x_{21}x_{22}$
A_2B_2/A_2B_2	x_{22}^2	–	–	–	x_{22}^2
	1	$x'_{11} = x_{11} - cD_0$	$x'_{12} = x_{12} + cD_0$	$x'_{21} = x_{21} + cD_0$	$x'_{22} = x_{22} - cD_0$

To determine how long it will take for an initial amount of LD D_0 to decay to a given amount of LD D_t the equation 1.30 can be solved to give:

$$t = \frac{\ln(D_t/D_0)}{\ln(1-c)} \tag{1.31}$$

(Hedrick 2010).

The measure of LD described is not the only one proposed (Hedrick 1987; Lewontin 1988; Devlin and Risch 1995). To examine the extent of linkage equilibrium over chromosomes, the r^2 and D' are often used and the extent of LD measured varies with the estimated amount of recombination over chromosomes (Dawson et al. 2002).

The rate of recombination c is estimated as the proportion of recombinant gametes produced from a parent with a known gamete constitution (Hedrick 2010). The amount of recombination can vary because of a few

factors. Recombination can vary between the sexes, on different chromosomes, and between different regions on the chromosomes. Regions of higher or lower levels of recombination than are expected are termed hot spots and cold spots (Arnheim, Calabrese, and Nordborg 2003; Kauppi, Jeffreys, and Keeney 2004). Patterns of LD can be used to try to putatively identify such hot and cold recombination regions and estimate rates of recombination (Stumpf and McVean 2003; Ptak, Voelpel, and Przeworski 2004; Auton and McVean 2007), and many other methods of recombination detection in DNA sequences exist. In chapter 2 more methods for detecting recombination are discussed along with presentation of the HybridCheck software.

LD can be generated by multilocus selection. For example, tightly linked members of a multigene family or supergene (Darlington and Mather 1950) may be under selection that generates linkage disequilibrium as each gene of the family is related in its adaptive function. Multigene family members are created by serial gene duplication, followed by divergence through mutation, drift, and differential selection. Therefore, they have historical association, but interacting effects between them may cause selection to maintain their association, keeping them in disequilibrium. The MHC of vertebrates has properties of both supergenes and multigene families and is in linkage disequilibrium (Edwards and Hedrick 1998; Beck and Trowsdale 2000).

LD can be influenced by genetic drift (Hill and Robertson 1968; Ohta and Kimura 1969). The effects of drift on LD can be considered by imagining the two-loci two-state model as four alleles at one locus. Drift will alter the frequency of the gametes from generation to generation similar to that of a single loci model. Thus, drift in small populations can lead to nonrandom associations between alleles at different loci (Hedrick 2010). Recombination reduces the effect of drift, reconstituting some gametes. The expected value

of the LD measure r^2 for a given effective population size N_e and a given rate of recombination between two loci c, can be expressed as:

$$E(r^2) \approx \frac{1}{1 + 4N_e c} \tag{1.32}$$

(Hill and Robertson 1968; Ohta and Kimura 1969).

With large $N_e c$, $E(r^2)$ moves towards 0, with smaller $N_e c$ $E(r^2)$ approaches 1. Just as with the single locus model, founder events and population bottlenecks can also influence LD. If N_e was small at some point in the past, the LD caused may still be present if the LD has not decayed (Hedrick 2010). With large $N_e c$, equation 1.32 is approximately

$$E(r^2) = \frac{1}{\rho} \tag{1.33}$$

where ρ is $4N_e c$ or the population recombination rate. This is analogous to the population mutation rate $\theta = 4N_e \mu$ (Wall 2000; Stumpf and McVean 2003; Padhukasahasram et al. 2006), and the expected amount of LD decreases as ρ increases (assuming that drift is the only thing affecting LD) (Pritchard and Przeworski 2001; Hedrick 2010).

Mutations may also generate low levels of LD, however recurrent mutation is unlikely to cause higher LD because as they are unlikely to occur associated with the same allele repeatedly, and any buildup of LD through mutation would occur more slowly than the process of recombination reducing LD (Hedrick 2010). However, mutation coupled with recombination and gene flow are the source of new haplotypes in populations. New genetic variants can increase in frequency by selection and drift, and hence all these factors in concert may create additional LD (Hedrick 2010). Mutations may also break up LD if the mutation rate is high enough. Assuming an allele A_1 which mutates to a disease allele A_2, creating a new gamete A_2B_1, if mutations from B_1 to any other B allele occur at rate μ, assuming

no recombination, the association between a disease allele A_2 and B_1 is broken down. This effect has been found to be especially significant for microsatellite loci, which are characterized by a high mutation rate relative to SNP and indel mutations (Payseur, Place, and Weber 2008).

Gene conversion can also affect LD, but typically only affects shorter DNA segments. Assume there is gene conversion around a gene B in an $A_1B_1C_1/A_1B_2C_1$ individual, gene conversion could result in a $A_1B_2C_2$ gamete. B_1 has been converted to B_2. This would decrease LD between A and B, and B and C. However, it would not affect LD between A and C. Many close sites do not have complete association, suggesting that reduction in LD is occurring through gene conversion (Ardlie et al. 2001). Note however that consecutive mutations can also explain the incomplete association between linked sites. For example, consider three haplotypes $A_1B_1C_2$, $A_1B_2C_1$, and $A_1B_2C_2$ in a 100bp fragment in a population sample. This observation is consistent with recombination (between the 1st and 2nd haplotype, with breakpoint between B and C, creating the 3rd haplotype). It is also consistent with gene conversion (e.g. a C_1 in an ancestral 2nd haplotype might have been converted by the C_2 of the 1st haplotype, thereby creating a novel 3rd haplotype). Finally, it is also consistent with mutation ($B_2 \rightarrow B_1$) in the ancestral 3rd haplotype ($A_1B_2C_2$), creating the 1st haplotype, and a second mutation ($C_2 \rightarrow C_1$) in another copy of the ancestral $A_1B_2C_2$ haplotype (before or after the first mutation at any point in time) resulting in the 2nd haplotype. In other words, and in contrast to Ardlie et al. 2001, the observation that many close sites do not have complete association should not be taken as evidence for gene conversion because other evolutionary forces can explain this observation more plausibly.

Gene flow can also affect LD. The amount of disequilibrium when two

populations are mixed to produce a third can be expressed as

$$D = m_x m_y (p_{1 \cdot x} - p_{1 \cdot y})(q_{1 \cdot x} - q_{1 \cdot y}) \tag{1.34}$$

where $p_{1 \cdot x}$ and $p_{1 \cdot y}$ are the frequencies of of the A_1 allele in the two populations being mixed (population x and population y), and $q_{1 \cdot x}$ and $q_{1 \cdot y}$ are the frequencies of the B_1 allele in the two populations (Hedrick 2010). For LD to be generated, the frequencies of both loci must be different in the two populations. The greater the difference, and the more equal the contributions are from each population, the more LD is generated (Hedrick 2010).

Population subdivision reduces the rate of LD decay. The reduction in heterozygotes in subdivided populations due to the Wahlund effect (Wahlund 1928) reduces the opportunity to create recombinant gametes. If the amount of gene flow is small, then it can determine the rate of LD decay (Nei and Li 1973). The amount of linkage disequilibrium has been expressed as $D \approx m/c$ (Barton et al. 2007) i.e. it is a balance between the rate of gene flow creating LD, and the rate of recombination reducing LD. Since many factors including selection, drift, gene flow and mutation affect LD, it can be difficult therefore to attribute a cause of LD without historical knowledge or data.

Since alleles are linked and selection occurs at one or more loci we say that alleles have a genetic background (Hedrick 2010). Multilocus phenomenon may explain some observations encountered in evolutionary genetics. Apparent heterozygous advantage at a given marker locus may actually be caused by association of alleles at a linked locus to the alleles at the marker locus (Ohta 1971). For example, Oosterhout 2009 proposed that the genetic variation at the MHC may be maintained by a linkage of the genetic load (or sheltered load) present at the peri-MHC region.

Recessive deleterious mutations associated with a given haplotype prevent the fixation of that haplotype in the population because these mutations would become expressed in homozygous state, reducing the fitness of that individual. In other words, an MHC haplotype is self incompatible because it expresses its genetic load in homozygous state. Assuming that each MHC haplotype has its own sheltered load of recessive deleterious mutations, this prevents their fixation in the population, and results in a balanced polymorphism (Oosterhout 2009). Recombination between MHC alleles is further reduced by negative epistasis, with selection operating against recombination because the recombinant haplotype are incompatible with both parental (non-recombinant) haplotypes.

Furthermore, changes in allele frequencies might be the result of selection acting on alleles at an associated locus to one being observed. This can result in genetic hitchhiking, selective sweeps or background selection (Charlesworth and Charlesworth 2010). Genetic hitchhiking, previously described as the mechanism by which hypermutator alleles can be indirectly selected for in clonal populations (section 1.1.3), is possible because of linkage. Neutral alleles can increase in frequency because of their association with a selected allele. The magnitude of hitchhiking depends on the extent of linkage, inbreeding, and the initial amount of LD (Thomson 1977; Hedrick 1980; Kaplan, Hudson, and Langley 1989). If there is no initial statistical association between the neutral and selected allele, there can be no hitchhiking, even if recombination rates are low. To fully understand the effect of hitchhiking, the rate of change in frequency of the positively selected allele must be known (Hedrick 2010). For example for a new advantageous recessive allele, initial increase in frequency due to selection will be low (see section 1.1.1), providing time for recombination to reduce initial LD, and thus reducing the amount of expected hitchhiking of neutral alleles. Hitchhiking can even create LD between two neutral loci

if they are associated with a third selected locus (Thomson 1977; Hedrick 1980). One of the most important effects of hitchhiking is the reduction in heterozygosity of neutral or nearly neutral variation in areas of low recombination (Maynard-Smith and Haigh 1974). This is called a selective sweep and leaves a characteristic signature in genome sequences, which can be detected to provide evidence of recent selection (Hedrick 2010).

The projects presented in this thesis are concerned with how recombination has influenced the adaptive evolution of the two species studied. Specific aspects of recombination, sex and linkage relevant to each project are introduced in detail in subsequent chapters. In the introduction to chapter 4 the advantages and disadvantages of recombination and sex are presented, to provide context to the question of why *F. cylindrus* might have abandoned sex (as is hypothesized at the start of the study). In chapter 3 the evolutionary advantages and disadvantages of introgression and hybridisation is discussed in the context of results, and there multilocus concepts are important.

1.1.6 Hybrid zones, introgression, and hybrid speciation

Gene flow and recombination can result in so called hybrid zones, a physical location where hybrid offspring of two diverged taxa occur (Hewitt 1985). A hybrid zone may form where divergence is occurring between adjacent populations of a species that was previously homogenous. Parapatric and peripatric speciation is most likely to result in hybrid zones because the divergence and speciation is driven not by geographical isolation. With parapatric speciation, changes in environmental conditions between the adjacent population can result in adaptations and reproductive isolation (Mayr 1942). Founder events and random genetic drift play an important role during peripatric speciation. Before reproductive isolation has evolved, ongoing gene flow and recombination between the two adjacent populations

could result in a hybrid zone. In this case, the hybrid zone is called a primary hybrid zone. Hybrid zones may also form as a result of secondary contact between two populations of diverged taxa which were previously allopatric and had diverged as a result of geographic isolation. In the latter case, partial pre-zygotic reproductive isolation has evolved, but this is broken down, for example due to changes in environmental conditions that could hinder conspecific mate choice. It is often difficult to distinguish between primary and secondary hybrid zones (Endler 1982).

Such hybrid zones have a cline in the genetic composition across the zone from one of the parental forms to the other, as novel alleles from either side (that is either parental population) flow into the hybrid zone. Such clines can either be gradual or stepped, and they can be observed by recording the frequency of diagnostic alleles for the parental populations, across the transect between the two parental populations (Hewitt 1985). When quantifying the cline in this way, the frequency of diagnostic alleles is often characterized by a sigmoid curve, and the width of the cline is dependent on the ratio of hybrid survival to rate of recombination (Hewitt 1985). In addition to a cline of genetic composition, hybrid zones often exhibit a higher variability in fitness within the zone. In the middle of the cline hybrizymes may also be found. Hybrizymes are rare alleles from both the parental taxa, which reach high frequencies where hybrids are formed, due to genetic hitchhiking of those alleles with alleles that contribute to hybrid fitness (Schilthuizen, Hoekstra, and Gittenberger 1999).

It is possible for alleles to flow back into the distinct parental populations through introgression (subsequent backcrossing of a hybrid individual breeding with a parental individual). As a result, they appear to present a problem for the biological definition of a species if it is defined as a population of (potentially) interbreeding individuals that produce fertile offspring, however if the two parental populations remain identifiably distinct then

there is no problem for the alternative concept of a species as taxa that retain their identity, despite gene flow (Mayr 1942).

When introgression occurs, each generation is less able to replace itself with genetically similar individuals as a result of the influx of alleles from across the hybrid zone, and this may lead to genetic assimilation and homogenization of the two parental populations (Robbins et al. 2014). However, hybridisation does not always lead to the merging and homogenizing of the two populations involved. The different evolutionary outcomes of hybridization occur through different pathways in addition to introgression, like consequences of ecology such as hybrid vigour or hybrid inferiority (Edmands 1999; Johansen-Morris and Latta 2006; Rieseberg and Carney 1998).

Hybrid vigour can lead to a slowing of the growth rates of the two parental populations, because of the competition with the more fit hybrids (Slattery et al. 2008). But equally, if the increased hybrid fitness only applies in the hybrid zone, then a stable situation occurs in which the two parental populations are not threatened with assimilation, and instead hybrid speciation may occur, whereby hybridisation leads to hybrids which are reproductively isolated from either of the two parental populations. Some hybrid zones can persist for thousands of years (White et al. 1966). This is possible as the hybrid zones are so called tension-zones. In tension zones, there is a balance between ongoing hybridisation, dispersal of parental forms, and natural selection against hybrids (hybrid inferiority). If those forces are in equilibrium, a stable tension zone persists (Bazykin 1969). Recent studies identifying the signature of admixture across the genomes of native westslope cutthroat trout, and an invasive rainbow trout, revealed genome-wide selection against the invasive alleles, and that this was consistent across environments and populations (Kovach et al. 2016). It is important to note when considering the possible paths the evolution of a

hybrid zone may take, that the different outcomes are not exclusive either/or scenarios: For example, even though a hybrid zone may be maintained by negative selection acting on hybrids, and whilst some alleles from a parental population will be prevented from flowing into the other parental population as a result of negative selection, other alleles that are neutral or positively selected for may be able to flow across the hybrid zone and into the other population (Hewitt 1985). Both of these processes are occurring at once, with the outcome varying across the genome, depending on the alleles. In this way, a hybrid zone acts as a semi-permeable barrier to the flow of alleles. Analysis of genetic and phenotypic variation across a hybrid zone of Antirrhinum, populations near the French-Spanish border is one such example demonstrating this (Whibley et al. 2006): The hybrid zone has a very steep cline in flower colour and morphology across the hybrid zone. After crossing plant morphs to determine the contribution of the EL, ROS, and SULF alleles to magenta and yellow flower colouration, they used image analysis to score the levels of pigment in the plant and a principal component analysis on pixel scores together allowed the creation of a 3D genotypic space or landscape controlling flower colour (Whibley et al. 2006). Sequencing of natural samples across the hybrid zone allowed them to identify three main haplogroups. One haplogroup was specific to the yellow morph, and the other two were found only in magenta morphs, the flower colour cline coincided with a cline in the frequency of these haplotypes. The researchers then sequences loci not involved in flower colour determination, the PAL and DICH loci, which are linked to the ROS colour determination locus. They sequences PAL and DICH loci from 18 individuals either side of the hybrid zone. They found PAL alleles fell into two distinct haplogroups, whilst DICH had no haplogroup structure (Whibley et al. 2006). Sequencing PAL and DICH alleles from individuals across the hybrid zone revealed no cline in the frequencies of these alleles, showing

they are subject to different evolutionary forces. These alleles also had no correlation with flower colour. They concluded the distribution of the two alleles reflects historical gene flow, thus the hybrid zone is a barrier to alleles determining flower colour, as F2 hybrids are less fit according to their 3D fitness landscape, but other alleles are able to pass through (Whibley et al. 2006).

Hybridization and introgression, is thought to occur in roughly 10% of animal species and 25% of plant species (Mallet 2005). Hybridization may lead to hybrid speciation, which is where new hybrid lineages become reproductively isolated from parental populations, and so are considered separate species. Genomic studies have allowed determination of the sizes of parental chromosomal blocks in introgressed populations and hybrid species (Buerkle and Rieseberg 2008; Morrell et al. 2005), as they allow observation of associations among alleles of one species in the genetic background of another, indicating recent introgression. Genome-wide studies of introgression and hybridisation have also supported the conclusions supported by the work of Whibley et al. 2006, that there is variation in the amount of introgression across genomes, and so some regions of the genome are more permeable to foreign alleles than others (Martinsen et al. 2001; Macholán et al. 2007; Scotti-Saintagne et al. 2004; Turner, Hahn, and Nuzhdin 2005; Yatabe et al. 2007).

Substantial changes can occur to a genome immediately after hybridisation, such as gene loss or silencing, changes in expression of some genes (Adams and Wendel 2005). Analysis of three synthetic sunflower hybrids and three natural sunflower hybrid species has shown large karyotypic changes can occur over a handful of hybrid generations (Karrenberg, Lexer, and Rieseberg 2007; Lai et al. 2005). The natural hybrid species also exhibit increased genome sizes of up to nearly 50% compared to the parental species (Baack, Whitney, and Rieseberg 2005). All species showed similar

increases in genome size because of the proliferation of retrotransposons (Ungerer, Strakosh, and Zhen 2006).

The evolutionary consequences of hybridisation are complex. F1 hybrids are often larger and more fit than their parents due to the effects of heterosis (Lippman and Zamir 2007), due to either overdominance or the reciprocal complementation of deleterious alleles (Clark Cockerham and Zeng 1996), this explains the establishment of hybrids but does not determine the longer term evolutionary success or failure of hybrids, which is more complex and is discusses in more detail in chapter 3. In chapter 3, processes of hybridisation and introgression, and the evolutionary outcomes of such processes are discussed in more detail, and in the context of the work presented in that chapter, which focuses on the role of such processes in the adaptive evolution of a plant pathogen species as it adapted to many hosts.

1.2 The role of Bioinformatics in population genetics

Deoxyribonucleic acid was demonstrated as the genetic material by Oswald Theodore Avery in 1944 (Russell 1988). Watson and Crick demonstrated its double helix structure composed of four nucleotide bases in 1953 (Watson and Crick 1953). This led to the central dogma of molecular biology. In most cases, genomic DNA defined the species and individuals, which makes the DNA sequence fundamental to the research on the structures and functions of cells. Sequencing of genomes then is now an essential task to complete, yielding essential data biologists need to understand biology and evolution of organisms. The automated Sanger method was considered a first-generation sequencing technology (Sanger and Coulson 1975; Sanger, Nicklen, and Coulson 1977), and since then newer methods have been

developed making sequencing cheaper and increasingly high throughput, these are referred to as next-generation sequencing (NGS) technologies (Goodwin, McPherson, and McCombie 2016).

With the development of NGS technology, algorithms and tools for bioinformatics and evolutionary study have developed rapidly. Here, I present a brief overview of the principles of several key bioinformatics tasks that population genetic studies with NGS data require. The processes below assume quality control of NGS reads is completed.

1.2.1 Sequence Alignment

An alignment of two sequences aims to discover or highlight how similar the two sequences are. The concept of alignments is a natural one in settings where one sequence, changes over time into a second sequence, through a series of simple operations (called edit operations) like insertions of characters, deletions of characters, and a substitution of one character for another (Mäkinen et al. 2015). It is unsurprising therefore, that alignments are a common first step in many evolutionary analyses. An alignment of the characters in two sequences, which have stayed the same over time, could be defined as the list of pairs of positions (i, j) such that the ith position in the first sequence is considered a match to the jth positions in the second sequence (Mäkinen et al. 2015).

In a practical setting, the two sequences (A & B) are typically short homologous regions of the genomes of two different individuals, or species/taxa, and are considered to have evolved through a series of changes (edit operations), from some unobserved common ancestor (Lemey, Salemi, and Vamdamme 2009). DNA sequence alignment algorithms typically require a scoring matrix which which they score potential alignments. These matrices typically define scores for aligning any two characters in two sequences,

and have some basis in biologically reality. For example, the BLOSUM scoring matrix was derived from data of conserved regions of protein families (Lemey, Salemi, and Vamdamme 2009). The score of any given pairwise alignment is the sum of the scores that were assigned by the scoring matrix for each position of the alignment.

A local alignment algorithm attempts to find the best alignments for subsequences of a query sequence with a reference sequence (Lemey, Salemi, and Vamdamme 2009). Whereas global alignment algorithms attempt to find the best end to end alignment between a query sequence and a reference sequence. Traditionally, pairwise sequence alignments were computed using dynamic programming algorithms such as the Needleman-Wunsch (global sequence alignment) (Needleman and Wunsch 1970), and the Smith-Waterman (local sequence alignment) algorithms (Smith and Waterman 1981), but efficient and accurate techniques for sequence alignment is an active area of research, and so many advances, and different techniques and software packages have been developed. Multiple alignment is the generalisation of pairwise sequence alignment to more than two sequences, this is a hard problem which becomes computationally unfeasible for many sequences without use of heuristics, such as the progressive alignment method, which first constructs a guide tree (Löytynoja and Goldman 2005).

Sequence alignments can be used to align multiple gene or protein sequences together, align reads from high throughput sequencing platforms to a reference genome assembly (Li and Durbin 2009), or to align different genome assemblies together (Paten, Earl, and Nguyen 2011). In all cases these alignments may be used to run variant calling algorithms to infer the presence of mutations and structural alterations that are present in the genomes of different taxa, individuals, or populations, and can be used to genotype individuals, and compute population genetics and evolutionary

analyses.

1.2.2 Variant Calling

Variant calling yields genotype data which may then be used in population genetics study. Identification of SNPs (sometimes called Single Nucleotide Polymorphisms or simply mutations) can be done with a read pileup output after aligning reads to a reference genome (Li et al. 2009; Li 2011). If a position j in the reference genome is covered by n reads, and of those reads, p per cent of them indicate that position j is an A, and the rest indicate that position j is a G, then it is possible to reason whether this is because the sample that was sequenced is polymorphic, or because of an alignment error or sequencing error (Mäkinen et al. 2015). Such errors are easy to identify, as they are independent events, and as such exist in a very low frequency, because the probability of observing many errors in the same location decreases exponentially. Therefore, so long as the sequencing is done to a sufficient depth of coverage, one can identify the polymorphic positions in a genome and rule out the errors with reasonable accuracy (Mäkinen et al. 2015).

Larger variants can also be detected from the read pileup. If there is a deletion in the genome of the sample from which the reads were sequenced, then if it is larger than the error threshold in the alignment, then there should be regions of the read pileup where the reference is uncovered by reads (Mäkinen et al. 2015). The region should have the same length as the deletion. If there is an insertion in the genome of the sample from which the reads were sequenced, then if it is longer than the error threshold of the alignment, then in the pileup there would be a series of consecutive positions (j, $j+1$) for which no read covers both j and $j+1$ (Mäkinen et al. 2015). This is a simplistic approach to indel detection because in reality software implementations and algorithms also take into account errors,

noise, base call qualities, and can have additional complexities such as utilizing data from many samples, and from linked sites. (Li 2011; Nielsen et al. 2011; Mielczarek and Szyda 2016).

Another approach to indel detection is to take advantage of sequencing technology platforms, which produce paired-end, or mate pair reads. Sequencers can produce pairs of reads for each DNA molecule, one begins from one end of the molecule, and the other begins from the other end, and both extend towards the middle of the molecule. When paired-end read pairs are aligned to a reference genome, they have an expected distance k between them, this expected distance is known in advance according to the protocol used to prepare the DNA library for sequencing (Mäkinen et al. 2015). Its possible to compute the actual distance for each paired-end read pair, and then compute the mean and variance of those distances. Once the mean distance k' and variance is known, each paired-end read pair can be tested to see if its distance is significantly different to the average distance. If the distance is significantly different an indel is inferred between those reads with length of $k - k'$ (Mäkinen et al. 2015).

1.2.3 Haplotype phasing

Genotypes are the unordered combination of alleles at each site of an organisms genome. The haplotype are the sequences of alleles that have been inherited together from one parent. For example, diploids possess two copies of each chromosome, therefore, in addition to being interested in which variants they possess (the genotype), one is also interested to know to which of a diploids two haplotypes each variant belongs is the variant in the organisms maternal copy of a DNA molecule, or is it in the paternal copy? The process of identifying all the variants which are situated along the same haplotype of an organism is called haplotype phasing. In an individual, variants which are clearly homozygous may be assigned to

both haplotypes very simply as both haplotypes must possess them.

Given that when there are N heterozygous sites in a sequenced DNA molecule, there are a total of $2N - 1$ possible haplotypes, that could result in those haplotypes (Mäkinen et al. 2015). Haplotype phasing was known to be a hard problem even before the development of high throughput sequencing technology. However, advances have been made and several software packages now exists to perform this task. The most accurate and widely used methods employ Hidden Markov Models to infer haplotypes Mäkinen et al. 2015. For some time, a software implementation called PHASE was considered the superior method. PHASE took ideas from coalescent theory about the joint distribution of haplotypes (Marchini et al. 2007; Marchini and Howie 2010; Howie, Marchini, and Stephens 2011). PHASE was limited by its speed however and since the development of PHASE other methods implemented in packages like IMPUTE2 and SHAPEIT1 & 2 have made improvements to the efficiency and accuracy of haplotype inference algorithms (Stephens and Donnelly 2003; Delaneau, Marchini, and Zagury 2012; Delaneau et al. 2013; Delaneau, Zagury, and Marchini 2013; O'Connell et al. 2014).

The flow of aligning high throughput sequencing reads to a reference, running variant calling and possibly haplotype inference, followed by downstream population genetic analysis on the genotype or haplotype data, is now a standard work-flow. The choice of which software packages and algorithms should be used for each task can be a subjective decision which should aim to follow best-practice for each case in question. For example, the best algorithm to use on human data, may not be the best one to use on an organism like wheat which has a radically different genome.

CHAPTER 2

HybridCheck

This chapter is based on the published scientific paper:

Ward, B. J., & van Oosterhout, C. (2016). Hybridcheck: Software for the rapid detection, visualization and dating of recombinant regions in genome sequence data. Molecular Ecology Resources, 16(2), 534-539.

The project and items of work were initially set out by my supervisor, but the work I present in this chapter is entirely my own work. I drafted the pseudo-code for the project, improved the dating algorithm presented in the chapter from it's original inefficient design, wrote all the software code, documented the package, and conducted all simulations used to test the software package, and created a website, github repository, and a web-app which provides an interface for the package.

2.1 Introduction

Recombination is one of the five evolutionary forces and is important for the formation of novel genotypes, haplotypes and alleles, thereby playing a key role in adaptive evolution (Grauer and Li 2000). Recombination is also crucial for separating deleterious mutations from their genomic background, and in combination with purifying selection it helps to curtail the mutational load (Lynch and Gabriel 1990). Recombination plays a fundamental role in the repair of damaged DNA, when homologous recombination replaces a damaged DNA strand with its intact counterpart. In all likelihood, it was this function of recombination that was important in early prokaryotic life and evolution (Cavalier-Smith 2002). With respect to adaptive evolution, however, the principal consequence of recombination is that it generates novel combinations of nucleotides, which in turns allows for selection to act a much finer scale, i.e. at the level of nucleotides rather than the entire genome. Given its fundamental importance in the biology, various mechanisms have evolved that facilitate recombination; with some depending on sexual reproduction whereas others also occur in asexually reproducing taxa. As evolutionary biologists/molecular ecologists studying gene and genome sequences, it is important to understand how the various mechanisms can result in recombination.

Homologous recombination is a process that occurs in both eukaryotes and prokaryotes, and it is an essential process through which single strand and double strand breaks, as well as base mismatches in DNA molecules are repaired. With homologous recombination, there is an equal exchange of homologous DNA sequences between the two chromatids (Lemey, Salemi, and Vamdamme 2009). In eukaryotes, this can occur through Double Strand Break Repair (DSBR) and Synthesis Dependent Strand Annealing (SDSA) (McMahill, Sham, and Bishop 2007; Sung and

Klein 2006). In prokaryotes, the RecBCD pathway, and the RecF pathways are the primary mechanisms (Madigan et al. 2012; Smith 2012). Although these pathways differ mechanistically, they all result in the invasion of donor DNA into a recipient DNA molecule through the formation of Holliday junctions, branch migration, ligation, and the repair of the DNA strands (Alberts, Johnson, and Lewis 2002).

The precise outcome of recombination and its effect on the donor and recipient DNA molecule depends on how the Holliday junctions are cut and resolved (Mimitou and Symington 2009). Crossing-over or reciprocal homologous recombination occurs when there is an equal exchange of sequence variation between the two homologous chromosomes (Grauer and Li 2000). Gene conversion is a type of non-reciprocal homologous recombination in which there is an unequal exchange of one sequence (the donor) to another (the recipient), such that the donor sequence replaces the recipient DNA (Grauer and Li 2000). Whereas crossing-over does not affect nucleotide variation, gene conversion tends to reduce nucleotide variation by making the donor and recipient sequence identical to one another. However, even though gene conversion tends to homogenise nucleotide variation, this process too can increase haplotype and genotype variation in the population, just like crossing-over (Spurgin et al. 2011). Both reciprocal and non-reciprocal recombination can occur between non-homologous sequences (Lemey, Salemi, and Vamdamme 2009). In addition, recombination can occur when distinct species or biotypes hybridise, in which case it is referred to as genetic introgression (McMullan et al. 2015). Genetic exchange between even more distantly related taxa can result in horizontal gene transfer (Eisen 2000; Ochman, Lawrence, and Groisman 2000). This too is considered a form of recombination, which occurs after gene flow between distinct taxa, and bacterial geneticists most commonly use the term 'horizontal gene transfer'.

Recombination can complicate evolutionary genetic, phylogenetic and phylogenomic analyses because neighbouring nucleotides within a single genome can differ markedly in their ancestry and coalescence. In the absence of recombination, the ancestry of a multiple alignment of homologous sequences can be represented by a single gene phylogeny. However, after a single recombination event, the sequences could have a different phylogenetic history and thus different phylogenies either side of the breakpoint (Lemey et al. 2009). Each recombinant region between two breakpoints could have a distinct ancestry and be represented by a different phylogeny. With high recombination rates, the history of a set of sequences becomes increasingly complex as different portions of the genome are shuffled, resulting in overlapping regions with distinct coalescence (Jouet, McMullan, and Oosterhout 2015). If recombination occurs in a single panmictic population, however, there will be relatively little variation in the ancestry of recombinant regions because all sequences coalesce relatively recently. On the other hand, recombination in structured populations (e.g. between distinct biotypes, strains or races) may result in the genetic introgression of diverged donor sequences, and this can lead to a mosaic-like genome structure (McMullan et al. 2015). In such cases, it is inappropriate to force a single phylogenetic tree onto a mosaic-like sequence, and it has been shown that this can significantly bias estimates of coalescent times (Jouet, McMullan, and Oosterhout 2015). Not only phylogenetic analyses are hindered by recombination, but also population genetic statistics can become biased if recombination is not accounted for, for example resulting in an upwards biased estimate of theta (and hence the effective population size) (McVean, Awadalla, and Fearnhead 2002; Watterson 1975), and the erroneous identification of positive selection (Shriner et al. 2003).

Given that recombination can potentially affect population genetic, evolutionary genetic and phylogenetic analyses, it is important to examine

whether recombination has left a signature in the sequence data. There are probably three questions one might address when analysing recombination in genome sequence data:

1. Is there evidence of recombination?
2. Where are the breakpoints / regions of recombination located in the sequence?
3. What is the rate of recombination scaled relative to the mutation rate or theta?

To detect the evidence for recombination, graphical exploratory tools can be used such as Splitstree (Huson and Bryant 2006), which visualises the impact of recombination on the phylogenetic relationship between alleles or sequences. However, to formally test the evidence of recombination, statistical tests need to be used, and many algorithms have been developed for this purpose (Lemey, Salemi, and Vamdamme 2009; Lemey et al. 2009). The general rationale of these tests is that recombination can insert novel nucleotides into a sequence alignment, making it appear that these polymorphisms have arisen there by mutation. A single nucleotide polymorphic (SNP) that is shared between two sequences, but which is not shared with their common ancestor is called a homoplasy. Such homoplasies are explained either by recombination or convergent evolution (Maynard Smith and Smith 1998). Statistical methods for detecting recombination are based on detecting phylogenetic incompatibilities that result from homoplasies (Bruen, Philippe, and Bryant 2006; Posada, Crandall, and Holmes 2002), or by finding clusters of identical substitutions in sequences (Posada, Crandall, and Holmes 2002). Measures that are computed by such methods, such as for example the homoplasy test (Maynard Smith and Smith 1998), the informative sites test (Worobey 2001), the refined

incompatibility score (Bruen, Philippe, and Bryant 2006), and the ABBA BABA test (Martin, Davey, and Jiggins 2014; Green et al. 2010) can be used to evaluate whether recombination has taken place. For example, ABBA BABA tests classify homoplasious SNPs as having one of two possible parsimonious ancestries, and they calculate the Pattersons D statistic that is based on the ratio of both types of ancestries. In case there is a significant excess of one type of ancestry over the other, this is considered evidence of recombination.

Once it has been established that recombination is affecting a nucleotide sequence, one can employ methods to identify where in the genome recombination has taken place. Those methods generally implement a scanning sliding window, and they calculate for each window the distribution of nucleotide substitutions or the genetic distance, or they assess the phylogenetic relationships between sequences at the window (Lemey et al. 2009; Posada, Crandall, and Holmes 2002). The former two methods typically attempt to find inversions or sudden changes in substitution pattern or distance values across the windows, and they do not rely on a phylogeny. Phylogenetic methods, on the other hand, infer recombination by detecting changes in the topologies, i.e. the shape of the tree. If adjacent sections of DNA sequence are phylogenetically incongruent, this is evidence for a recombination event or breakpoint (Lemey, Salemi, and Vamdamme 2009). Methods that rely on sliding windows tend to be hampered by an increased false positive rate (Type I error rate) due to multiple testing (Lemey, Salemi, and Vamdamme 2009). Bayesian approaches (Paraskevis et al. 2005) have been developed to avoid such sequential testing problems, and in addition, they can identify breakpoint positions and the parental (donor) sequences (Suchard et al. 2002).

One may also want to quantify the rate of recombination, either as a relative rate compared to the mutation rate, or as a measure of the number

of bases or recombinant regions in a DNA sequence. Measures that assess the evidence of recombination like the homoplasy test or the refined incompatibility score (Bruen, Philippe, and Bryant 2006; Maynard Smith and Smith 1998) can also be used to estimate the number of recombination events. For example, the refined incompatibility score for two sites in a sample can be interpreted as either the minimum number of convergent mutations, or the minimum number of recombination events that have occurred between a given pair of sequences (Bruen, Philippe, and Bryant 2006). The homoplasy test written by Maynard Smith and Smith 1998 calculates whether there is a statistically significant excess of homoplasies derived from the dataset, compared to the number of homoplasies that would be expected by mutation, without the occurrence of any recombination. Essentially then, simple measures and calculations of recombination rate estimation are based on trying to count the number of recombination events that have occurred during the evolutionary history of the collected sample (Stumpf and McVean 2003).

However, given that these measures do not take into account the time to the most recent common ancestor of the sample, they simply count the number of recombination events rather than estimating the recombination rate (Posada, Crandall, and Holmes 2002). In addition, recombination events do not necessarily leave a detectable trace in the DNA sequences (Lemey, Salemi, and Vamdamme 2009). To overcome this limitation, recombination can be modeled explicitly using coalescent approaches (Stumpf and McVean 2003). Using the coalescent as a framework, it is possible to estimate the population recombination rate ($\rho = 4Ner$) in software such as LAMARK (Hudson and Kaplan 1988; Hudson 2001; Kuhner 2006). This value is comparable to the population mutation parameter theta ($\Theta = 4Ne\mu$). Calculating ρ and Θ allows one to calculate the effect of recombination on nucleotide polymorphisms relative that of mutation (ρ/Θ).

Having identified a recombination region or block between a recombinant sequence and its parental (donor) sequence, it is possible to estimate when recombination did occur. This is can be done by calculating a divergence time estimate of the block in the recombinant and parental (donor) sequence. The simplest estimates of divergence time assume a molecular clock (Li 2008; Metzgar, Scripps, and Jolla 2007), i.e. a mutation rate that is constant through time and across lineages. The nucleotide divergence between the two sequences is equivalent to $2\mu t$, in which μ is the base mutation rate and t the number of generations that have elapsed since divergence. Sequence evolution may deviate from a molecular clock, and hence, methods have been developed that can take into account variation in mutation rates between taxa, genes and evolutionary time (Brown and Yang 2011; Drummond et al. 2012; Drummond and Suchard 2010; Thorne, Kishino, and Painter 1998). The popular software BEAST allows dating estimates to be made using their Bayesian estimation framework using both strict and relaxed molecular clock models (Bouckaert et al. 2014).

The HybridCheck project was created with the aim to help researchers understand the effects of recombination on genome sequence data. The software was written as a package for the R language, and it allows users to do the following.

1. Evaluate the evidence of recombination in sequences.
2. Identify recombination breakpoints and blocks.
3. Estimate the age of recombinant blocks.
4. Generate graphs to visualise the effects of recombination on the pattern of nucleotide similarity between sets of three sequences.

The development of the package involved the following three stages:

1. The R package was written to implement the functionality:

(a) Conduct ABBA-BABA tests of introgression and calculate Patterson's D, and Fd for four taxa or populations.

(b) Scan alignments of 3 sequences for putative regions of recombination and generate plots of recombination signal from these Triplet Scans.

(c) Automatically return putative regions of recombination from Triplet Scan data.

(d) Calculate the probability that the high level of sequence similarity between two putative recombination regions is consistent with the mutation rate and sequence dissimilarity observed elsewhere in the sequence.

(e) Estimate the 95% confidence interval for the coalescence time of a recombination region between two sequences (the donor and recipient). The algorithm assumes a molecular clock, and uses the binomial cumulative frequency distribution function.

(f) Draw figures to visualise the (mosaic-like) genome structure and level of nucleotide (dis)similarity between sets of three sequences.

2. A user-friendly interface was developed by creating a web-app front-end for the R package. This used a framework called Shiny. This enables users that are unfamiliar with R to use the package as a web-app with a graphical interface, as well as an R code package.

3. The performance of HybridCheck was evaluated using simulated data, and the package was assessed for the following criteria.

(a) **False positive rate**: The detection of recombination regions in simulations without recombination.

(b) **False negative rate**: A failure to detect recombination regions or portions of recombination regions in simulated sequence data with known recombination regions.

(c) **Accuracy of block age estimates**: The accuracy of the estimated coalescence time of detected recombinant blocks.

2.2 Implementation

2.2.1 Four Taxon Tests

A Four Taxon Test (FTT) is implemented in HybridCheck to allow the user to answer the question: *Is there evidence of recombination in my sequences?* FTTs use two SNP patterns called ABBA and BABA to identify introgression and require four sequences or populations, denoted as P1, P2, P3, and P4. In addition, FTTs assume a phylogeny where P1 and P2 coalesce first to form a taxonomic unit, which then coalesces with P3, and finally P4/A is the out-group with the longest branch. The ABBA SNP pattern is expected to be in abundance when introgression has occurred between P2 and P3 and the two populations share the derived allele i.e. the allele that is not ancestral (the A in ABBA and BABA). Conversely, the BABA SNP pattern is expected to be in abundance when introgression has occurred between P1 and P3. Statistics computed for a FTT quantify the abundance of these two SNP patterns. The FTT implemented in HybridCheck calculates two statistics; Pattersons D, and F (Durand et al. 2011).

Pattersons D in equation 2.1 tests for an excess of ABBA or BABA SNPs between four populations:

$$D(P1, P2, P3, A) = \frac{\sum_{i=1}^{n} C_{ABBA}(i) - C_{BABA}(i)}{\sum_{i=1}^{n} C_{ABBA}(i) + C_{BABA}(i)} \tag{2.1}$$

$C_{ABBA}(i)$ and $C_{BABA}(i)$ are defined as a binary count of whether the

ABBA or BABA pattern is observed or not at site i if four sequences are used. Alternatively if population samples are used $C_{ABBA}(i)$ and $C_{BABA}(i)$ are more generally defined using equations 2.2 and 2.3.

$$C_{ABBA}(i) = (1 - \hat{p}_{i1})\hat{p}_{i2}\hat{p}_{i3}(1 - \hat{p}_{i4}) \tag{2.2}$$

$$C_{BABA}(i) = \hat{p}_{i1}(1 - \hat{p}_{i2})\hat{p}_{i3}(1 - \hat{p}_{i4}) \tag{2.3}$$

Where $\hat{p}_{ij}$ is the frequency of the derived allele at site i in population j. Pattersons D is expected to be 0 where no introgression has occurred between the populations (Durand et al. 2011).

The $\hat{F}_d$ statistic is defined as the fraction of the genome shared through introgression (Martin, Davey, and Jiggins 2014). The equation uses the same numerator as that of the formula for Pattersons D, which is given the name S and denotes the difference between the number of ABBA sites and BABA sites, as per equation 2.4.

$$S = \sum_{i=1}^{n} C_{ABBA}(i) - C_{BABA}(i) \tag{2.4}$$

The formula for the $\hat{F}_d$ statistic compares this observed value of S, denoted as $S(P_1, P_2, P_3, P_4)$, with a value of S estimated under a scenario of introgression (Martin, Davey, and Jiggins 2014). Specifically HybridCheck considers two scenarios and computes $\hat{F}_d$ for both: Complete introgression between populations 2 and 3 and complete introgression between populations 1 and 3. These two scenarios are denoted as $S(P_1, P_D, P_D, P_4)$ and $S(P_D, P_2, P_D, P_4)$ respectively. In both scenarios, P_D is the donor population and is chosen by finding which of the introgressed populations has a higher frequency of the derived allele (Martin, Davey, and Jiggins 2014). Therefore, the two formulas for $\hat{F}_d$ that are used by HybridCheck are given

as equations 2.5 and 2.6.

$$\hat{f}_d = \frac{S(P_1, P_2, P_3, P_4)}{S(P_1, P_D, P_D, P_4)} \tag{2.5}$$

$$\hat{f}_d = \frac{S(P_1, P_2, P_3, P_4)}{S(P_D, P_2, P_D, P_4)} \tag{2.6}$$

When calculating the FTTs, HybridCheck will break up the sequence alignment into a user definable number of blocks of a given length, and will compute for each block:

1. Pattersons D.
2. The two $\hat{f}_d$ statistics (one for each of the two scenarios of introgression).
3. A Pvalue based on the binomial distribution.
4. The number of sites that have a higher ABBA score.
5. The number of sites that have a higher BABA score.

These blocks are then used perform a jackknife to compute jackknife estimates, standard deviation, and Z scores for the four populations of the whole alignment. The binomial P-values computed for each block used with Fishers combined probability formula to calculate an overall binomial based P-value for the entire alignment.

HybridCheck can be directed by the user to use certain populations in the place of P1, P2, P3, and P4. Alternatively it can automatically generate combinations of four populations and then decide which of the populations should be assigned which of the four positions, using the distances between the sequences. The statistics calculated in the four-taxon tests have been described and their performance evaluated in previous work by Martin, Davey, and Jiggins 2014. HybridCheck can be directed by the user to use

certain populations in the place of P1, P2, P3, and P4. Alternatively it can automatically generate combinations of four populations and then decide which of the populations should be assigned which of the four positions, using the distances between the sequences.

2.2.2 Sequence triplet scans for recombination signal

A sliding window scan of pairwise sequence similarity for three sequences (hereafter referred to as a triplet) was implemented in HybridCheck to allow the user to answer the question: *Where are the breakpoints / regions of recombination located in the sequences?* HybridCheck was designed to generate and scan every possible triplet for a multiple sequence alignment. In addition, HybridCheck can be set to ignore triplets that include two or more sequences that are highly similar, reducing the number of scans to be performed. HybridCheck can also analyse a user-defined subgroup of sequences, or use the results of the four-taxon tests to generate the sets of triplets that need to be analysed. All non-polymorphic sites are removed from each triplet prior to the sequence scans.

Potential recombinant regions are identified from the sliding window similarity scan data based on significantly elevated levels of sequence similarity. The cut-off point to identify elevated similarities is found by calculating the kernel density distribution of all raw sequence similarity data and identify peaks that fall outside this distribution. The start and end points of peaks are recorded (in base pairs) as well as the number of mutations within the block.

The exact probability that the nucleotide similarity within a block is significantly higher than the overall sequence average can be calculated by modeling the accumulation of mutations as a Bernoulli trial. The probability of observing k or fewer mutations in a nucleotide sequence alignment of two sequences of length n is given by equation 2.7.

$$Pr(X \leq k) = \sum_{i=0}^{\lfloor k \rfloor} \binom{n}{i} p^i (1-p)^{n-i} \tag{2.7}$$

In this equation (2.7), p is the proportion of observed single nucleotide polymorphisms (SNPs) between the two aligned sequences (including the non-informative sites). If the probability falls below the Bonferroni corrected critical value $\alpha = 0.05$, the amount of polymorphism in the block is inconsistent with the level of polymorphism that is expected from the accumulation of mutations. In this case, recombination is taken to be a valid explanation for the number of observed substitutions.

2.2.3 Estimating the age of recombinant regions

HybridCheck can estimate the coalescence times of the introgressed blocks. This time is estimated assuming a strict molecular clock and using the observed number of SNPs in the introgressed block. In order to correct for mutation saturation, homoplasy, back mutations and transition / transversion ratios, HybridCheck converts the number of SNPs into the number of mutations using a JC (Jukes and Cantor 1969), K80 (Kimura 1980), F81 (Felsenstein 1981), HKY (Hasegawa, Kishino, and Yano 1985), or GTR (Tavare 1986) correction.

Considering the mutation accumulation process as a Bernoulli trial, and the coalescence time can be found by finding the root of the equation 2.8.

$$f(n, k, 2t, Pr(X \leq k)) = \left(\sum_{i=0}^{\lfloor k \rfloor} \binom{n}{i} 2\mu t^i (1-2\mu t)^{n-i} \right) - Pr(X \leq k) \tag{2.8}$$

In equation 2.8, μ is the mutation rate, t the time in generations, k the observed number of SNPs, and n the total number of base pairs in the block. The R function uniroot computes the value for $2\mu t$ by finding the root

(i.e., the zero value) of function 2.8. (Brent 1973). In order to calculate the median and 5-95%CI, the function is solved for $2\mu t$ when Pr is set to 0.5, 0.05 and 0.95.

2.2.4 Performance Testing

HybridCheck was tested on sequence triplets of 50kb in length which contained no introgression events to quantify its false positive rate α (i.e. erroneously identifying recombination). The simuPOP Python module (Peng and Kimmel 2005) was used to simulate three populations with 500 individuals that derived from a single panmictic ancestral population, and which continued to evolve in genetic isolation. The populations diverged for between $0.01 \leq \mu t \leq 0.1$ generations (this is equivalent e.g. to t = 1 to 10 million generations with $\mu = 10^{-8}$ base mutation rate). Sequence triplets were generated by randomly sampling one sequence from each of the three populations. A total of 100 independent sequence triplet replicates were generated for each simulated level of divergence (μt).

HybridCheck was also tested on 50kb sequence triplets which contained set known introgression events of various ages to assess the sensitivity of the software to detect hybridization and the false-negative (β) rate. These triplets were also generated by simuPOP simulations in which two parental sequences diverged for between $0.02 \leq \mu t \leq 0.08$ generations, exactly as in the false positive error simulations described above. However, unlike the false positive error simulations, two subsequent steps were simulated: The parental sequences recombined at a user-defined breakpoint at 35 kb, generating a third recombinant sequence. Then, in order to age the introgression blocks, the three sequences diverged for another $\mu t = 0{:}0.1$ generations under a JC69 model, and during this time, the signal of introgression becomes eroded by mutation.

Finally, the accuracy of the dating algorithm was tested using a regression analysis. This used the same simulated data as was generated for evaluating the type II error rate. The estimated age calculated by HybridCheck was the response variable in the regression, and regressed the known coalescence time of the recombinant blocks in the simulations, was the explanatory variable.

2.3 Results

The false positive rate is presented in Figure 2.1, plotted on the y-axis against the amount of divergence (expressed as μt) between sequences on the x-axis. Depending on the divergence time of the populations, the false positive rate decreased with increasing sequence divergence but remained consistently less than α=0.05. This means that if a triplet of sequences is analysed for recombination with HybridCheck, the more diverged they are from each other, the less likely it is that blocks will be falsely identified as putatively recombinant, when in fact no recombination has taken place. From this, one may conclude that recombination detection analyses can be confounded when populations or sequences analysed are not very diverged from each other, and that apparent recombination blocks or signals may be explained by other factors. Such facts include ancient population admixture or incomplete lineage sorting, and this will be addressed in more detail in the discussion.

The false negative rate is presented in Figure 2.2. The false negative rate is plotted on the y-axis, against the amount of time since recombination occurred (expressed as μt). The data is partitioned into series, according to the amount of divergence (expressed as μt) between parental sequences prior to hybridisation. Figure n+1 shows that HybridCheck was able to detect $>$95% of recent introgression events even if the two parental populations

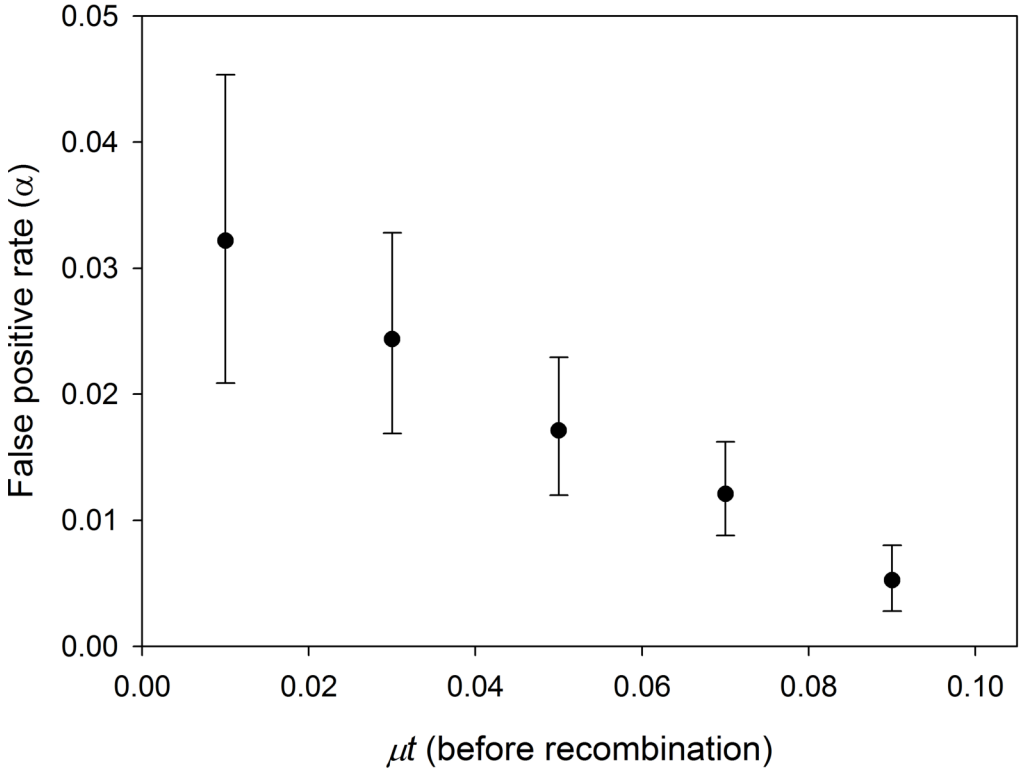


Figure 2.1: The mean(±5 - 95%CI) false positive rate (α) of HybridCheck as a function of the ancestral divergence time μt (i.e. the amount of time of the sequences diverged before recombination). As sequences become more diverged, the false positive rate decreases.

had diverged only moderately. However, more ancient introgression events were detected only if both parental populations had significantly diverged.

The accuracy of the dating estimates HybridCheck calculates for our simulated scenario is presented in Figure 2.3. This analysis shows that when the ancestral sequenced had diverged significantly ($\mu t \geq 0.2$), the age estimates calculated by HybridCheck are a good approximation of the actual time passed since recombination (Linear Regression: Estimated age = 0.000795 + 0.968 t, R2=99.3%). However, when the exchanges occurred between sequences that were only moderately diverged ($\mu t < 0.2$), the age of the recombination events are underestimated when recombination happened in the distant past ($\mu t > 0.05$) (see Figure 2.3). In such cases, mutations accumulated after the recombination event fragmented the blocks, resulting in an underestimate of the number of SNPs in the blocks that were

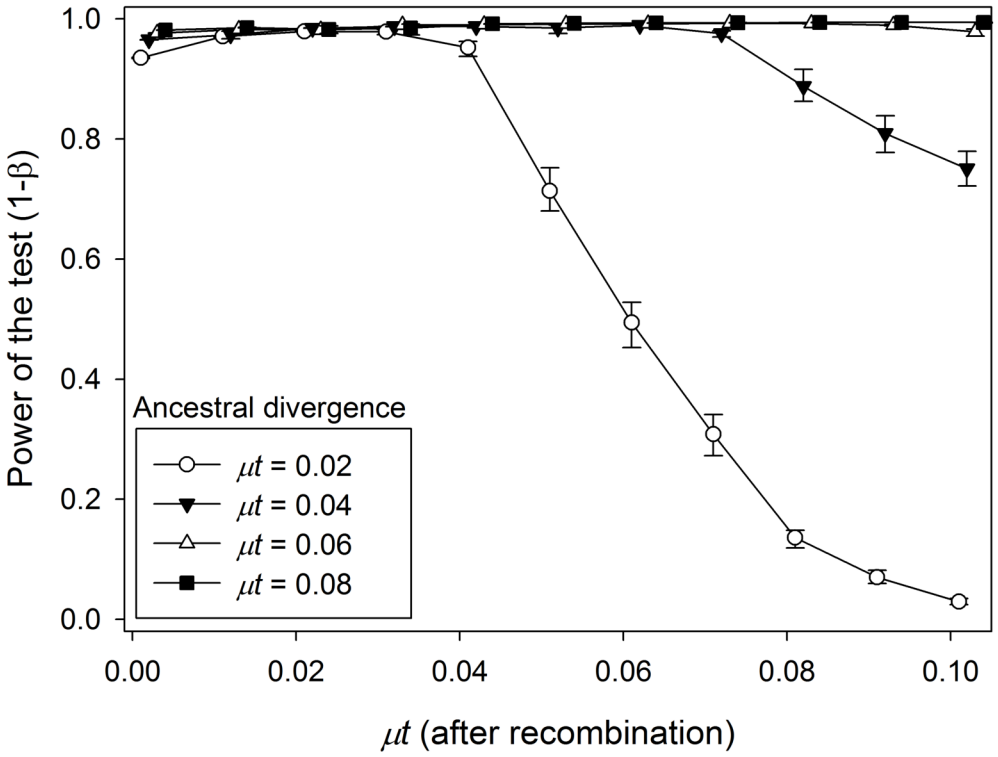


Figure 2.2: The mean($\pm$5 95%CI) statistical power (1 - β) of HybridCheck as a function of the divergence time of the sequences after recombination (expressed in μt) for sequences with ancestral divergence times μt = 0.2, 0.4, 0.6 and 0.8 generations. Recombination between moderately diverged sequences can be detected in $>$95% of the cases, as long as the recombination event was relatively recent.

detected.

2.4 Discussion

In this project, the objectives were to create and test a software package for the exploratory analysis of large sequences for evidence of introgression and hybridization. The package is designed to take the researcher through the following questions:

1. Is there evidence of recombination / introgression?
2. Where are the recombination regions in the sequences?
3. What is the divergence time of recombinant blocks that are detected

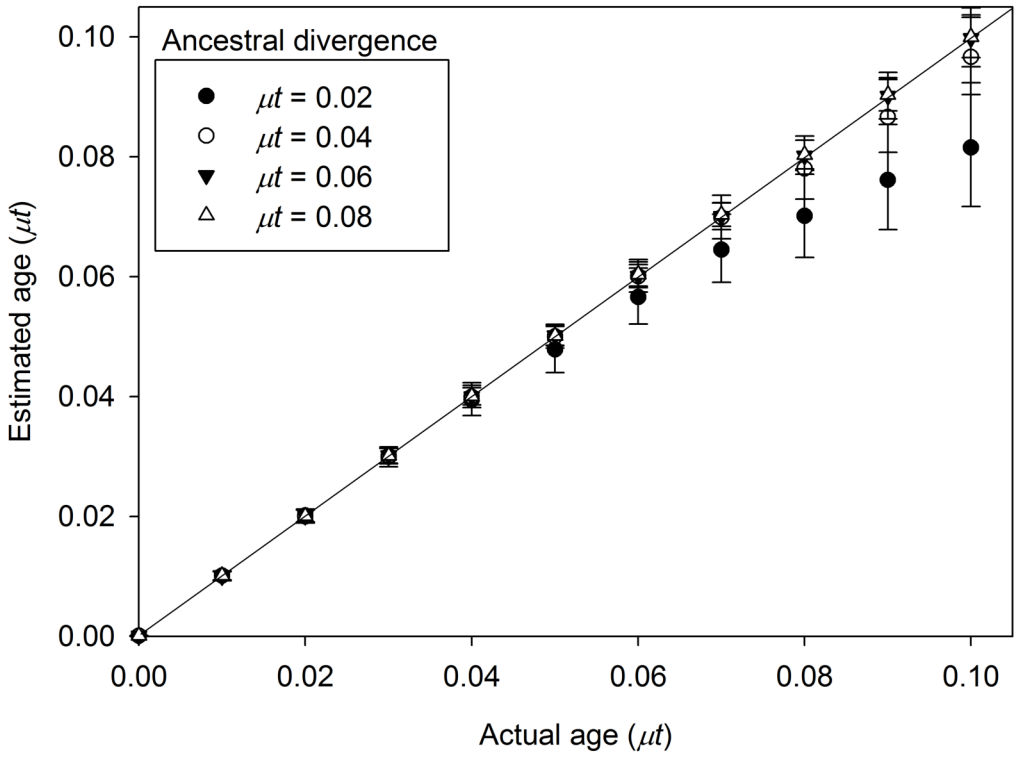


Figure 2.3: The mean ($\pm$SEM) estimated age (expressed in μt) of recombinant blocks calculated using the dating algorithm with a JC correction in HybridCheck, versus their actual age. In most of the scenarios, HybridCheck returns an unbiased estimate of the divergence time. However, the age is underestimated in cases of ancient recombination between populations that have ancestral divergence of 0.2.

by the package?

2.4.1 Performance of detecting recombinant blocks

The data demonstrate that for the simulated scenarios, HybridCheck performs best when sequences are diverged sufficiently prior to hybridization, and the hybridization or recombination event was relatively recent. However, when the parental sequences of the hybrid sequence were sufficiently diverged recombinant blocks were clearly detected long after the recombination event ($\mu t > 0.06$). In addition, when divergence between parental sequences of a hybrid sequence was high then dating estimates of the recombinant blocks remained more accurate for older recombination blocks.

If two parental sequences are significantly diverged prior to hybridisation,

the introgressed regions will be more apparent in the sequence similarity scans of HybridCheck because their high nucleotide similarity stands in sharp contrast with the genomic background. With a lower level of ancestral divergence, the increase in local sequence similarity caused by recombination is more difficult to distinguish from stochastic variation in nucleotide divergence, around a higher average level of sequence similarity. As a result, the algorithm HybridCheck employs to decide on a suitable sequence similarity threshold can be confounded as it tries to identify regions with sequence similarity that fall outside of the mean noise levels of sequence similarity. Therefore, HybridCheck would struggle to analyse a study system in which populations or taxa analysed are too closely related and have not diverged for long enough to accumulate unique polymorphisms which will be shared between parental and hybrid sequences.

Previous studies have shown that many window based recombination detection methods perform better when the divergence is above 0.05 (expressed as a proportion of the sequence length) (Posada and Crandall 2001). Furthermore, simple implementations such as MAXCHI, and site incompatibility based methods usually perform better than phylogenetic based methods because the latter only tend to detect recombination if it changes the tree topology (Posada and Crandall 2001).

HybridCheck window scans attempt to find elevated similarity between genome sequences / contigs of two taxa which are unrelated. Such elevations in similarity are indicative of, and often coincide with incongruence between differing gene tree topologies. However, such signatures can have causes other than recombination, and elevated levels of sequence similarity could also be due to stabilizing selection conserving sequences between populations. Alternatively, diverging populations of organisms could show increased levels of divergence in regions of the genome that are under adaptive selection, and if there is gene flow between populations

the background will be homogenized compared to the regions that are subject to divergent selection (Nadeau et al. 2012). Such genomic islands of divergence appear to be less evident between populations that are further along the speciation process in butterfly species (Nadeau et al. 2012). A selective sweep is a phenomenon whereby positive selection in an allele reduces variation in neighboring regions due to linkage. This is also called hitchhiking (Hedrick 1980). If a selective sweep is strong and only one haplotype exists in high numbers in the population as a result, then a large reduction in variation is possible. Selective sweeps could create regions of sequence similarity similar to those created by hybridisation events. Note however, this scenario reduces variation around a positively selected allele within in a population.

HybridCheck attempts to overcomes these effects of selection in part by removing non-polymorphic sites prior to measuring the sequence similarity across sequences, but it is still possible that selection could be responsible, and the removal of informative sites by selection therefore reduces the power of HybridCheck to reliably identify introgression in those regions. Therefore HybridCheck is not recommended or useful if a researcher is interested in smaller regions subject to very strong selection, due to the resulting lack of information. If there are protein-coding regions in a detected recombinant region and selection is thought to be responsible, then the sequences should be analysed for evidence of purifying selection and/or selective sweeps within the detected region.

Elevated sequence similarity and incongruent tree topologies can also be caused by incomplete lineage sorting or deep coalescence (Rogers and Gibbs 2014). This occurs when an ancestral species is polymorphic for a given gene before the species tree splits into two daughter species. After the first species split, if the polymorphism does not become resolved into two separate monophyletic lineages before the next speciation event,

then the species tree will not match the gene trees of individual alleles (Rogers and Gibbs 2014). This problem is likely if a population size is very large, or if the time between branching events is low (Rogers and Gibbs 2014). Much of the genome of *Homo sapiens* shows evidence of incomplete lineage sorting. As a consequence, parts of the genome supported the phylogeny (chimpanzee, (human, gorilla)), whereas other regions of the genome supported the phylogeny (human, (chimpanzee, gorilla)) (Galtier and Daubin 2008). Both these phylogenies disagree with the species phylogeny of homonids (gorilla, (human, chimpanzee)) (Galtier and Daubin 2008; Rogers and Gibbs 2014). This discordance is because selection can cause similar sequences, or islands of divergence as previously described, and then incomplete lineage sorting results in gene trees that are discordant with the species tree and other gene trees, as a result of the incomplete and stochastic resolutions of polymorphisms, before subsequent speciation events (Scally et al. 2012).

However, HybridCheck can help discern recombination from incomplete lineage sorting by comparing the coalescence time of recombinant regions with the split of the species. If the age of a recombinant region is significantly younger than the split of the ancestral species, the pattern is inconsistent with incomplete lineage sorting. In this case, genetic introgression after hybridisation is a more plausible explanation for the observed increase in local sequence similarity. HybridCheck makes this practically possible for the researcher to do, for many recombinant blocks.

To summarise the performance of the HybridCheck when identifying recombinant regions, the HybridCheck use case is intended predominantly as an exploratory method to scan for signal between sequences from diverged populations or taxa, rather than within populations. Outside of this use case, HybridCheck may be unsuitable for some systems as a result of limited divergence between sequences, and selection, both of which

result in reduced information for the HybridCheck analysis method. Recent speciation and large population sizes may result in incomplete lineage sorting, which can affect patterns of divergence and ancestry in similar ways to recombination, however coalescent times computed by HybridCheck may help distinguish incomplete lineage sorting from recombination. When using HybridCheck for a study system outside of its designed use case, whilst it is useful for highlighting the regions of the genome affected by the above factors, regions should not be uncritically considered the result of hybridisation or recombination, and the alternative causes e.g. selection should be followed up and ruled out before any such conclusion.

2.4.2 Performance of estimating the age of recombination events

From the results it is evident that the dating algorithm used in HybridCheck tends to underestimate the divergence time of recombinant blocks in old recombination events. This is because recombination blocks can become fragmented by accumulation of subsequent mutations following the recombination event. Consequently, older recombination blocks tend to be smaller, when they are actually larger. Thus, not all mutations are accounted for, resulting in an underestimate of the divergence time particularly for old recombination events.

Furthermore, the dating algorithm used in HybridCheck makes several assumptions in order to be simple and fast. As a result however, if these assumptions are broken then this will affect how representative the estimates returned by HybridCheck are of the true age of a recombination event. The algorithm assumes that the mutation rate has been constant over time and identical in all taxa. This assumption is not always true, and more sophisticated approaches, such as the Fossilized-Birth-Death process allow for

the calibration of divergence time estimates during Bayesian phylogeny estimation (Heath, Huelsenbeck, and Stadler 2014). It uses all available fossils, and considers extant species and fossils of species part of the same macro-evolutionary process (Heath, Huelsenbeck, and Stadler 2014).

In addition, the algorithm uses a nucleotide substitution rate to infer the mutation rate. In order to correct for mutation saturation, homoplasy, back mutations and transition / transversion ratios, HybridCheck converts the number of SNPs into the number of mutations using a JC (Jukes and Cantor 1969), K80 (Kimura 1980), F81 (Felsenstein 1981), HKY (Hasegawa, Kishino, and Yano 1985), or GTR (Tavare 1986) correction. However, substitution rates do not solely depend on mutation rates, and they appear to be auto-correlated across sequences due to the effect of selection. Selection can vary between sites, genes and taxa, and selection and substitution rates can change through time as conditions change (Barrick and Lenski 2013; Bromham and Penny 2003).

Furthermore, the size of populations must be taken into account (Bromham and Penny 2003). Bayesian coalescent approaches incorporated in software such as BEAST (Bouckaert et al. 2014) should be used when using a relaxed clock or more advanced method of dating. However, these methods are computationally more demanding and might become unfeasible when estimating the divergence time of a large number of recombination events. In such cases, the age estimate returned by HybridCheck offers a good approximation when recombination occurred relatively recently ($\mu t < 0.05$), and also when the ancestral sequences have diverged significantly before hybridizing.

In conclusion, the HybridCheck project is intended as a simple all-inclusive tool to analyse recombination in genome sequence data. The implemented algorithms are not as sophisticated as methods that employ Bayesian estimation of parameters and coalescent simulations. However,

this means that the package is computationally fast, which makes it a useful first port-of-call for identifying recombination and assessing whether other explanations such as incomplete lineage sorting may apply.

CHAPTER 3

The role of introgression in the adaptive evolution of the generalist plant pathogen, *Albugo candida*

This chapter is based on the published scientific paper:

McMullan, M., Gardiner, A., Bailey, K., Kemen, E., Ward, B. J., Cevik, V., ... Jones, J. D. (2015). Evidence for suppression of immunity as a driver for genomic introgressions and host range expansion in races of Albugo candida, a generalist parasite. eLife, 4, 1-24.

This thesis chapter presents a research project that was a collaboration between many researchers. In this chapter in order to provide clear description of the work involved, some details regarding some work that has not been performed by myself are presented. Specifically, work described in sections 3.2.1 and 3.2.2 were completed by collaborators and not myself. My contributions to the work are described in sections 3.2.3 and 3.2.4, and it is results of this work that is presented in this chapter.

3.1 Introduction

Host specificity is a defining feature of pathogens, and can be defined as the inverse of the number of hosts that a given pathogen can infect (Poulin and Keeney 2008). Host specificity is negatively correlated to the probability of parasite extinction, and positively correlated to the ability of a parasite to colonise and adapt to a new host (Poulin and Keeney 2008). Host specificity is constrained by the physiology of the pathogen. Therefore the host specificity of a pathogen is constrained by factors including (but not limited to) the pathogen's method of transmission, method of obtaining nutrients and energy from the host, and the ecology of the pathogen and host (Poulin 2011). Such factors are proximal constraints on host specificity, but host specificity is ultimately constrained by the evolutionary and biogeographical history of the pathogen and its potential hosts (Poulin and Keeney 2008; Poulin 2011).

A highly specialist parasite occurs in only a single host species. They often require host-host contact for transmission, and their longevity and future is strongly linked to that of their host species (Poulin and Keeney 2008). Conversely, a parasite that is more generalist may survive the extinction of one host species, since there is another host species they can exploit to survive. Generalist parasite species may rely less on contact-transmission or close proximity between hosts. For example, they may be transmitted through food, or some other species vector (Pedersen et al. 2005). However it should be noted that even if a pathogen has a very high mobility, and dispersal, its host-specificity can be high (Poulin and Keeney 2008).

The availability of ecologically and evolutionarily related or similar hosts cohabiting the same habitat, may cause differences in the host-specificity of two otherwise similar pathogen species (Jex, Schneider, and Cribb 2006).

Furthermore, parasites put selective pressure on host populations to adapt and develop immunity, increasing the frequency of genetic and epigenetic variants that improve immunity, and pathogen detection in the host. As a result, these adapting host populations impose selection pressure on the pathogen populations, increasing the frequency of variants that maintain the pathogens efficiency of immune suppression. Over time, both host and parasite co-evolve and become intimately associated, as they both adapt to each other's latest antagonistic evolutionary innovations. This is called an evolutionary arms-race (Boutemy et al. 2011; Buckling and Rainey 2002; Cooper et al. 2008; Kemen and Jones 2012; Lamour et al. 2010).

Overall, the general pattern observed in nature, is that most parasite species are largely specialised and co-evolve with only a few, if not one, host species (McMullan et al. 2015). It should be noted however, that this generalisation is based on a measure of host specificity that is based on a simple measure of host specificity, namely the number of host species that are colonised by a parasite in natural populations. However this metric makes an oversimplification that does not reflect biological reality. For example, two pathogen species may have the same number of host species, but if one of the pathogens infects species of one genus, and the other infects species of multiple genera, then it is not realistic to conclude both of the parasite species are equally specialised. It is because of this problem, that Poulin and Mouillot 2003 defined a host-specificity measure that takes into account the taxonomic or phylogenetic distances between the hosts colonised by a parasite. Later the authors published an improved metric that incorporated the phylogenetic or taxonomic distinctness of a pathogens host species, but also weighted for the prevalence of the parasite on its different host species (Poulin and Mouillot 2005). The rationale for such a weighting is that a pathogen that is largely concentrated on only one of its multiple hosts should be classified as more specialised than a pathogen

that utilises and colonizes all of its host species evenly.

The organism of interest in this work is the obligate biotrophic plant pathogen, *Albugo candida*. Plant pathogens have a parasitic relationship with their host, and are classified according to the nature of this relationship with the host. Pathogens which obtain nutrients from decaying plant matter are classified as necrotrophs, whereas pathogens which require living host tissue in order to obtain nutrients are classified as obligate biotrophs (Kemen and Jones 2012). These biotrophs don't typically secrete abundant lytic enzymes, and cause little physical or structural damage to the host plant (Kemen and Jones 2012). Pathogens with a combination of these two lifestyles are classified as hemibiotrophic (Kemen and Jones 2012; Lamour et al. 2012). *Albugo candida* is an obligate biotroph, and whilst *Albugo candida* is a generalist, infecting species of the Brassica family, obligate biotrophs are typically specialists (McMullan et al. 2015).

After an obligate biotroph makes a host-jump, it is expected that selection will increase any adaptive genetic or epigenetic variant in the population that results in more efficient immune suppression of the new host (Dong et al. 2014; Kemen and Jones 2012; Poulin and Keeney 2008; Raffaele et al. 2010; Thines 2014). Furthermore, host-parasite co-evolution over time will result in both the host and parasite constantly adapting to each others latest antagonistic adaptations, and they will become more intimately associated historically (Morgan and Kamoun 2007; Raffaele and Kamoun 2012; Thines 2014). As both of these processes occur, new effectors and pathogenicity factors may be created, and existing ones may receive beneficial mutations, and they may also have their levels of expression changed epigenetic modification and inheritance (Dong et al. 2014; Gijzen, Ishmael, and Shrestha 2014; Raffaele and Kamoun 2012; Raffaele et al. 2010; Win et al. 2012). These will be fixed due to selection pressure if they are beneficial. These modifications enable more efficient immune

suppression and exploitation of one host species, but increase the risk of detection in other host species by triggering their immune system (Martin and Kamoun 2012). Thus, as obligate biotrophic pathogen populations become more adept at suppressing the immunity of one host, they will become less adept at infecting previous host(s) or other hosts it can infect.

Therefore, obligate biotrophs are typically known for being intimately associated with their hosts i.e. they have a high host specificity (Thines 2014). Yet there are generalist biotrophic parasites that appear to have overcome this evolutionary dilemma and show virulence on diverse hosts. *Albugo candida*, the organism that is the subject of this work, is one such generalist, but there are other generalist oomycetes, like *Phytophthora capsici* (Lamour et al. 2012).

Some generalist parasite species have solved the dilemma by evolving multiple specialised races, and each specialised race can infect a different host. For example, the eukaryotic order *Albuginales*, of which *Albugo candida* is a member, is completely comprised of obligate biotrophic pathogens that cause disease on a broad range of plant hosts (Biga 1955; Choi and Priest 1955; Walker and Priest 2007).

Albugo is the largest genus of the order *Albuginales*, and it was reported to consist of 33 specialist pathogens by Biga 1955. More recently, the estimate is that the genus comprises approximately 50 pathogens, and these are typically specialists. In addition new distinct *Albugo* species have been discovered that were previously thought to be members of *Albugo candida* (Pers) Roussel. (Choi et al. 2011; Choi, Shin, and Thines 2009; Ploch et al. 2010; Thines et al. 2009). This is because in the past decades, classification was based largely on morphology, and this led to the application of a broad species concept, that resulted in *Albugo candida* (Pers.) Roussel being regarded as the causal organism of all incidents of white blister rust on all *Brassicaceae* hosts (Choi et al. 2011). As late as

2011, it has been estimated that a dozen distinct species thought to be *Albugo candida* await discovery (Lamour and Kamoun 2009).

Albugo candida (Pers.) Roussel can infect 241 species of plants in 63 genera from the families of *Brassicaceae*, *Cleomaceae* and *Capparaceae* (Choi, Shin, and Thines 2009). *Albugo candida* infections are the causal agent of white blister rust disease, resulting in significant losses on *Brassica* crops of economical importance. For example, *Albugo candida* causes up to 56 of yield losses in Indian Mustard (Meena et al. 2002). *Albugo candida* consists of different physiological races, each usually featuring high host-specificity and approximately 24 races of *Albugo candida* have been defined, based on their host range (Saharan et al. 2014; Saharan and Verma 1992).

Albugo candida reproduces both asexually and sexually (Holub et al. 1995). During asexual reproduction, diploid zoospores are formed in zoosporangium beneath the leaf epidermis. The zoosporangium are visible when dehydrated and in large numbers, as white blisters (Holub et al. 1995). These sporangia then rupture the epidermis of the host leaf, to release zoospores for dispersal. During sexual reproduction, fertilization between two isolates creates non-motile, diploid, and thick-walled oospores (Holub et al. 1995). The oospores can resist extreme temperatures and desiccation. The relative importance of both reproductive modes is not well established, but the clonal (asexual) mode of reproduction allows rapid population expansion, especially given modern crop mono-culture growing practices. Although *Albugo candida* comprises distinct, specialised physiological races that colonize different host plants, and that distinct species have been identified that were initially thought to be *Albugo candida* (Choi et al. 2011), it is still considered a single species.

According to evolutionary and population genetic theory, the trade-offs associated with adaptation and host-specialisation, coupled with strong

population structuring, can result in adaptive radiation and speciation (Abbott et al. 2013; Stukenbrock 2013). *Albugo candida* then may be thought of as a currently ongoing adaptive radiation; The broad host range of *Albugo candida* is enabled by an ongoing specialisation of independent physiological races, and these races are likely heading for speciation (Dres and Mallet 2002). If strains or races of a parasite develop adaptations to specific hosts, and make trade-offs in doing so, specialising to the given host, does parasite specialization inevitably lead to speciation? Certainly, specialising on one or a few hosts, at the cost of being able to infect other hosts, will mean separation of specialised races, ecologically, and even geographically, over time such separation is expected to result in reproductive isolation.

Compared to other microbial plant pathogens, *Albugo* species are notable as infections strongly suppress host innate immunity. As a result, infections of *Albugo* species increase the susceptibility of the host to a secondary infection by pathogens that would otherwise be avirulent, including downy mildews (Cooper et al. 2008). It has been suggested that this immune suppression caused by *Albugo* infections might allow an accelerated adaptation of other pathogen species to host that is susceptible to *Albugo* species (Thines 2014).

However, whilst it has been suggested the immune suppression will accelerate the adaptation of other pathogens to the suppressed host, before this project, no evolutionary rationale was proposed explaining why rendering a host susceptible to other pathogens could be adaptive for the various Albugo species. Hypothetically, a pathogen which colonizes and adapts to the hosts of Albugo species due to the immune suppression of Albugo species infections, will become competition against Albugo species for the same resources (Cooper et al. 2008).

Suppression of host innate immunity would facilitate cohabitation of distinct physiological races that otherwise would not come into contact

due to their specialisation and adaptive trade-offs, as previously discussed. When the distinct physiological races come into contact, genetic exchange including introgression and hybridisation may occur between them. Here, introgression is defined as the introduction of nucleotide variation from a parental donor race into the genome of a recipient race, through the mechanism of recombination (Hedrick 2013). This flow of genetic variation from one donor physiological race, to a recipient physiological race, could slow down the genetic divergence of the races, and slow or prevent speciation. However, introgression between races that are specialised and adapted to exploit different hosts could be maladaptive, and therefore could be strongly selected against. This is because hybrids would inherit effector alleles derived from both parental races. Therefore, whilst the hybrid genomes would contain effectors that enable the immune suppression of multiple hosts, they could also contain effectors that trigger immunity on multiple hosts. Immune recognition of even a single effector is sufficient to trigger the immune response and stop an infection. Therefore any hybrid that possess an expanded repertoire of effector alleles are likely to have a strong fitness disadvantage on most potential host plants, as with larger effector repertoire's comes an increased likelihood of one of them triggering host immunity.

This chapter presents work conducted and contributed to a larger genome project analysis of *Albugo candida*, conducted by a team of scientists at the University of East Anglia, and The Sainsbury Laboratory. This project aimed to answer the following questions, in order to try and resolve this question of whether immune suppression and secondary infection is adaptive or maladaptive, and whether it is due to hybridisation:

1. Are the distinct physiological *Albugo candida* races genetically isolated and on the road to speciation?

2. Does suppression of host innate immunity enable cohabitation and growth of races with non-overlapping host ranges?

3. Are the genomes of *Albugo candida* affected by recombination and hybridisation?

The work presented in this chapter was primarily conducted with a goal of answering the third question of the project. During the collaborative project, genome sequence assemblies were created for five isolates that were collected from four host species (*Brassica oleracea*, *Brassica juncea*, *Capsella bursa-pastoris*, and *Arabidopsis thaliana*). This chapter presents analyses performed on the assembled sequence scaffolds for the detection of recombination, hybrididdation, and mosaic genome structure.

3.2 Methods

In order to perform the analysis of genome structure that is the focus of this chapter, prior work was conducted to isolate the *Albugo candida* races used in this study, test for virulence, extract and sequence DNA, and RNA, and perform genome and transcriptome assemblies. These procedures are subsequently described in detail in (McMullan et al. 2015), and given that these procedures are not the focus of this chapter, the reader is referred to this paper for details on the wet lab and molecular methods. A brief summary of these methods is described below.

3.2.1 Isolation and cultivation of races used in the study

In order to address the research questions presented in the previous section, genome sequence assemblies were required of five isolates of *Albugo candida*, the white rust fungus. These isolates were collected from four different host species: *Brassica oleracea*, *B. juncea*, *Capsella bursa-pastoris*,

and *Arabidopsis thaliana*. The isolates were collected by Erik Kemen, prior to the evolutionary analyses that are the focus of the present chapter.

The isolate designated AcNc2, was isolated from infected leaves of *Arabidopsis thaliana* Eri-1 field-grown plants in Norwich, England. The isolate was collected in 2007. The isolate AcEm2 was collected from wild *Capsella bursa-pastoris* in Kent, England in 1993. AcBoT was collected from infected cultivars of *Brassica oleracea* called 'Bordeaux F1', from Lincolnshire, England, in May 2009. AcBoL was harvested from infected *Brassica oleracea* leaves from Lincolnshire, but in the January of 2009. An isolate which is virulent on *Brassica juncea* called Ac2V was provided by M Borhan of Agriculture and Agri-Food, Canada. All of these isolates were single spore purified (Kemen et al. 2011).

3.2.2 Genome assemblies of isolates

The assembly of isolate AcNc2 was used as a reference. The assembly is 34Mb in size, and has 5212 contigs of approximately 160-fold coverage. The assembly was approximately 73% of an estimated genome size of 45Mb. The unassembled part of the genome (approximately 11Mb) is likely to contain repeats, approximately 8% of which represent collapsed regions, since they have coverage that is several times higher than the average. For each isolate, several assemblies were constructed with different k-mer lengths. Each assembly was assessed according to number of contigs, N50 (Bp and number), mean contig length, assembly size, GC content, average genome coverage, repeat content, and the number of predicted genes. High sequence similarity of the five *Albugo candida* isolates resulted in the conclusion that three races had been sequenced: AcNc2, and AcEm2 were isolates of the same race, and AcBoT and AcBoL were also two isolates which belonged to the same race. Therefore, detection of recombination and hybridisation in this chapter were first conducted on the three races

AcNc2, Ac2V, and AcBoT, each of which had a 33-34Mb assembly.

3.2.3 Detection of recombination events

Recombination events were statistically identified on contigs $\geq$10,000 Bp using the software RDP3 using five independent detection algorithms: RDP (Martin and Rybicki 2000), GENECONV (Padidam, Sawyer, and Fauquet 1999), Maxchi (Smith 1992), Chimaera (Posada and Crandall 2001), and 3Seq (Boni, Posada, and Feldman 2007). All of these tests are available in the Software Package RDP for Microsoft Windows (Martin et al. 2015). Tests were conducted using a critical value = 0.05 and p-values were Bonferroni corrected for multiple comparisons of sequences. Sequences were made linear using unphased base calling, i.e. where a sequence has a base position that is heterozygous, one of the nucleotides was assigned at random at that site.

Recombination events were only considered genuine if they were supported by at least three of the recombination detection methods in RDP, and recombination events detected using the methods in RDP were only counted if the parental sequences could be identified, and the start and end positions of recombination events were unambiguous.

In order to visualise the effects of recombination and hybridisation on the genome structure of the *Albugo candida* races, the software package HybridCheck was developed for the R programming language. The development and testing of this software package is described in detail in chapter 2, so only a brief description will follow. HybridCheck can analyse three sequences with a sliding window scan, and produce plots with use the RGB tricolour system to indicate where regions of hybridisation or recombination have occurred between sequences. Each sequence is designated one of the three primary colours, red, blue, or green. In regions of a given sequence that are unique, then those regions are coloured in with the unique

colour of that given sequence. However, in regions of the sequences in which all the SNPs are shared with another sequence, then the region is coloured with the hybrid colour of the two sequences (e.g. yellow if the two sequences have the unique colours red and green). All monomorphic sites are excluded in this computation. In cases where recombination is recent, the hybrid colouration is strong as most of the SNPs are shared between two sequences. However older events may have accumulated mutations since the recombination / hybridisation event. In such a case, there are less shared SNPs between two sequences, and the colour intensity is less strong.

3.2.4 Dating identified recombination events

Immediately after a recombination or hybridisation event has occurred, a hybrid or recombinant offspring's DNA sequence will have regions which are near identical to one parent, and regions which are near identical to the other parent. In those regions the molecular clock is effectively zeroed. Therefore, for a given recombinant region, the only substitutions which could be observed between the recombinant and the donor must have occurred since the recombination event took place.

This divergence between a donor sequence region, and the same region in the recombinant offspring was used to estimate the time since the recombination event. Two methods were used to calculate the number of generations since individual identified recombination events occurred. A binomial mass function was used, which was developed for the HybridCheck R package. The equations are described more fully in chapter 2. Briefly, the method computes a window of time, within which the recombination event is most likely to have occurred. It does this by taking into consideration the cumulative probability of observing the number of mutations that have occurred in the recombinant region, between donor and parent, given

the average mutation rate. The function assumes that the recombination event has evolved neutrally since the recombination event occurred, and that mutation rates between the two sequences were constant through time, and equal in both sequences. The mutation rate in oomycetes is unknown, and therefore the binomial mass function was used with two different mutation rates: = 106 and 107 per base per generation. This binomial mass function was used to analyse all detected recombination events.

In addition to the binomial mass function, an analysis was conducted in BEAST (Drummond and Rambaut 2007). Phylogenetic trees were estimated with a HKY + G model, a Yule tree prior, and a strict molecular clock assumption, where the mutation rate was assumed to be $\mu = 10^6$. Ten independent analyses were run, with an MCMC of 10 million steps, with a burn-in of 10%. Because of the computational complexity and time required for BEAST analyses, 20 recombinant regions were analysed in this manner. The results were compared to the date that was estimated for the recombinant region by the binomial mass function, and this confirmed that the binomial mass function provides a good approximation of the divergence time.

3.3 Results

3.3.1 Distribution of polymorphisms across races

Polymorphisms were found to be unequally distributed across the genomes of the *Albugo candida* isolates analysed. In some regions of the genome, there are stretches of identical sequence which are as long as 10kb in length. In other regions of the genome, stretches of lower sequence similarity may be found. For example, between the isolates AcBoT and

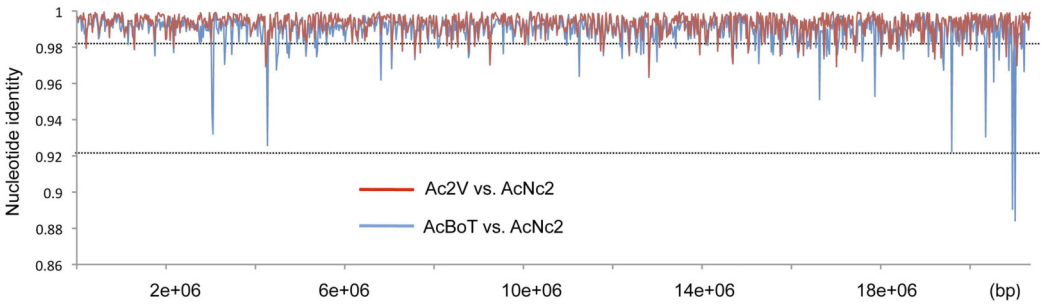


Figure 3.1: Nucleotide identity amongst the homologous genomic regions of Ac2V, AcBoT and AcNc2. The mean identity was calculated for the sliding window of 20 Kb.

AcNc2, a region of approximately 5kb was observed with 89% sequence similarity. This is demonstrated in Figure 3.1.

The distribution of the polymorphisms is highly suggestive of a mosaic-like genome as the polymorphisms are not only distributed unevenly, but they were distributed in a block-like manner. Stretches of nucleotide similarity are arranged in a block like structure; there are regions where AcNc2 is highly similar to AcBoT (and therefore diverged from isolate Ac2V), followed by regions where isolate AcNc2 is highly similar to Ac2V (and therefore diverged from isolate AcBoT). The HybridCheck software package visualises such genome structure in Figure 3.2. The figure visualises the effect on the genome by colouring regions yellow where races AcNc2 and AcBoT show near sequence identity, cyan where races AcBoT and Ac2V show near sequence similarity, and purple where races AcNc2 and Ac2V show near sequence similarity. Note that in the figures, there are also regions of unique colouration (red, green, and blue), and such regions represent diverged parts of the genome where the three races have large proportion of unique (races-specific) polymorphisms (Figure 3.2).

This observation of alternating blocks of high sequence identity between otherwise diverged (as represented by areas of red, green, and blue) genomes, provides supporting evidence for genetic introgression between diverged races that show a considerable (yet still incomplete) level of

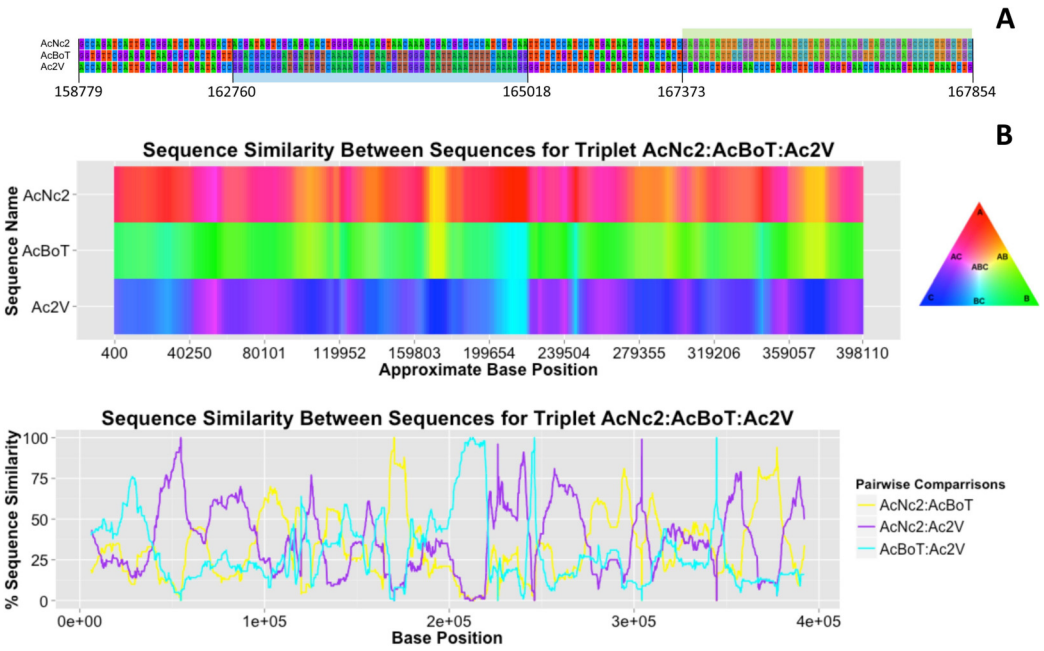


Figure 3.2: Extensive variation in sequence similarity between *Albugo candida* races. **A)** An sequence alignment between base positions 158,779 and 167,382 within contig 1 of *A. candida* races AcNc2, AcBoT and Ac2V. Two recombination blocks coloured blue and green are visible, displaying high sequence similarity between races. **B)** The sequence similarity across the length of contig 1, amongst three *A. candida* races. Similarity is visualised using the colours of a RBG colour triangle in the software HybridCheck. Areas where two contigs have the same colour (yellow, purple or turquoise) are indicative of two races sharing the same polymorphisms. The linear plot of the proportion of SNPs shared between the three pairwise comparisons between the races. Shown on the X-axis is the actual base position.

reproductive isolation. The recombination detection methods described in the previous section test for recombination blocks visualised here, formally.

3.3.2 Recombination blocks identified using RDP

All 133 contigs were analysed for presence of recombination blocks using algorithms in the software package RDP. Recombination analysis with these algorithms identified 675 recombination blocks on 127 sequence contigs which were significant, even following correction of the alpha with a Bonferroni correction. These identified blocks were reported as significant for at least three different recombination detection tests. If the length of all the significant blocks is summed in a linear fashion, then approximately

25% of the total length of all contigs analysed is identified as recombinant, this is equal to 3Mb. These blocks represent regions of the genome which are derived from either another race, or the ancestor of another race. Algorithms in RDP were able to report such donor sequences in some cases. The full data-set from the RDP output is publicly available from http://dx.doi.org/10.7554/eLife.04550.015.

3.3.3 Estimated ages of recombination events

Dating analysis of the significant recombination blocks using the Hybrid-Check binomial algorithm indicated that the recombination events detected occurred at a range of different dates. If one assumes a $\mu = 10^{-8}$ substitution rate which is constant across cell cycles, and that there are 100 cell cycles per year, then the most recent introgression event occurred approximately 220 years ago, and the oldest detected event occurred almost 200,000 years ago. The mean age for all the detected recombination events is approximately 6237 years ago, with a standard error of 12,594 years. Furthermore, there is no significant difference between the average estimated dates across different contigs.

The wide range in age estimates of the introgressed regions provides evidence for the hypothesis that recombination and hybridisation between diverging *Albugo candida* races has been a consistent and ongoing evolutionary process, affecting the entirety of the genome. This finding rules-out the hypothesis that one or a few recombination/hybridisation events in the distant past are responsible for creating the mosaic structure observed. This also helps explain the cause of the mosaic genome structure that has been observed: occasional introgression events across a range of evolutionary times is expected to result in genome containing introgressed blocks of sequence from a donor race, interspersed inside the distinct genomic background of the recipient race (i.e. the very pattern observed in *Albugo*

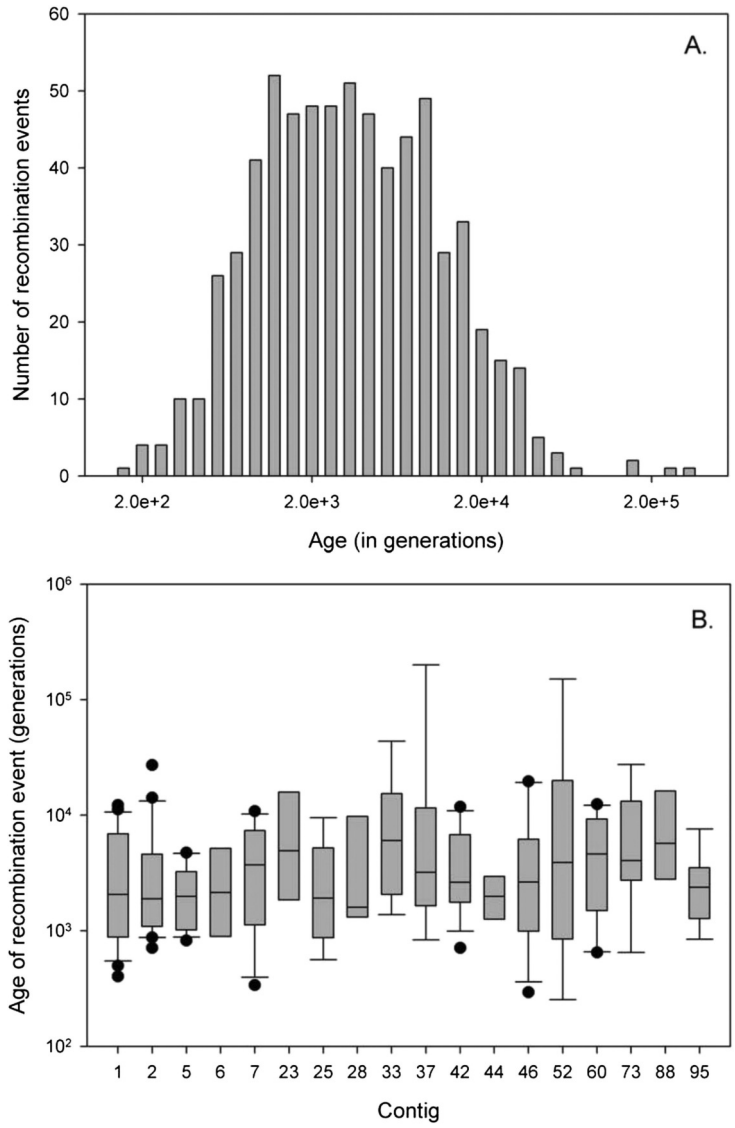


Figure 3.3: A) Age of the 675 recombination blocks, identified across the whole genome, estimated using the HybridCheck binomial mass function, assuming a substitution rate of $\mu = 10^6$; **B)** A box plot of the median (plus first nation blocks and third quartile) log-age of recombination events in contigs. Only contigs with eight or more events are shown. There is no significant difference in age of events between contigs (GLM: $F22, 233 = 1.06$, $p = 0.387$).

candida).

3.4 Discussion

The genome of *Albugo candida* appears to have a mosaic-like genome structure: 675 regions were identified in 127 analysed contigs, which were consistently identified by multiple and independent recombination detection

methods. The mosaic-like structure reflects discordant phylogenetic signals of genomic regions with distinct coalescence, and this suggests that introgression has occurred at a range of time points throughout the evolutionary history of the *Albugo candida* races.

3.4.1 Hybridisation and clonal reproduction of *A. candida*

Albugo candida is an obligate biotroph, growing and reproducing on living plant tissue, and virulence experiments confirm that the *Albugo candida* races isolated in this study are indeed host specific (McMullan et al. 2015). To explain the observed mosaic genomes, two distinct and host specialised *Albugo candida* races would have to make contact by colonizing the same host plant in order to hybridize, although ex-situ hybridisation cannot be ruled out. Yet, any *Albugo candida* race landing on a non-host plant is likely to trigger host immunity before it can mate with another distinct race. So, given that the genome structure expected from recent introgression between distinct races is observed, how have they made contact? One potential explanation was that infected host plants could form secondary contact zones for *Albugo candida*: if a host plant was infected by a compatible (infectious) *Albugo candida* race its immunity would be suppressed. With a suppressed immune system, non-specialised *Albugo candida* races might be able to colonise the already infected host, enabling both races to make contact and hybridise through sexual reproduction. This hypothesis was tested with experimental infections of host plants with multiple races. These experiments confirmed that a virulent race of *Albugo candida* could suppress the immunity of its host plant, such that other non-virulent races of *Albugo candida* could co-colonise it (Cooper et al. 2008; McMullan et al. 2015).

Following formation of a viable hybrid, clonal reproduction would allow fast dispersal of the pathogen and population expansion. This aspect of the model was supported by analysing genomic identity between isolates which infect the same host species (i.e. within different races) and quantifying the shared proportion of heterozygous sites. Genotypic similarity at heterozygous sites of pairs of independent isolates that infect the same host plant was exceptionally high; AcBoT and AcBoL shared 97% of their heterozygous sites in common, and AcEm2 and AcNc2 shared 99.95%. Sharing of this proportion of heterozygous sites rules out Mendelian segregation and sexual reproduction, and confirms that these isolated were reproduced clonally. Given that AcEm2 and AcNc2 were sampled 100 miles apart geographically, and ten years apart in time, clonal reproduction appears to be the principal mode of reproduction of this race of agronomically important pathogens.

The largest contig of the reference assembly, (contig 1; 400kb) was used to analyse polymorphism distribution and detect recombination blocks. The proportion of heterozygous sites in contig 1 was calculated for each isolate. Very few sites of contig 1 were heterozygous within AcNc2 (0.03%), AcEm2, and Ac2V (0.01%). Within isolates AcBoT and AcBoL, the proportion of heterozygous sites was higher (both 0.65%). The high levels of genotypic identity observed between isolates which infect the same the host species would not be expected if sexual reproduction and Mendelian segregation was the primary mode of reproduction, especially given that isolates AcEm2 and AcNc2 are separated by approximately 100 miles and 10 years. Furthermore, the high proportion of heterozygous sites (for contig 1) in isolates AcBoT and AcBoL is more consistent with asexual population expansion: A diploid organism reproducing asexually/clonally most of the time will accumulate mutations between each pair of homologous chromosomes. This will generate more heterozygous sites over time, resulting in

allelic divergence and increased observed heterozygosity. However, the observation of a low level of observed heterozygosity in AcEm2 and AcNc2 is not expected in organisms where asexual and clonal reproduction is the primary method of reproduction. Given there is no evidence of self-fertilisation (or any other form of asexual reproduction), it is likely that gene conversion has been operating to reduce within genome diversity in the races over time. The phenomenon is called Loss of Heterozygosity or LOH, and it has been observed in other plant pathogen species such as *Phytophthora capsici* (Lamour et al. 2012), as well as at a whole genome scale in yeast (Diogo et al. 2009). In both studies it was hypothesized the Loss of Heterozygosity observed has facilitated rapid adaptive evolution and genome plasticity.

To summarise, it appears that the generalist pathogen *Albugo candida* is comprised of distinct physiological races, which are diverging as they specialise on different host species. Secondary contact between distinct races on an immunosuppresed hosts results in inter-specific sexual reproduction between races, producing new hybrid offspring. These hybrids may be able to spread rapidly by clonal reproduction on their own, or introgression may occur.

3.4.2 Biology of genetic introgression and hybridisation

Introgression is defined as the transfer of genetic information (DNA or RNA) from one species (or OTU, race, or biotype) to another as a result of hybridization between them followed by repeated backcrossing (Ridley 2004; Abbott et al. 2013).

Hybridisation and introgression can lead to a mosaic-like genome structure, with regions of different parental lineages interspersed throughout the genome (Baack and Rieseberg 2007; Stukenbrock et al. 2012). Those

regions will have different ancestry or coalescence, and hence, be represented by different phylogenetic trees. Introgression has the potential to augment the adaptive evolutionary potential of populations and introduce a source of genetic variation into genomes. As a source of genetic variation, mutations have longer waiting times, and lower initial frequencies. In contrast, introgression can occur multiple times, thereby increasing the probability of fixation of the variant. Furthermore, whereas mutations tend to be neutral (Kimura 1968), or have (slightly) deleterious fitness effects (Ohta 1973), introgression inserts pre-selected variation of one of the parental (donor) lineages into the hybrid line (Hedrick 2013). Adaptive introgressed variants can be new, have less pleiotropy, less strong linked effects, and less recessivity (Hedrick 2013). In contrast to mutation, multiple simultaneous changes across multiple loci are possible with hybridisation and introgression, but whether these multiple changes are deleterious or not depends on the details of the molecular interactions within the hybrid.

The view of Wright is that selection favours favorably interacting gene combinations, resulting in a highly integrated genome which contains coadapted gene complexes (Wright 1931; Wright 1932; Dobzhansky 1970). However, Fisher argued that selection acts on individual genes, and would favour genes which increase fitness on average across all possible genetic backgrounds of a given lineage, such genes were called ”good mixers” (Fisher 1930). Both of these views are compatible with the concept of negative epistasis (Hedrick 2013; Burke and Arnold 2001) in a hybrid genetic background (also called hybrid incompatibility): In any two separated lineages, fixation of alleles in one lineage occurs independently and there is no selection for compatibility with any other lineage. Hybridisation produces novel genotypes which have not previously been subject to selection, and if they are less well adapted than the parental genotypes, selection would act against such less fit hybrids. This reduction in fitness of segregating hybrids

has been taken as evidence for unfavorable interactions between genomes of parental individuals, negative epistasis, and hybrid incompatibility. The most widely accepted model of such incompatibility was developed by Bateson, Dobzansky and Muller (Dobzhansky 1936; Muller 1942). Negative epistasis has been confirmed empirically in several animal and plant organisms in the past, including (but not limited to) *Drosophila spp.* (True, Weir, and Laurie 1996; Palopoli and Wu 1994; Hollocher and Wu 1996; Cabot et al. 1994), *Helianthus spp.* (Rieseberg et al. 1996; Rieseberg, Whitton, and Gardner 1999), *Tigriopus californicus* (Burton 1990b; Burton 1990a; Burton, Rawson, and Edmands 1999), and *Iris spp.* (Cruzan and Arnold 1994; Burke, Voss, and Arnold 1998), and is a primary cause of hybrid inferiority.

However, hybrids can be superior to their parental lineages. Hybrid fitness can occur by several means. F1 hybrids are commonly larger in body size and have higher growth rates and yields (Baack and Rieseberg 2007; Hedrick 2013; Burke and Arnold 2001). Such vigour is called heterosis, and is explained by the dominance and the over-dominance hypotheses (Baack and Rieseberg 2007; Lippman and Zamir 2007). Other explanations posit that synergistic interactions between different alleles at different loci (i.e. positive epistasis and inheritance of complete co-adapted linkage blocks), and changes in gene expression can also contribute to heterosis (Baack and Rieseberg 2007; Swanson-Wagner et al. 2006). Heterosis may contribute towards the establishment of an asexual or allopolyploid hybrid. Fitness resulting from Heterosis may be short lived, for introgressed hybrid lineages. This is because sexual reproduction over several generations would cause loss of heterozygosity in the subsequent (backcrossed) generations. Instead, long term success depends largely on the fixation of novel favorable gene combinations from the two parents (Baack and Rieseberg 2007; Burke and Arnold 2001). The genes in such combinations must either

interact favorably with other genes in the combination to increase fitness, or increase fitness in an additive way, with little or no interaction. Thus, selection and niche differentiation play a central role in the establishment of these relatively fit hybrids, because otherwise competition and gene flow with parental populations may overwhelm them (Buerkle et al. 2000; Rieseberg, Archer, and Wayne 1999). Just as evidence of negative epistasis has been found empirically in several species, empirical evidence of epistasis producing relatively fit hybrids has also been found for several species. For example, in addition to confirming cases of hybrid inferiority in *Helianthus spp.*, Rieseberg and colleagues also found beneficial epistatic interactions in hybrid of *Helianthus annuus* and *Helianthus petiolaris* (Gardner et al. 2000; Rieseberg et al. 1996). Evidence of favorable cytonuclear interactions was found in hybrids of *Iris fulva* and *Iris brevicaulis*, indicating that as well as interactions between genes, interactions between the nucleus and the cytoplasm can also determine the success of a hybrid (Burke, Voss, and Arnold 1998). Hybrid lineages may also exhibit transgressive segregation i.e. they may have more extreme trait values than either of the parents, when the parents possessed alleles of opposing effects. This may be beneficial or deleterious, depending on the nature of the trait and may be caused by epistasis, or, as QTL analyses have demonstrated, through additive effects (Baack and Rieseberg 2007; Burke and Arnold 2001). Hybridisation could also help purge mutational load by the masking deleterious alleles in heterotic F1 individuals, followed by introgression of favorable alleles (Ingvarsson and Whitlock 2000).

3.4.3 Introgression and evolution of *Albugo candida* in the wider context

Given the potential advantages of introgression, it has been hypothesised that introgression it is instrumental in generating novel combinations of pre-selected virulence effectors from different diverged races in *Albugo candida* (McMullan et al. 2015). Not all such combinations may be successful or viable, but successful genotypes would be important in facilitating the colonisation of new hosts i.e. a host jump. As a hypothetical example, the *Albugo candida* race Ac2V is proposed to possess an effector allele, which interacts with an *Arabidopsis* R gene called WRR4. This prevents Ac2V from colonising *Arabidopsis*. It is unknown which effector interacts with WRR4, but if the effector allele segregated away in hybrid offspring, or was removed through loss of heterozygosity, the hybrid offspring may be able to overcome *Arabidopsis* resistance.

The impact of introgression and hybridisation has been demonstrated in other species. For example, in sunflower species *Helianthus anomalus* (Ungerer et al. 1998). *Helianthus anomalus*, like *Albugo candida*, has a genome which appears to be composed of distinct parental blocks. However, unlike *Albugo candida*, the introgression was dated as occurring over a short timespan of 10 - 60 generations, which provides support for the idea that hybrid speciation is a punctuated process (Ungerer et al. 1998). The dating analysis of blocks present in *Albugo candida* suggests that introgression has occurred between different races at different times, and repeatedly throughout the evolution of the species. Furthermore, unlike *Albugo candida*, introgression in the sunflower species occurred between two different species, and resulted in a new hybrid species. For *Albugo candida*, whilst the races are isolated from each other most of the time, repeated introgression between them during secondary contact on immunosuppressed

host plants likely acts to prevent them becoming completely isolated, new species. A classic example of an adaptive radiation is Darwin's Finches (*Geospiza*, *Certhidea*, *Pinaroloxias*, and *Camarhynchus/Platyspiza* spp.), and even here hybridisation has been demonstrated (Lamichhaney et al. 2015): Recent whole-genome resequencing, and phylogenetic analysis based on autosomal, mtDNA, and sex-linked loci of 120 birds representing all of the Darwin finch species and two other related species revealed discordant phylogenies (Lamichhaney et al. 2015). Calculations of Patterson's D, supported the hypothesis of gene flow and hybridisation throughout the radiation (Lamichhaney et al. 2015). Rare introgression is thought to have facilitated the exchange of mimicry genes between *Heliconius* butterfly species, post isolation (Martin et al. 2013).

Studies from hybridisation with yeast provide findings which corroborate the findings of this study. For example, genetic exchange between 3 strains of *Saccharomyces cerevisiae* has been quantified, and indicates that for these strains out-crossing has only occurred 314 times during approximately 16 million cell cycles (Ruderfer et al. 2006). This is approximately one out-crossing event per 50,000 cell cycles. Thus while the strains of yeast do mate and recombine in the wild, this is not a frequent occurrence (Ruderfer et al. 2006). This is also what has been inferred for *Albugo candida* as the result of this study. In addition, the genomes of wine strains of *Saccharomyces cerevisiae* contain introgressed blocks from the species *Saccharomyces paradoxus*, *Saccharomyces kudriavzevii kudriavzevii*, *Saccharomyces uvarum uvarum*, and *Zygosaccharomyces bailii* (Dujon 2010). The blocks in the genome of *Saccharomyces cerevisiae* are almost identical to the corresponding regions in the genomes of the donor species, indicating that the introgression events have been recent (Dujon 2010). This is similar to what this study has demonstrated for *Albugo candida*. It appears that introgression is a general phenomenon in yeast genomes,

but one review concluded that its importance in its evolution has yet to be determined.

The importance of introgression in the evolution of *Albugo candida* is hypothesized to be as follows: Isolation, divergence and specialisation of races will generate repertoires of tried and tested effectors for a specific race. Those adapted race-specific repertoires are then brought together when two races hybridize to generate novel repertoires of novel combinations of these effectors. Specific avirulence effectors that trigger host immunity may be lost through segregation and the Loss of Heterozygosity (LOH) effect hypothesized to be taken place in oomycetes by Lamour et al. 2012, and documented here and in McMullan et al. 2015. These hybrids, having new combinations of effectors, and having lost effectors which impeded their colonisation of other hosts previously, may expand their geographical range and population size clonally. Such new hybrids may be able to colonise new hosts, explaining the phenomenal host range of species such as *Albugo candida* (and possibly other generalists). Hybridisation between races has been shown to expand host range in other plant pathogen species such as *Phytopthora* spp. (Ersek, English, and Schoelz 1995), and the transfer of virulence genes leading to host range expansion has also been demonstrated in bacterial and fungal pathogens (Ford Doolittle 1999; Mehrabi et al. 2011). Sexual oospores of *Albugo candida* are tolerant of strong environmental pressures, which raises the prospect: might hybrid spores produced by reproduction between two races lie dormant, forming banks of hybrid genotypes, waiting for conditions better suited to their genotype and phenotype?

The ability to expand host range and generate novel genotypes through hybridisation, and then reproduce rapidly clonally may be especially favored in a monoculture based agro-ecological environment, characterized by different, large, homogeneous regions of (often clonal) host plants of

one species (Stukenbrock and Bataillon 2012). Recently Stukenbrock et al. 2012 demonstrated that the plant pathogen species *Zymoseptoria pseudotritici* was formed by the hybridisation of two distinct fungal individuals, and that the genome is characteristic of bottleneck and selection following the hybridisation event which occurred approximately 380 sexual generations ago, resulting in the generalist grass pathogen. The obligate biotroph and powdery mildew, *Blumeria graminis f. sp. Hordei* also has a mosaic genome of alternating monomorphic and polymorphic DNA sequence blocks (Hacquard et al. 2013). Pathogen adaptation to agro-ecological environments is characterized by high genome plasticity of pathogens (a successful pathogen needs to keep up in the co-evolutionary arms race with its host), but a reduction in diversity for recently emerged lineages (selection is strong and new and recently emerged lineages are often bottlenecked) (Stukenbrock and Bataillon 2012). Pathogens such as late blight of potato, *Phytophthora infestans*, wheat yellow rust *Puccinia striiformis*, and *Magnaporthe oryzae*, which are specialised, may represent an end-result of a much broader process of pathogen adaptation and evolution. The results gained from this work provide insight into how recombination and hybridisation plays a role in generating novel virulent races, and into their subsequent spread and geographical range expansion by clonal propagation. These findings are of particular relevance to modern, monoculture based agriculture.

CHAPTER 4

Allelic divergence in the polar diatom *Fragilariopsis cylindrus*

This chapter is based on a submitted scientific paper:

Mock, T., Otillar, R. P., Strauss, J., Allen, A. E., Dupont, C. L., Frickenhaus, S., ... Grigoriev, I. V. (Submitted). Extensive genetic diversity and differential bi-allelic expression in a Southern Ocean diatom. Nature.

This project was a very large collaboration spanning many years to sequence the genome of the *Fragilariopsis cylindrus* organism. In order to clearly describe my work and set it in context, some work that was not performed by myself is described. In particular, any work mentioned in the introduction is not my contribution to the work, but was completed by colleagues. My contributions to the work are described in sections 4.2.2.1, 4.2.2.2, 4.2.2.3, 4.2.2.4, and the results section presents data that was the outcome of my work only. In the discussion some further preliminary work is described. A figure showing this work is provided as an appendix, and this work was done jointly and equally between myself and a colleague.

4.1 Introduction

4.1.1 Sexual reproduction and recombination

Sex as a mode of reproduction has a two-fold cost. Firstly, most sexually reproducing species only have one gender capable of bearing offspring (Visser and Elena 2007). Secondly, in sexually reproducing organisms, any individual will only contribute approximately half of its genetic information to each offspring; i.e. in diploid sexuals, gametes are haploid (Agrawal 2001). In contrast, an asexually reproducing, clonal organism contributes all of its genetic information to each offspring, and every individual is typically capable of bearing young (Schlupp, Taebel-Hellwig, and Tobler 2010). This generalization applies to most sexual organisms however, there are exceptions. For example, not all sexually producing organisms have the two-fold cost problem. Yeasts are sexual organisms with two mating types and both types are capable of producing offspring. In addition, a species of poecilids can reproduce through a process of gynogenesis; a process similar to asexual reproduction through parthenogenesis, but is distinct as the presence of sperm is required to stimulate egg development (Schlupp, Taebel-Hellwig, and Tobler 2010). Hybridisation has also given rise to a Hermaphroditic Cichlid individual which can self (Svensson et al. 2016). In addition, some species shuttle between asexual and sexual reproduction, and the frequency at which this happens directly affects the factors raised above.

All else being equal, an asexual species should outperform a sexual species over time because of its faster population growth rate. However, sexual and asexual species do co-exist together, sometimes with similar fecundity (Schlupp, Taebel-Hellwig, and Tobler 2010). However, despite this, sexual reproduction is very widespread, especially among the eukaryotes.

These observations led researchers to think that the benefits of sexual reproductions must be evolutionary and lead to the production of offspring with benefits that outweigh to costs. To summarize most of the commonly cited reasons sexual reproduction is maintained, it may be described as a mechanism, through which:

1. Beneficial mutations can spread through a population more quickly.
2. Novel genetic combinations are generated.
3. Deleterious mutations can be purged or masked.

These benefits are possible because sexual reproduction brings together into one individual, the chromatids (and alleles they contain) in the gametes of two parental individuals from separate genealogical lines (out-crossing). In addition, when parental individuals generate gametes, meiotic recombination will result in new combinations of genes (Felsenstein and Yokoyama 1976). This in turn contributes to the generation of novel genetic (or rather, genotypic) variation. As a result, two or more beneficial mutations from separate genealogical lines may occur together within the same individual, thus facilitating the spread of beneficial mutations through the population to fixation.

This is formalized by the Hill-Robertson effect (Hill and Robertson 1966), and is demonstrated by considering two loci with the haplotype A_2B_2 with a fitness of 1. It is then assumed two mutants at both loci (A_1B_2, A_2B_1) can occur after a time period with fitnesses of $1 + s$, and that fitnesses are multiplicative such that A_1B_1 has fitness $(1 + s)^2$. With no or low recombination, the ancestral haplotype is lost by selection and both advantageous mutants will exist in the population for some time until one is lost by drift (Coop and Przeworski 2007). But with recombination, a haplotype A_1B_1 is possible, bringing both mutants together in one haplotype before one of the

mutants is lost by drift, thus both mutants get fixed rather than one Coop and Przeworski 2007. With low recombination rates selection increasing the frequency of the mutant alleles is less effective, this is the Hill-Robertson effect (Hill and Robertson 1966).

The effect is more likely to occur when selection is not too strong, recombination rates are low, and when the favorable mutants have negative disequilibrium i.e. they initially occur on different haplotypes (Hedrick 2010). An asexual lineage, in contrast would have to acquire one beneficial mutation, followed by another, a limitation called clonal interference (Gerrish and Lenski 1998).

Similarly, deleterious mutations accumulating throughout the population in different genealogical lines may occur together within one individual, which suffers stronger negative selection pressure and is eliminated from the population (Crow 1994). A third possibility is a deleterious allele is inherited from one parent, and the corresponding allele inherited from the second parent is not deleterious. In that case, the affects of the deleterious allele may be alleviated or masked, as the offspring individual still possesses a non-deleterious copy. Chromosomal crossover during meiosis may also result in the removal of deleterious mutations (Crow 1994).

The maintenance of sexual reproduction has also been attributed to its role in DNA mismatch repair (Bernstein, Bernstein, and Michod 2011). The repair and complementation hypothesis proposes that sexual reproduction is an adaptive response to incorrect DNA replication, through mutation and damage to the DNA molecule (Bernstein et al. 1984; Bernstein 1985; Bernstein, Hopf, and Michod 1987). Recombination repair is the only mechanism currently known which removes double stranded damages to the DNA molecule and such double strand damage is common and could be lethal if not repaired: in human cells such damage occurs approximately 50 times per cell cycle (Vilenchik and Knudson 2003).

Recombination and sexual reproduction also plays a role in eliminating detrimental variation from the population, which otherwise would accumulate over time and decrease the fitness of the population (Muller's ratchet) (Muller 1932). Recombination produces individuals containing fewer deleterious mutants, helping to reverse the decline in fitness.

The Red Queen Hypothesis also offers an explanation as to why sex has repeatedly evolved in all life forms (Paterson et al. 2010). It states that in a rapidly changing environment, alleles that were previously neutral or deleterious and the rapid change makes sexual reproduction advantageous. Such rapid changes are proposed to be particularly evident during co-evolution between a parasite and its host (Decaestecker et al. 2007).

However, despite the advantages of sex, evidence of ancient asexuality has been identified in the genomes of some organisms including root-knot nematodes and bdelloid rotifers (Lunt 2008; Welch and Meselson 2000; Meselson and Welch 2007; Pouchkina-Stantcheva et al. 2007). The classic hallmarks of ancient asexuality are diverged alleles and a lack of phylogenetic incongruence caused by recombination (Schurko, Neiman, and Logsdon 2009).

4.1.2 *Fragilariopsis cylindrus* and Diatoms

Fragilariopsis cylindrus is a species of Diatom: microscopic eukaryotic phytoplanktons, which are found throughout all the worlds oceans wherever there is sufficient light and nutrients to support them (Armbrust 2009). Diatoms are so named because of their shape and method of reproduction: Their cells are covered by a silica cell wall made of two halves, and they reproduce by asexual mitotic division, decreasing in size each time. Diatoms occasionally reproduce by forming an auxospore, which reverses the decline in size resulting from reproduction by mitotic division (Armbrust 2009). Auxospores also play a role in sexual reproduction, forming after

haploid gametes fuse to form a diploid zygote. Diatoms are an important group of organisms of study because of their role in the ecosystem and in marine biogeochemical cycles (Assmy et al. 2013; Thomas and Dieckmann 2002; Pondaven et al. 2000).

Diatoms provide an important ecosystem service by performing photosynthesis. It has been estimated that of all photosynthesis that occurs on earth, one fifth is performed by Diatom species. Each year diatoms generate as much organic carbon as that produced in total by all the terrestrial rainforests on Earth (Armbrust 2009). The organic carbon that is produced by diatoms by photosynthesis is input into food webs: in coastal regions diatoms support fisheries (such as anchovies in the Peruvian ocean) and in the open-ocean, much of the organic matter produced sinks and becomes food for deep-sea organisms (unless is reaches the ocean floor, where it may become sequestered in sediment and rock) (Armbrust 2009; Bowler, Vardi, and Allen 2010). As a result, a significant amount of petroleum deposits under the ocean floor are derived from diatoms sinking.

As Diatoms are found throughout all the worlds oceans, they populate interesting and dynamic environments in which environmental factors change and can become extreme. They are known to be adapted to limited iron, extremes in temperature (Arrigo et al. 2012; Bayer-Giraldi et al. 2011; Bowler, Vardi, and Allen 2010), salinity (Krell 2006), and temporal variation in the environment: seasons cause rises and falls in temperature, and freezing and melting sea ice also means the environments structure can be heterogeneous through time. All these extremes occur in the environment of *Fragilariopsis cylindrus*, which is particularly successful in the Southern Ocean, and is often found to form large populations in the bottom layer of sea ice and the wider sea-ice zone including open waters (Kang and Fryxell 1992). Such ice is characterized by temperatures below the freezing point of sea water, high salinity caused by the semi-enclosed pores within the ice,

and low diffusion rates of dissolved gases and inorganic nutrients (Thomas and Dieckmann 2002). The environment is not limited in dissolved iron however, unlike the surface ocean (Wang et al. 2014). Furthermore, the environment is dynamic: every winter, phytoplankton in the Southern Ocean get locked into sea ice and are released again in the following summer, when most of the sea ice melts (Vancoppenolle et al. 2013). However, only a subset of these phytoplankton them have evolved adaptations to cope with this dramatic environmental change, including *F. cylindrus*, which is known to thrive in both habitats (Bayer-Giraldi et al. 2011; Vancoppenolle et al. 2013).

How Diatoms have adapted to such conditions, and become so successful in the oceans, is of interest to evolutionary biologists and genome sequencing has provided insight. Complete genome sequences are available for two Diatom species (*Thalassiosira pseudonana* and *Phaeodactylum tricornutum*), containing between 10 and 14 thousand genes. However, of those genes only approximately half can be assigned a putative function based on experimental knowledge (Bowler, Vardi, and Allen 2010). Furthermore, approximately 35% of the genes found are specific to each Diatom, which suggests some of them encode adaptations to specific environmental conditions (Bowler, Vardi, and Allen 2010). As secondary genome sequences became available, the origin of Diatoms seems to be a secondary endosymbiosis between red algae and a heterotrophic eukaryote, and surprisingly many bacterial genes were identified, highlighting the role of HGT in the evolution of Diatom species (Bowler, Vardi, and Allen 2010; Raymond and Kim 2012).

Diatom specific genes were found to have high diversification rates, and since *Thalassiosira pseudonana* and *Phaeodactylum tricornutum* diverged approximately 90 million years ago, and the two have diverged as much as metazoans had diverged in approximately 550 million years (Bowler, Vardi,

and Allen 2010). It is thought that diversification in Diatoms has been driven by transposable elements, which increased the rate of insertion, deletion, and recombination events (Bowler, Vardi, and Allen 2010). In contrast, diversification of genes in metazoan genomes during the aforementioned 550 million years, is thought to have occurred largely through whole and segmental gene duplication events (Bowler, Vardi, and Allen 2010). Some of the diatom specific transposons are activated in response to stresses such as Nitrogen starvation, suggesting diversification of Diatom genes may be stimulated by environmental cues (Bowler, Vardi, and Allen 2010). The resulting mix-and-match genomes (Armbrust 2009) of Diatom species has brought together unique combinations of genes facilitating adaptation to a range of environments, including that encode unique pathways of nutrient assimilation. comparing the genome of a psychrophile such as *F. cylindrus* with that of diatoms evolved in temperate oceans provides an opportunity to obtain first insights into how this species has evolved to conditions of Southern Ocean waters, and managed to persist for millions of years, underpinning the ecology of an unique food web.

Recently the first large-scale genomic sequencing of *Fragilariopsis cylindrus*, a eukaryotic psychrophilic organism of ecological importance, including whole-genome sequence, transcriptome and population genetic analyses, was completed. In this thesis chapter I present my contribution to the population genetic analyses of this large body of collaborative work. This goal of the work described in this chapter was conducted in order to evaluate hypotheses about the evolutionary history of *Fragilariopsis cylindrus*. These hypotheses were proposed during the genome project, to explain observations about the genome data, and the hypotheses that I tested in this project.

4.1.3 The *Fragilariopsis cylindrus* genome project

The draft of the *F. cylindrus* genome was approximately 60Mb in length, which is larger than the sequences for the nuclear, plastid and mitochondrial genomes of the cosmopolitan diatom *T. pseudonana* (34Mb), and the whole-genome sequence of *P. tricornutum* (27Mb) (Armbrust 2009; Mock et al. 0). The draft genome of *F. cylindrus* is smaller in size compared to the toxigenic coastal species *Pseudo-nitzschia multiseries* (300 Mb) (Armbrust 2009).

Assembler programs typically use single end or paired end reads to find overlaps in sequence fragments, joining them to form contigs. Since it is known that paired end reads are generated from the same DNA fragment, this can help link contigs onto scaffolds, which are ordered assemblies of contigs, with gaps in between them (Baker 2012). However, assemblers are not always accurate: one common problem is that if one suspects that the read depth for an assembled region is too high, then it may be that the assembler has merged multiple regions because of their high sequence similarity (typically these are repeat rich regions or duplications) (Baker 2012). A second problem is if one suspects that regions of an assembly have a lower read depth than the rest of the assembly, then it may be that those regions represent single polymorphic loci, which have been assembled as two distinct loci (Baker 2012). 30.2Mb of the scaffolds of *F. cylindrus* could not be collapsed into a single haplotype, because they had greater than 1.5% nucleotide discrepancies. The genome contains just over 20,000 protein-encoding genes, and of those, 28% of them represent alleles that could not be collapsed (Mock et al. 2017). The genome contains 46 Mb of collapsed haplotype and 15.1 Mb of diverged haplotype that represents the diverged alleles of the same genetic loci.

The genome contains 21,066 predicted protein-encoding genes, 6,071

genes were represented by diverged alleles, and each pair of diverged alleles had both coding and non-coding regions, and were up to 6% polymorphic in the non-coding regions. Comparison of the diverged allele, and non-diverged allele gene ontologies (GO) revealed that genes in the categories catalytic activity (GO:0003824), transporter activity (GO:0005215), metabolic process (GO:0008152), transport (GO:0006810) and integral to membrane (GO:0016021) were significantly enriched in the diverged alleles set (Mock et al. 2017). Furthermore, biological process GO categories metabolic process (summarising lipid-catabolic process (GO:0016042), glucose metabolic process (GO:0006006), oxidation-reduction process (GO:0055114) and translation (GO:0006412)) as well as GO category transport-related categories protein transport (GO:0015031) and proton transport (GO:0015992) enriched in metatranscriptome sequences from Southern Ocean sea ice, and these sequences had high similarity to sequences contained in the diverged alleles of *F. cylindrus* according to BLASTX analyses (Mock et al. 2017).

Differential expression experiments and RNA-Sequencing suggested that 40% of the non-collapsed, diverged allelic pairs showed a 4 fold unequal bi-allelic expression (Mock et al. 2017). This suggested an allele-based adaptation to different environmental conditions. The differential expression in alleles suggested they were controlled by separate regulatory systems. Alleles showing the strong unequal bi-allelic expression were found to have an elevated rate of non-synonymous mutations, which suggests significant positive / adaptive selection and evolution of these allelic pairs (Mock et al. 2017). It was concluded therefore, that positive selection has been a driving force in the evolution of these alleles and hence the adaptation of this diatom to the environmental conditions it faces.

An evolutionary explanation of the 28% of genes that could not be collapsed (i.e. diverged genes) is desired, as it would explain one of the

mechanisms through which this diatom appears to have adapted to its polar environment. However, this signature of positive selection alone does not provide a sufficient evolutionary explanation: Meiotic recombination, which occurs during sexual reproduction, should act to homogenize any two alleles of one gene in the diatom genome.

Allelic divergence is a classic signature in genomes of organisms called ancient asexuals (Little and Hebert 1996; Pouchkina-Stantcheva et al. 2007; Schurko, Neiman, and Logsdon 2009). By its definition asexuality is a negative proposition, based on an apparent lack of sexual reproduction in an organism, and since absence of evidence is not equivalent to evidence of absence, ancient asexuality is a difficult proposition to demonstrate in an organism absolutely (Schurko, Neiman, and Logsdon 2009). Indeed the existence of ancient asexuals has been debated and doubted in the past (Judson and Normark 1996; Little and Hebert 1996), and this is perhaps unsurprising considering current theory explaining the benefits of, and maintenance of sexual reproduction.

If the divergence of alleles is due to ancient asexual reproduction, then the recombination rate between these alleles should be reduced. It was also expected that phylogenetic networks would have a very clear structure, with deep branches. To test these predictions and evaluate empirical data I performed population genetic simulations. More detail is presented in the methods section, but briefly, sequence data was available to test for the evidence of recombination based on an environmental sample of *F. cylindrus*, that was amplified by PCR and sequenced using Sanger sequencing. It resulted in 200 high quality sequences from alleles of Ferrichrome ABC transporter and Large Ribosomal Protein L10, and the signature of recombination between these alleles was analyzed as well as several other population genetic parameters.

This project had the aim of establishing whether ancient asexuality and

a lack of recombination is evident, by establish whether recombination has occurred by analyzing the aforementioned DNA sequences.

The specific aims were:

- Use LAMARC to establish a population recombination rate and population Theta parameter.
- Use the incompatible sites test to detect evidence of phylogenetic incompatibility (and therefore recombination) between closely related sequences.
- Visualize recombination signal of choice sequences with the HybridCheck package.
- Conduct a comparative phylogenetic network analysis.
  - Construct un-rooted phylogenetic networks of alleles present in the natural sea-ice populations.
  - Construct un-rooted phylogenetic networks from *silico* populations simulated using simuPOP. Some of these *silico* populations were simulated under asexual (clonal) regimes of reproduction, and some were simulated under a sexual reproduction regime, with different mutation and recombination rates.
  - Compare the empirical networks with those simulated, to try and suggest the mutation and recombination rates the Diatom population may have in nature.

4.2 Materials and Methods

4.2.1 Materials

4.2.1.1 Sequence Data: PCR Amplified Alleles

In this study, subsequently described analyses were performed using the same dataset. Two genes (ABC Iron Transporter (Protein ID 240308) and Large Ribosomal sub-unit (Protein ID 240308)) of an environmental sample of *F. cylindrus* were amplified by PCR and sequenced using Sanger sequencing to yield high quality sequences. A total of 93 and 103 alleles were found in both genes, respectively. The DNA extraction, and PCR amplification, was completed by Dr. Jan Strauss. Sanger sequencing was performed by (Mock et al. 2017). These two sequence datasets shall be referred to hereafter as FcABC (ABC Iron transporter), and FcLR (Large Ribosomal Subunit).

4.2.1.2 Sequence Data: Allelic pairs from the genome

Previously, a set of diverged alleles was defined for any downstream analyses: The genome assembly was aligned against itself using BLAST, with a 95% nucleotide identity threshold, and greater or equal to 50% alignment coverage for smaller scaffolds. Syntenic scaffolds that were homologous across their whole length were analyzed with Mauve. Diverged alleles on large scaffolds were referred to as allele 1, the corresponding allele on the smaller scaffold was referred to as allele 2. For more details, the reader is referred to the paper (Mock et al. 2017). The allelic pair set was used to estimate coalescence times between alleles, as described in the next section.

4.2.2 Methods

4.2.2.1 Estimating Coalescence times of alleles

Because the FcABC and FcLR sequences were used for recombination detection, and the calculation of networks for the simulation and network analysis portion of this study, it was important to determine the two sequence datasets were representative of the allelic pairs identified in the genome data. Therefore, coalescence times were calculated A) Between the two sequences of each allelic pair identified from the genome data (see above), B) between pairs of FcABC sequences, C) between pairs of FcLR sequences. If the distributions of coalescence times for A) FcABC, and B) FcLR, overlap the distribution of coalescence times calculated for the genome data, then the FcABC and FcLR sequence datasets could be considered representative of the allelic pairs from the genome data.

Coalescence times were estimated using the algorithm available in the HybridCheck R package (https://github.com/Ward9250/HybridCheck). The algorithms and design of HybridCheck is described in chapter 2 of this thesis. Briefly, the algorithm used estimates coalescence time of two aligned sequences based on the number of mutations that are observed between two sequences. HybridCheck models a Bernoulli trial with a strict molecular clock, which assumes a constant mutation rate (μ = 10e-9) and a Jukes and Cantor model for base substitutions.

Coalescence time estimates calculated by the HybridCheck algorithm are expressed in terms of generations, as described in chapter 2. An estimate in terms of real time (years) was desired to attempt to put the divergence of the allelic pairs into a historical context. Estimates were converted to years using an estimated division rate of 12.472 per year. This yearly division rate assumed a division rate of 0.1 per day, and a growing season of four months per year, where each month consisted of 30.4368

days. 946 allelic pairs were successfully pulled, aligned, and dated from the genome sequence data.

4.2.2.2 Testing for recombination in the PCR amplified alleles with the PHI-test

We tested for recombination in both the FcABC and FcLR sequence datasets using the PHI-test for recombination (Bruen, Philippe, and Bryant 2006). The test accepts a multiple sequence alignment and is based on the principle of refined compatibility: For a given pair of informative sites in a multiple sequence alignment, they are deemed compatible if there is a phylogenetic history that can be inferred parsimoniously, on the condition that there is no recurrent mutation, or convergent mutations (Le Quesne 1969).

If the condition is not satisfied then the sites are classified as incompatible. Incompatible sites are explained either by homoplasies, or by recombination. The PHI-test extends this notion by using the refined incompatibility score, which allows for consideration of situations in which multiple homoplasies can be parsimoniously inferred a pair of sites (Bruen, Philippe, and Bryant 2006). The PHI-test then computes the mean refined compatibility scores of nearby sites and a p-value is calculated parametrically (Bruen, Philippe, and Bryant 2006). The analyses were repeated with window sizes of 100, 50, and 10 base pairs.

4.2.2.3 Population recombination rate and theta parameter estimation with LAMARC

A population recombination rate, and the population mutation rate Θ (Theta), was inferred for the FcABC and FcLR sequence datasets, using the LAMARC software for coalescent analysis (Kuhner 2006). Five independent runs were run for both datasets, in which 20 sequences were randomly sampled

from each sequence dataset, and analysed with LAMARC, using uninformative priors and default settings, as much about *F. cylindrus* populations in the wild is unknown. These results informed the choice of the Θ parameter used in simulations as described below.

4.2.2.4 Comparative Phylogenetic Network Analysis

Population Genetic Simulations.

All simulation scenarios were written as simuPOP scripts (Peng and Kimmel 2005). Since we are interested in assessing whether *F. cylindrus* has an asexual past causing allelic divergence, when the word recombination is used in the section is specifically refers to meiotic recombination unless otherwise stated.

Two scenarios were simulated:

1. A scenario in which individuals reproduced clonally (i.e asexually) and no recombination could take place.

2. A scenario in which individuals reproduced sexually every generation, and in which the rate of meiotic recombination could be specified.

In all three of these simulations, individuals in the simulated population were diploid and so contained one pair of chromosomes each (two homologous copies). The chromosomes were 750bp in length and the pairs of chromosomes begin as identical. By initializing individuals in this manner and then evolving them, each individual containing a pair of 750bp acted as an evolving allelic pair.

When running each simulation design, various combinations of effective populations size, and mutation rates were used in a balanced manner such that Θ for the simulated populations should result in a similar Θ estimated for the FcABC and FcLR sequences by the LAMARC analysis. This permitted the preservation of the Θ parameter of the population but allowing more

reasonable compute time. Θ values of 0.66, 0.066, 0.0066, were chosen based on the LAMARC analysis, with the value 0.066 being closest to the estimates returned by LAMARC.

It was assumed that the census size set in the simulations is a reasonable approximation for the effective population size, given that in our simulations the population was panmixtic, i.e.:

- There are always an equal number of males to females.
- No one individual is more likely to produce offspring than any other.
- Mating is random when sexual reproduction occurs any male can potentially be paired with any female.
- The number of breeding individuals is always the same for all generations.

For the simulations where recombination occurs, various recombination rates (relative to μ) were used, from no recombination ($r = 0$), to $r = 0.1\mu$, $r = 0.5\mu$, $r = \mu$, $r = 5\mu$, and $r = 10\mu$.

All simulations ran on the computer for a number of generations equal to the intended effective population size multiplied by 20. The mating scheme kept the population size constant during mating, one male and one female virtual diatom is randomly picked from the population. The number of offspring they produce is drawn from a Poisson distribution with mean and variance equal to 2. This is repeated over and over until the new offspring population is of equal size to the parental population. Individuals could be randomly selected for mating more than once.

In every simulation performed, 96 individuals were randomly sampled and exported at various time points throughout all the simulation runs, and converted to FASTA sequence files. These FASTA files could then be used for generation of networks with SplitsTree (Huson 1998).

Preparation of PCR amplified allele sequences.

The population genetic simulations described above were simulated with the absence of selection pressure. Therefore, before constructing phylogenetic networks of the FcABC and FcLR sequences to compare with networks constructed from the simulated sequences, it was necessary to reduce the influence of selection as much as possible. Therefore, when constructing phylogenetic networks for the FcABC and FcLR sequences, only the 3rd codon positions were utilized. To do this, a script translated every sequence in every possible reading frame and scored the number of stop codons or unknown proteins present in the translation. It is assumed the correct reading frame for the alleles is the one in which there are no stop codon in the middle of the sequence. Furthermore, this reading frame should be the same for almost all sequences. Sequences that resulted in uncertain translations in every reading frame were not used, and only sequences that had showed one reading frame with no stop codons were used to build networks.

Calculating Phylogenetic Networks.

Phylogenetic networks were computed for the FcABC, FcLR, and simulated sequence datasets generated by each of the population genetic simulation scenarios previously described. All networks have been generated with the SplitsTree software (Huson 1998), and the methods used in the package to compute and draw the networks were the *Uncorrected_P* character transform, the *NeighbourNet* distances transform, and the *EqualAngle* splits transform.

These networks constitute an expectation of what may be seen in the networks of the *F. cylindrus* alleles under various scenarios of sexuality or asexuality. If *F. cylindrus* has a past history of asexual reproduction, we would expect networks of sequences generated by an asexual simulation to show greater similarity to the networks of the *F. cylindrus* alleles. If *F.*

cylindrus has a past history of low levels of sex then its network would show more similarity to the network derived from the model in which there is lower levels of recombination, and so on. By comparing the *F. cylindrus* networks to those modeled networks it is possible to assess whether strict asexuality or infrequent sex is a likely possibility. It is important to note any simulated scenario with sexual reproduction with a zero recombination rate is not the same as asexual reproduction as the clonal reproduction scenario as the latter does not follow Mendelian inheritance, whereas sexual reproduction, with a recombination rate of zero, does follow Mendelian inheritance.

In comparing networks of simulated allelic pairs and networks of the sequenced *F. cylindrus* sequences, characteristics regarding the structure of the network, can be expressed quantitatively. To quantitatively assess the networks, we calculated the p-distance matrices for all the sets of simulated scenario sequences, and for the real *F. cylindrus* sequences. In particular we calculated the mean and the variance both of which were expected to be higher for networks of sequences evolved with lower recombination rates, showing signs of allelic divergence. The distances reflect the mean branch length in the network and are principally affected by the mutation-drift equilibrium, and hence Θ. In order to assess the effect of recombination relative to the mutation rate (R/μ), we quantified the number splits in the network, again comparing the simulated networks with those of the *F. cylindrus* alleles.

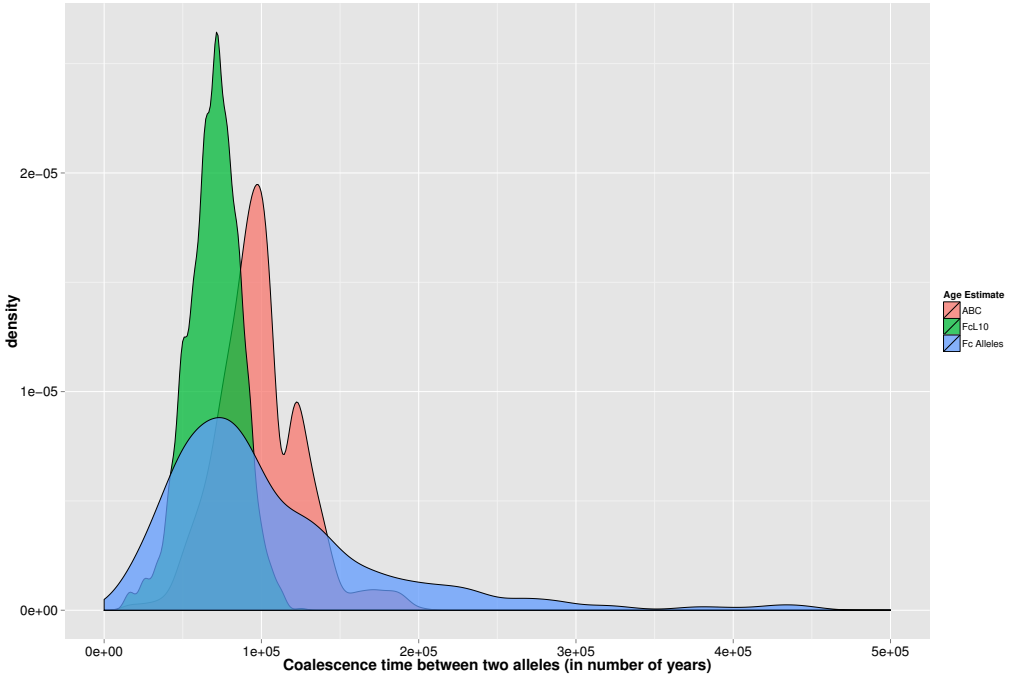


Figure 4.1: Smoothed density plot of the maximum coalescence times (in generations) calculated for allelic pairs of the ABC Iron Transporter (red), Large Ribosomal Subunit (green) and allelic pairs from the genome (blue).

4.3 Results

4.3.1 Estimating coalescence times of alleles

Figure 4.1 shows the distances calculated between the allelic pairs simulated from the ABC Iron Transporter and Large Ribosomal Subunit sequence pools, and between the allelic pairs identified from Fc Alleles RNAseq data. The three distributions show considerable overlap, which implies that the divergence between allelic pairs identified from the genome is representative of the divergence between alleles from two known genes (Figure 4.1).

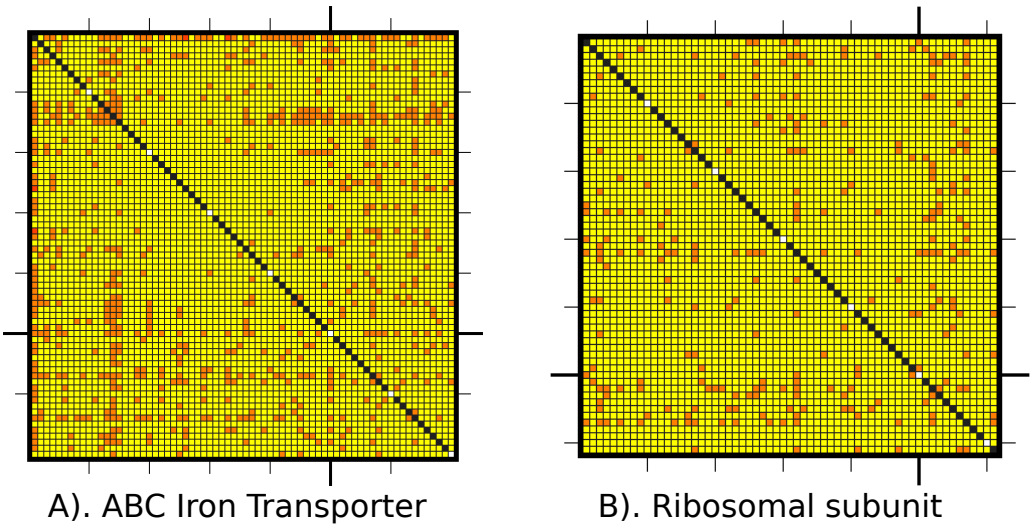


Figure 4.2: Incompatibility score matrices computed for A). The ABC iron Transporter and B). The Large Ribosomal Subunit. Yellow boxes indicate two informative sites are compatible, and darker boxes indicate the two sites are incompatible. The presence of incompatible sites in the alignments is suggestive of recombination.

4.3.2 Testing for recombination in the PCR amplified alleles with the PHI-test

PHI Scores calculated for the sequences of the ABC Iron transporter and the Large Ribosomal Subunit (Table 4.1), and Figure 4.2 shows the refined incompatibility matrices between informative sites computed for the ABC Iron Transporter (A.), and the Large Ribosomal Subunit (B.). Yellow squares indicate pairs of informative sites that are compatible, darker squares indicate a pair of sites that are incompatible. The presence of incompatible sites in these sequences, and the PHI-Scores and NSS scores shown in Table 4.1 suggests recombination has indeed affected these sequences.

4.3.2.1 Comparative analysis of phylogenetic networks

Figure 4.3. Shows an example network generated from sequences produced by the population genetics simulation scenario, in which individuals reproduced by asexual (clonal) reproduction. This network is clearly characterized by two distinct clades, separated by long branches.

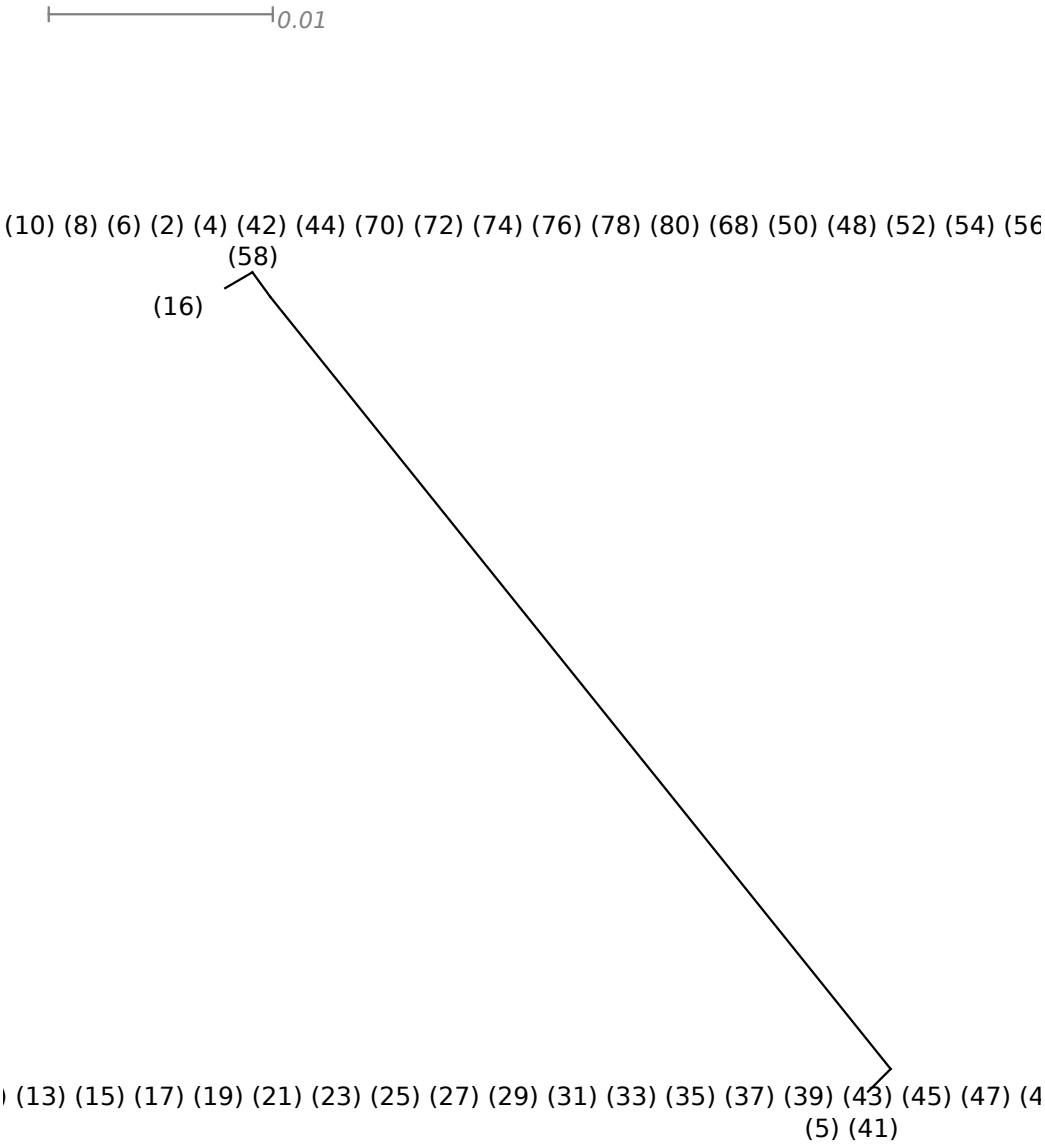


Figure 4.3: Network of simulated allelic pairs, evolved under an asexual reproduction scheme. The first copies of each allelic pair form a clade, and the second copies of each allelic pair form a clade. This is because there is no recombination during gamete formation, as with clonal reproduction, offspring are clones of their parent.

Table 4.1: PHI-Score and Neighbor Similarity Scores of the PCR amplified sequences for three different window sizes.

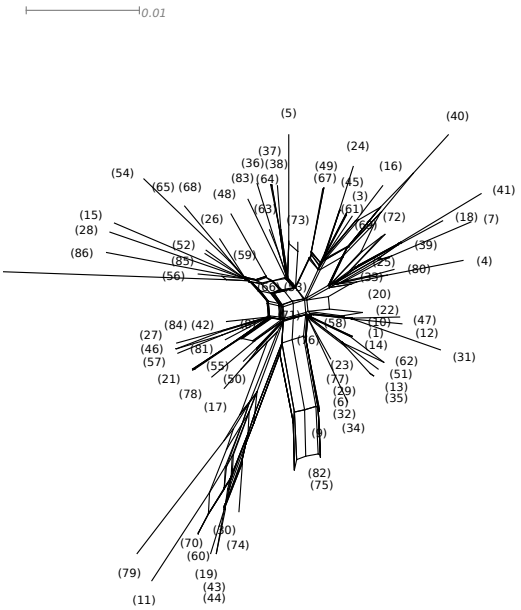
Sequences	Window Size	PHI Score	P-Value	NSS	NSS P-Value
FeABC	100	0.0930	0.0000405	0.81056	0.005
FeABC	50	0.0955	0.0041100	0.81056	0.004
FeABC	10	0.0870	0.0814000	0.81056	0.006
Fcl10	100	0.0930	4.0500000	0.81056	0.005
Fcl10	50	0.0385	0.0184000	0.88306	0.342
Fcl10	10	0.0500	0.2650000	0.88306	0.338

If *F. cylindrus* has a history of asexual reproduction and ancient allelic divergence, then it is expected that the networks calculated for the PCR amplified sequences of the ABC Iron Transporter and the Large Ribosomal Subunit will have a similar structure to that of the network in Figure 4.3.

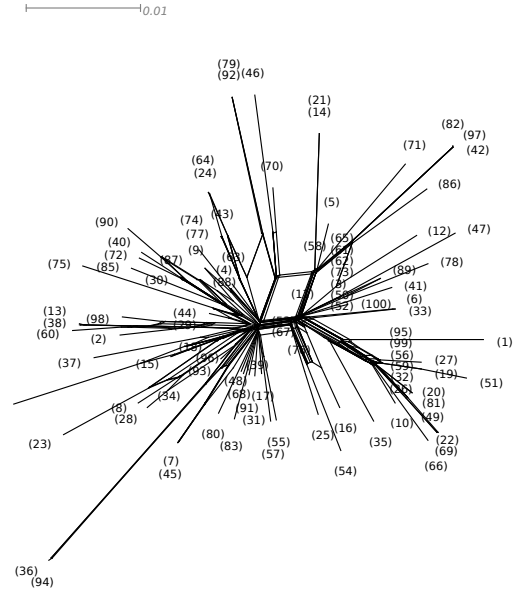
Panels a and b in Figure 4.4 show the phylogenetic networks calculated for the PCR amplified sequences of the ABC Iron Transporter (a), and the Large Ribosomal Subunit (b). These two networks are clearly different qualitatively to the kind of network in Figure 4.3 that would be expected if *F. cylindrus* had a history of asexual reproduction without meiotic recombination. They do not show a clear partition between two clades or clusters, instead they have average branch lengths of around 0.1, and contain around 255 splits.

Panel a of Figure 4.5, demonstrates the effect of increasing or decreasing θ in population genetic simulations, on the resulting sequences, and thus the networks produced: The average branch lengths in networks, is positively related to the θ parameter set in the simulation.

Figure 4.6 presents this relationship qualitatively with the networks produced by Splitstree. From figures 4.5 and 4.6 it can be seen that the networks best matching the real sequence networks (figure 4.4) in terms of branch lengths, are those produced by simulations where $\theta = 0.066$, which is close to the value which LAMARC has estimated.

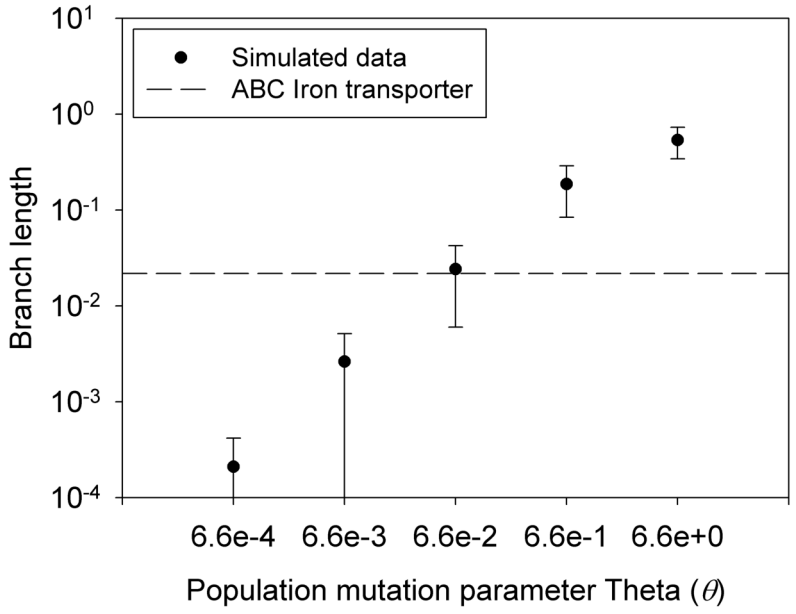


(a) ABC Iron Transporter

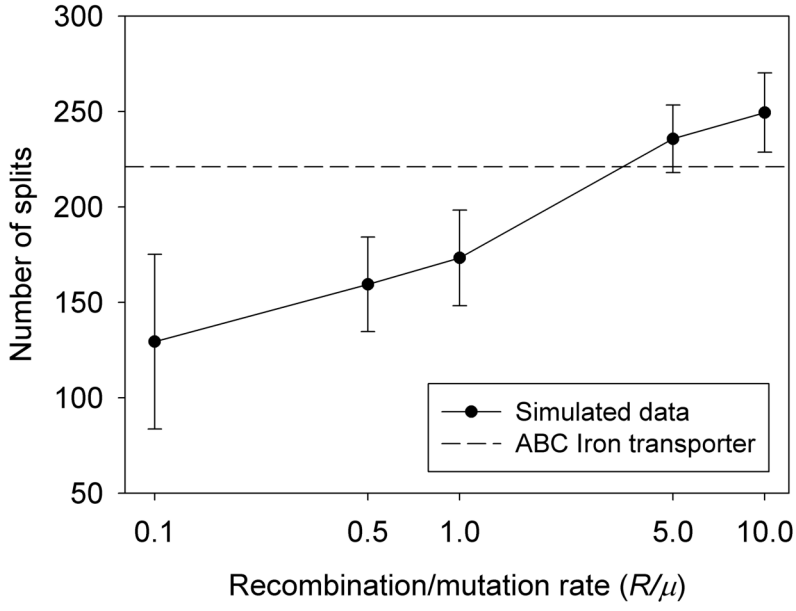


(b) Large Ribosomal Subunit

Figure 4.4: Split Networks of the ABC Iron Transporter and Ribosomal Subunit sequences have average branch lengths close to 10^{-2} and contain 225 splits.

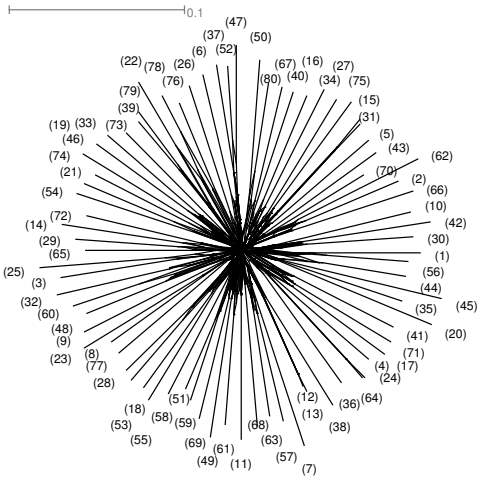


(a) The effect of θ on network branch lengths

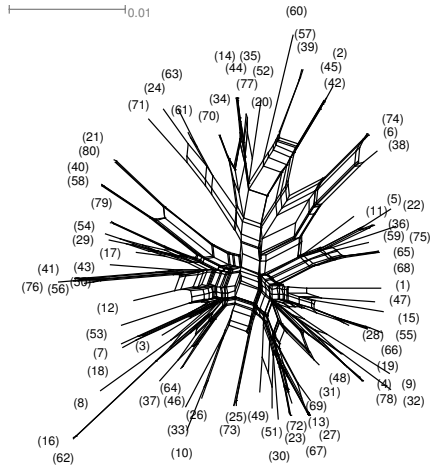


(b) The effect of recombination rate on splits

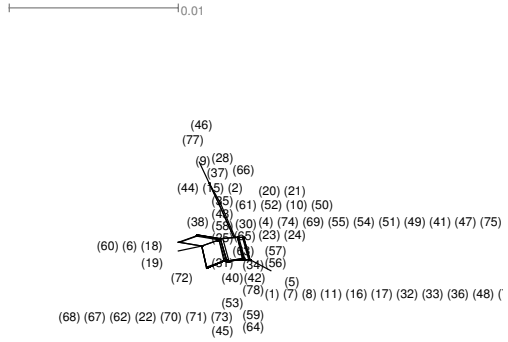
Figure 4.5: Quantifying the branch lengths and number of splits in networks produced from simulations with varying levels of recombination and values of θ. Larger values of θ cause longer branches (a), and higher recombination rates result in more splits (b).



(a) $\theta = 0.66$



(b) $\theta = 0.066$



(c) $\theta = 0.0066$

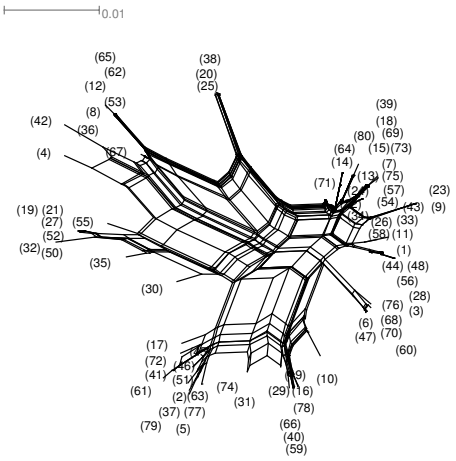
Figure 4.6: Networks computed from simulations with three different values of θ. Larger values of θ result in longer outer branches of networks.

Panel b of figure 4.5 demonstrates the effect of varying the recombination rate relative to the mutation rate in population genetic simulations, on the sequences and networks produced: The number of splits in networks is positively related to the recombination rate, relative to the mutation rate. This relationship is shown qualitatively in the networks drawn in figure 4.7.

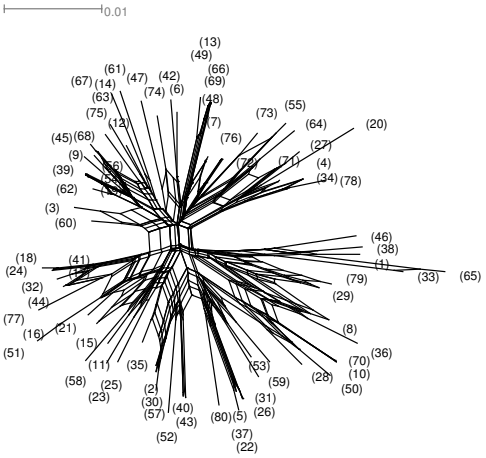
4.4 Discussion

The phylogenetic networks resulting from population genetic simulations support several assumptions we had about how recombination, and population mutation rate (θ) may be inferred from phylogenetic networks. Specifically, (1) the levels of Theta affect the average branch lengths of the networks, and (2) the extent of recombination affects the number of splits in phylogenetic networks. These two assumptions are not controversial: a higher population mutation rate leads to more mutations in a population the same amount of time, and thus would lead to longer branches in any phylogeny or network computed for sequences sampled from the population (Frankham 1996; Hein, Schierup, and Wiuf 2004; Wakeley 2009). Phylogenetic Split Networks (Huson 1998) were conceived of as a way to detect and represent reticulate evolution. Wherever there is a non-tree like structure or loops, recombination may be inferred. The networks resulting from the simulations confirm these assumptions, and so give confidence in any inferences made about the population and evolution of *F. cylindrus* from the networks of the ABC Iron Transporter sequences, and the Large Ribosomal Subunit Sequences.

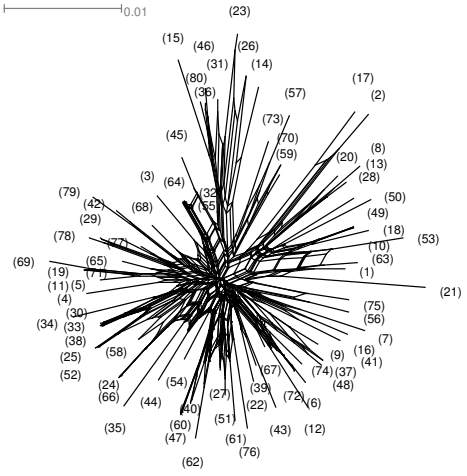
Secondly, from comparisons between the networks of the ABC Iron Transporter sequences, Large Ribosomal Subunit Sequences, and simulated networks, it was concluded that LAMARC (Kuhner 2006) estimate of Θ was a reasonable estimate for the population of *F. cylindrus*. It was



(a) $R/\mu = 1$



(b) $R/\mu = 5$



(c) $R/\mu = 10$

Figure 4.7: Networks computed from simulations with three different levels of recombination, relative to the mutation rate μ. Larger values of R result in more splits in networks.

3224 also concluded that these networks provide evidence of recombination for
3225 the sequences of the ABC Iron Transporter, and in the sequences of the
3226 Large Ribosomal Subunit. Evidence of recombination does not have to
3227 mean that an organism is reproducing sexually, meiotic recombination is
3228 associated with sexual reproduction, but mitotic recombination could also
3229 explain the recombination signal detected in these sequences. However,
3230 whilst mitotic or meiotic recombination may explain the recombination signal
3231 in the sequences, it was concluded that ancient asexuality is not a likely
3232 explanation, because of the lack of similarity of the ABC Iron Transporter,
3233 and the Large Ribosomal Subunit networks, to the networks generated by
3234 simulations of ancient asexual evolution.

3235 4.4.1 Sex and the diatom reproductive cycle

3236 Even though sexual reproduction has not been observed in the lab cultures
3237 of *F. cylindrus*, this diatom does not appear to be an ancient asexual. This
3238 might not be surprising given what is already known about Diatom biology
3239 and sexual reproduction. The typical cell cycle of Diatoms is diplontic i.e.
3240 the vegetative cells are diploid, and the haploid gametes are short lived
3241 (Chepurnov et al. 2004).

3242 The Diatom life cycle features two key phases which may be summarized
3243 by the following:

3244 The first phase is a long vegetative phase; this phase can last for months
3245 or years. During this phase, vegetative cells divide by mitosis, gradually
3246 becoming smaller. The cell size decrease during the vegetative phase of
3247 the diatom life cycle is due to the shape and structure of Diatom cell walls
3248 and the division pattern of the Diatoms. The cell wall is made of sillicated
3249 components, which together are termed the frustule. The frustule is made
3250 of two overlapping halves or thecae (Chepurnov et al. 2004; Davidovich and
3251 Bates 1998; Poulickova 2008). These thecae are not the same size, the

larger of the two thecae is called the epitheca, and the smaller of the two is called the hypotheca. When mitosis occurs, cytokinesis splits the diatom where the two thecae overlap. The two resulting daughter cells inherit one of the parent cells two thecae as its own epitheca, and they grow their own hypotheca (Chepurnov et al. 2004). Since one of the daughter cells inherits a hypotheca as it epitheca, it will be smaller in size to its parent cell. Thus the average cell size of a population of diatoms decreases as mitotic cell division occurs.

The second phase is shorter, and includes sexual reproduction and the production of new vegetative cells, restoring the cell size (Chepurnov et al. 2004). Production of gametes during the sexual reproduction phase has been demonstrated to occur by classical meiosis in many Diatom species. Diatoms restore their cell size through the production of auxospores, which result from sexual reproduction (Davidovich and Bates 1998). During auxosporulation, recombination and cell size restitution occurs: gametes fuse to form the auxospore, which expands and a new cell is produced within. The cell walls of the gamete producing cells are lost, and so the auxospore must then form the shape of the vegetative cells de novo (Chepurnov et al. 2004). If a population of Diatom cells did not undergo sexual reproduction to produce the auxospores to restore their cell size, the population would gradually decrease in cell size until they become critically small. At this point the population would die, and this has been observed in experimental cultures. Diatom cells can only become sexualized when they are sufficiently small, but they may also not be able to become sexualized if they become too small or hit the critical cell size before they die (Chepurnov et al. 2004; Davidovich and Bates 1998; Poulickova 2008). The maximum size of initial diatom cells, the maximum and minimum sizes of cells capable of sexual reproduction, and the minimum size before death are strict for each diatom species and are termed cardinal points (Chepurnov et al. 2004).

However, despite the role of sex in the restoration of cell size in diatoms, it is not always necessary for cell size restoration. For some diatom species, asexual auxosporulation is a possibility, presumably it is some secondary modification of a developmental pathway that was sexual, and some species do not even undergo auxosporulation and exist as entirely as asexual populations, and their cell size is restored by vegetative cell enlargement (Chepurnov et al. 2004; Gallagher 1983; Nagai et al. 1995; Sabbe et al. 2004; Werner 1977). Species such as *Caloneis amphisbaena* and *Sellaphora pupula* "lanceolate" have been found to exist in populations of a very limited range of cell size, and this cell size has remained unchanged after many generations of observation (Mann 1989; Mann et al. 2004).

Therefore, whilst sex is a common feature of the diatom life cycle, and is important for cell size restoration in many species, it is not unreasonable to suggest the hypothesis that a diatom like *F. cylindrus* could have evolved asexually for a long period of time. However, the network reconstructions and evidence of recombination demonstrated by this study cast doubt on that hypothesis as an explanation for the diverged alleles.

4.4.1.1 Allelic Divergence in diatoms may be explained by population size

If the ancient asexuality hypothesis is rejected as the explanation of the diverged alleles in *F. cylindrus*, then an alternative explanation of how this diatom evolved diverged and functionally differentiated alleles is desired. These alleles show signatures of positive selection, and they are differentially expressed. The question is; assuming sexual reproduction and recombination, why does recombination not homogenize the sequence variation between two alleles over time?

An alternative hypothesis explaining the adaptive evolution of *F. cylindrus* is a large population size, which would lead to bigger coalescence

times between maternal and paternal loci. In combination with a low recombination rate, this would result in independent adaptive evolution and divergence of the different haplotypes. This is intuitive if one considers a coalescent process back through time of an idealized population, because the coalescent relates genetic diversity to demographic history. In such a process, the probability that any two lineages extant at time t, coalesce in the previous generation $t1$, is the probability that they share a parental DNA sequence. For a diploid population there are $2N_e$ alleles in every generation, assuming a constant population size (Hein, Schierup, and Wiuf 2004). Assuming random mating and neutral evolution, the probability any two alleles coalesce in the previous generation (i.e. they share the same parental sequence) is $1/(2N_e)$. Therefore, the probability those two alleles do not coalesce, is $1(1/(2N_e))$. These probabilities are dependent on the size of the population in question (Wakeley 2009). Larger populations, result in a smaller probability that two alleles coalesce in the previous generation, and a greater probability that they do not. With each successive previous generation, the probability of coalescence is geometrically distributed (Hein, Schierup, and Wiuf 2004; Wakeley 2009). This means that it is the product of coalescence at the generation of interest and the probability of non-coalescence at the preceding generations i.e.

$$P_c(t) = \left(1 - \frac{1}{2N_e}\right)^{t-1}\left(\frac{1}{2N_e}\right) \tag{4.1}$$

From this equation, it can be seen that with larger populations, the probability that two alleles coalesce further back in time is greater i.e. the expected coalescence time between two alleles is larger, therefore alleles are expected to be more diverged.

This explanation is consistent with the estimation of a Θ of 0.066 by the

LAMARC (Kuhner 2006) analysis, which is also supported by the simulations. The population mutation rate Θ, is proportional to the product of the mutation rate and the effective population size and so the value predicted by LAMARC could be the result of a very large population. Furthermore, prior research has been performed to estimate the abundance of *F. cylindrus* in water columns around the Antarctic (Kang and Fryxell 1992). During the summer, numbers of 7.9×10^{10} cells $^{m-2}$ were observed, and during the winter, numbers of 1.1×10^{8} cells $^{m-2}$ were observed. Marginal ice zones are known to be sites with much dynamic activity such as jets, eddies, currents, melting, freezing, and upwelling (Kang and Fryxell 1992). They are also known to be sites of increased phytoplankton biomass and primary productivity, due to their light levels, ice-distribution, and vertical stability.

Therefore, the hypothesis that a large population size explains the levels of diversity is consistent with both population genetic (coalescent) theory, results of this study, as well as the findings of other research. It is also attractive, because of its simplicity. It is much more plausible that a phytoplankton species has very large populations; than it is that the species had abandoned sex as a reproductive strategy: Sex is a common aspect of the diatom life cycle and is often essential for cell size restoration and population survival. Furthermore, as was explored in the Introduction, there is a substantial body of theory explaining why sexual reproduction evolved two become a widespread reproductive strategy, and is advantageous, despite the apparent costs.

4.4.1.2 Study limitations and subsequent FALCON assembly

However, this study has some limitations which should be acknowledged when considering these results. First, whilst evidence of recombination in the form of the splits networks and the presence of incompatible sites is obtained from these sequences, it was not possible to examine any

larger recombination events or blocks as was possible for *Albugo candida* in chapter 3. Indeed, the number of informative sites was too small for the HybridCheck software (which was implemented to analyse large contigs) to effectively run. Secondly, the analyses were only performed on two genes. Whilst it was concluded that the two genes were representative of the larger set of diverged alleles (figure 4.1), we do not know if the PCR primers amplified only maternal and paternal alleles of those genes or if some of the sequences amplified also represent paralogues.

The question of whether the diverged alleles observed in the assembly were truly diverged alleles was resolved for the assembly described in the introduction experimentally. Single haplotyped fosmids were Sanger sequenced by collaborators, providing contiguity information and they were compared with the assembled genomic scaffolds, and an annotated protein set from the diverged regions in the genome. Data from these comparisons revealed a clear separation between allelic pairs and gene duplications based on 100% identity to the haplotyped Sanger sequenced fosmids. Additionally, the nucleotide similarity of the diverged alleles (mean = $97.01 \pm 0.03\%$) is significantly (p-value $< 10^{-09}$) higher than for gene duplicates (mean = $84.07 \pm 0.36\%$). Therefore, whilst it may be that some uncollapsed regions of the assembly could be duplicates, there is high confidence that the allelic pairs identified are indeed diverged alleles and not duplicates.

Since this work has been completed, an assembly has been completed using PacBio long read sequencing technology, which has also supported that true diverged alleles have been identified and that they are not duplicated sequences (although duplicated sequences are indeed present in the genome). The sequencing work and library preparation was completed by the platforms and pipelines team. A 20kb fragment length library was constructed, and a 4kb insert size library was also created. Both libraries were sequences using the PacBio RS2 instrument, using SMRT cells with

the c4P6 chemistry. The 20kb fragment length library yielded 1.37Gb of data, and the 4kb insert size library yielded 3.85Gb of data. The final N50 of read length varied between 8215 to 8898bp for the 20kb fragment length library, and the N50 ranged from from 2558 to 2680bp for the 4kb insert size library.

Assembly was completed by collaborator Pirita Paajanen, who combined the data from the SMRT cells and filtered the shortest reads, yielding 3.8Gb of data which gave 63x coverage. Assembly was completed using the diploid aware PacBio assembler, Falcon 0.3.0. The output of the Falcon assembler was divided into two parts. The haploid assembly resulted in primary contigs from which a genome size of 59.7Mb was deduced. However, the assembler also produced alternate contigs which were the result of the assembler being unable to decide between two possible routes through the genome graph the genome. Such 'bubbles' in the genome graph represent diverged haplotypes, containing diverged alleles.

The haplotype divergence differed between chromosomes: The longest chromosome 000000F had only one alternate contig with a length of 6047bp. In contrast, contig 000002F was 1246645bp long and had 14 associated alternative contigs, of a total sequence length of 633764bp. For each of the 14 alternative contigs of chromosome 000002F, I extracted and aligned the two haplotype sequences using the pairwise alignment algorithm available in the Bio.jl software package (https://biojulia.github.io/Bio.jl), using an EDNA scoring matrix. Once aligned, a non-overlapping sliding window was moved across the sequences, and the p-distance between the sequences within each window was calculated. For each computation, the width of the sliding window was set as 1% of the width of the pairwise alignment in (bp). The results of this analysis are included as extra information in the appendix, figure A.1. The figure demonstrates different levels of divergence across the diverged haplotype pairs, including the appearance of indels

between some haplotype pairs. Further work will show how the sequences of allelic pairs align to the FALCON assembly, revealing which pairs align to different haplotypes of a FALCON 'bubble' (true allelic pairs), and which pairs align to the same haplotype of a FALCON 'bubble' (potentially gene duplicates).

At the time of writing, multiple population samples of *F. cylindrus* are not available, and so analyses presented here used sequences from cultures, and so further population genetic analyses should be conducted in the future as more data becomes available, for example to assess the population structure of *F. cylindrus* and investigate if gene flow is occurring between subpopulations of *F. cylindrus*.

The fact that the genome assembly contains some duplicates, and that some of the allelic pairs analysed in this study may be found to be duplicates is not problematic for the hypothesis that this Diatom species has adapted through alleleic divergence, as it may be argued allelic divergence could lead to gene duplication and the conditions for the divergence of alleles and the divergence of duplicates overlap: When diverged alleles are maintained in a population due to heterozygote advantage, duplications may rapidly spread through the population, causing an individual to act as a genetic heterozygote yet still breed true. Proulx and Phillips 2006 argued that genetic redundancy is the mechanism usually cited as allowing duplicate genes to diverge, but redundancy is present in a diploid before duplication: Dominance creates the same kind of redundancy duplicates have, but for alleles of single copy genes. Therefore mode of inheritance is the thing then which most distinguishes duplicates from single copy genes: Segregation prevents the fixed inheritance of alternative allelic variants at a single locus (Proulx and Phillips 2006). In other words, heterozygotes at one locus are broken up by segregation during sexual reproduction, whereas duplicate loci in an individual can carry copies of alternate alleles

at different loci. Their results show that fitness relationships that allow divergent alleles to evolve at one locus overlap significantly with those that allow the divergence of previously duplicated genes at two different loci (Proulx and Phillips 2006).

4.4.1.3 Conclusion

The genome of the polar diatom *Fragilariopsis cylindrus* contains diverged alleles that are differentially expressed in different environmental conditions. Evidence of recombination was found which contradicts the ancient asexuality hypothesis explaining how these diverged alleles may have evolved. An alternative, competing hypothesis is proposed, supported by the evidence presented, that a large population size has allowed diversifying selection to differentiate the alleles of genes despite the homogenizing effect of recombination. Additional population samples, and analysis of larger contigs made possible by improved genome assembly for recombination, will help answer the question of how *F. cylindrus* has evolved this remarkable strategy to cope with varying environmental conditions.

CHAPTER 5

General Conclusion

5.0.1 Summary and Conclusions

In this thesis, work focused on how recombination facilitates the adaptive evolution of a plant pathogen and a polar marine diatom. Both of these organisms were of evolutionary interest due to aspects of their lifestyles and/or physiology: The plant pathogen *Albugo candida* was of interest because whilst it was an obligate biotroph, it has a very large host range, and the diatom *Fragilariopsis cylindrus* was of interest because the genome sequencing project and differential expression experiments revealed genes with diverged alleles that were differentially expressed in different environmental conditions.

Recombination is important for the formation of novel genotypes, haplotypes and alleles, therefore is plays a key role in adaptive evolution (Grauer and Li 2000). Recombination separates deleterious mutations from their genomic background, in combination with purifying selection this reduces the mutational load (Lynch and Gabriel 1990). Recombination also brings beneficial mutations from separate lineages into one individual or lineage. However, recombination also plays a fundamental role in the repair of damaged DNA, when homologous recombination replaces a damaged DNA strand with its intact counterpart, and it was likely this function of

recombination that was important in early prokaryotic life and evolution (Cavalier-Smith 2002). With respect to adaptive evolution, however, the principal consequence of recombination is that it generates novel combinations of nucleotides, which in turns allows for selection to act a much finer scale, i.e. at the level of nucleotides rather than the entire genome.

The potential of recombination to generate novel allelic combinations is important for host and pathogens which are engaged in an evolutionary arms race to adapt and counter adapt to each others molecular mechanisms of pathogenicity or immunity. The red queen hypothesis explains the advantage of sexual reproduction in such terms. The variability generated by sexual reproduction (and meiotic recombination) results in genetically unique offspring, which permits a faster response to selection (Paterson et al. 2010). As a result sexually reproducing species are able to improve their genotype in changing conditions. Co-evolutionary interactions between host and parasite select for sexual reproduction in hosts in order to reduce the risk of infection. Oscillations in genotype frequencies are observed between parasites and hosts in an antagonistic co-evolutionary way without necessitating changes to the phenotype, and in host-parasite co-evolution systems with multiple hosts, Red Queen dynamics may affect which host and parasite types become common (or rare) (Charlesworth and Charlesworth 2010).

It was hypothesized that the *Albugo candida* species was composed of several host-specialised races, each locked in an evolutionary arms race with their specific host. Such a race with a specific host would lead to further divergence and possibly speciation of the races. However, *Albugo* is known to be able to suppress non-host resistance. Infections of *Albugo sp.* could suppress the runaway cell death phenotypes of plants, allowing formerly avirulent strains of downy mildew to infect (Cooper et al. 2008). Assuming that this ability extended to other non-host species, *Albugo*

may be modeled as a 'microbial hub': taxa that are integral and highly connected to the network of a hosts microbial community. Such hubs may affect community compositions through microbe-microbe interactions or, as seems to be the case with *Albugo*, suppression of host defense responses (Agler et al. 2016). Therefore, non-host immune suppression would enable host-specific races of *Albugo candida* to overcome the ever increasing barrier to gene flow that specialisation imposes, and sexual reproduction between races, followed by introgression by back-crossing, would permit the generation of a range of novel genotypes. Consequently the species could evolve its wide host range.

To assess this hypothesis it was necessary to scan the genome of *Albugo candida* isolates to identify recombinant regions. Furthermore, to distinguish such regions as recombinant and not the result of incomplete lineage sorting due to rapid divergence, the regions identified needed to be tested for significance and the coalescence times estimated.

Scans of the genomes for recombination revealed a highly recombined mosaic genome, and therefore a rapid coalescence estimation method for all of the recombination blocks was desired, in addition to a method of plotting which effectively demonstrated the high degree of mosaic-ism in the *A. candida* genome. Therefore, rapid detection and dating of recombination blocks was implemented, and the software package HybridCheck was created and tested using simulated data as in chapter 2. HybridCheck was also tested for consistency with RDP analyses of *A. candida*, which identified recombination, and BEAST estimates of coalescence times for a subset of the identified recombination regions (chapter 3). The evidence presented in chapter 3 confirmed the model of *Albugo candida evolution*: Isolation, divergence and specialisation of races generates repertoires of effectors for a specific race. Those adapted repertoires are then brought together when two races hybridize. The result if the generation of novel

repertoires of novel combinations of these effectors. Specific avirulence effectors that trigger host immunity may be lost through segregation and through loss of heterozygosity (Lamour et al. 2012; McMullan et al. 2015). Hybrids, with new combinations of effectors, and having lost effectors which impeded their colonisation of other hosts previously, may expand their geographical range and population size clonally. Some of these hybrids may be able to colonise new hosts, expanding the host range.

The genome assembly project of *F. cylindrus* revealed that the genome contained 21,066 predicted protein-encoding genes, 6,071 genes were represented by diverged alleles, and each pair of diverged alleles had both coding and non-coding regions, and were up to 6% polymorphic in the non-coding regions. Furthermore, differential expression experiments and RNA-Sequencing suggested that 40% of the non-collapsed, diverged allelic pairs showed a 4 fold unequal bi-allelic expression (Mock et al. 2017).

Alleles showing the strong unequal bi-allelic expression were found to have an elevated rate of non-synonymous mutations, which suggests significant positive / adaptive selection and evolution of these allelic pairs (Mock et al. 2017). It was concluded therefore, that positive selection has been a driving force in the evolution of these diverged alleles and hence the adaptation of this diatom to the environmental conditions it faces.

An evolutionary explanation was hypothesized: The alleles of an allelic pair could diverge as a result of positive selection because there was a long history of asexual reproduction in the organism, and hence an absence of recombination acting as a homogenizing force between alleles.

However, results from recombination detection analysis, and phylogenetic network construction of PCR amplified sequences from DNA extracted from *F. cylindrus* cultures conflicted with results of the same analyses performed with DNA sequences obtained by population genetics individual based simulations of ancient asexuality. Indeed the results for *F. cylindrus*

were more consistent with those of simulations of a scenarios of sexual reproduction, and a large Θ value. This result suggests an alternative competing hypothesis, that very large effective population sizes could have led to the divergence of the alleles in each allelic pair as a result of positive selection, in the face of the homogenizing influence of recombination through sexual reproduction.

5.0.2 Impact and potential future directions

5.0.2.1 *Albugo candida*

A paper describing the extent of the introgression identified within the *A. candida* genome was published in eLife (McMullan et al. 2015). According to Google Scholar, the study has been cited 11 times at time of writing. Citations include reviews of the role of hybridisation and introgression in the adaptive evolution and emergence of new fungal and filamentous plant pathogen strains (Depotter et al. 2016; Dong, Raffaele, and Kamoun 2015; Stukenbrock 2016), research demonstrating the role of recombination in the evolution of the Rp1 resistance genes in grasses (Jouet, McMullan, and Oosterhout 2015), and a study presenting evidence that for *Coleosporium ipomoeae*, any genotypes can infect multiple hosts from non-local communities, but only are highly host specific when tested on hosts from local communities, calling into question theoretical results of single-pathogen single-host studies which suggest that selection favours genotypes with a broad host range (Chappell and Rausher 2016). Following the 2015 eLife paper, Belhaj et al. 2015 published a more extreme example of the ability of *Albugo spp.* to suppress the host immune system. They found that *Phytophthora infestans*, which is typically a potato and tomato specialist pathogen, was capable of infecting the plant model organism *Arabidopsis thaliana* when *Albugo laibachii* has also colonized the plant. The nature of

the *P. infestans* infection was similar to that of an *Albugo laibachii* infection: Transcription profiling of *P. infestans* infections revealed a significant overlap between the sets of secreted proteins of *P. infestans* during infection of *Arabidopsis thaliana* and during infections of potato. This suggests there is similar gene expression dynamics on the two species, and it raises the question. Is gene flow between two different Oomycete species possible? And could this contribute to adaptive evolution of these pathogens.

It is well established that *Albugo* suppresses hon-host immunity in hosts it infects, and as a result of work presented in this thesis it was concluded that this lowers barriers to gene flow and permits introgression, facilitating the generation of novel pathogen haplotypes and enabling *Albugo candida* to evolve a wide host range. However, this model of *Albugo candida* evolution raised a conceptual problem: This phenomenon appears to extend to other pathogen species that were not *Albugo spp.* (Belhaj et al. 2015), and therefore *Albugo spp.* may act as a microbial hub as previously noted. If this is the case, how is it that *Albugo spp.* (obligate biotrophs with a vital dependence on the host) can compete in this limited niche, whilst at the same time enable non-host colonization for other pathogen species who are then presumably competitors for the same resource. An answer to this problem was provided by a paper from Ruhe et al. 2016. Shotgun proteomics was completed of the apoplastic fluid of samples of lab-grown *Arabidopsis thaliana* that were infected with *Albugo spp.*, and samples which were uninfected. Work was repeated for wild-grown *Arabidopsis thaliana* and they found that whilst both lab-grown and wild-grown *Arabidopsis thaliana* supported extensive *Albugo* colonization (Ruhe et al. 2016). However, no or low levels of defense-related proteins were detected in lab samples, but regardless of *Albugo spp.* infection status, wild plants showed a broad spectrum of defense-related proteins at high abundances and lab-grown plants did not. These results suggest that *Albugo spp.*

do not strongly affect immune responses and leave distinct branches of the immune signaling network intact (Ruhe et al. 2016). This suggests that the pathogens of the *Albugo* genus, including *Albugo candida* in the wild are fine tuned to avoid triggering strong host defense reactions, but also to avoid a broad-spectrum host defense suppression, thus allowing them to avoid competition from other species growing in the same niche (Ruhe et al. 2016). Since races of *Albugo candida* are members of the same species, they may still colonize the same host plant at the same time, allowing introgression to occur (explaining the introgression signal observed), but other more distantly related competing pathogens may be excluded by this precise host immunity manipulation observed by Ruhe et al. 2016, and so may not get to compete with *Albugo spp.*. However this experiment only examined *Arabidopsis thaliana* as a host, and crops grown in monoculture are often uniform and subject to artificially maintained conditions and treatments, and this may be considered analogous to plants grown in laboratory conditions. So it is uncertain whether in monoculture environments *Albugo spp.* manipulate their host immune systems subtlety and precisely, thus avoiding colonization of competition, or whether as with lab-grown *Arabidopsis thaliana* they do significantly affect the secretome of the host, allowing competitors to colonize.

In the future, additional study of more strains and population samples of *Albugo candida* is desirable, since the study presented in this thesis only examined the genomes of three 'races', and more samples might increase the number of *Albugo candida* races we can analyse. Future potential work also includes disentangling the true branching order of *Albugo candida* races, and improving the detection and dating methods used to analyse *Albugo candida* genomes (see below).

5.0.2.2 HybridCheck

The HybridCheck software package was initially created out of a need specific to the *A. candida* project in chapter 3. Following the *A. candida* project, the HybridCheck software was published in a short software note in Molecular Ecology Resources (Ward and Oosterhout 2016), and other groups across the Norwich Research Park became interested in using it with their own study systems.

In particular, researchers at Norwich Medical School working on *Cryptosporidium* used HybridCheck to perform chronological assessment of recombination events identified in the genomes of three trains of *C. parvum* (IIaA15G2R1, IIcA5G3j, IIcA5G3a), and a single *C. hominis* (IbA10G2) GP60 sub-type strain (Nader 0). They found 104 unique recombination events, and a skewed distribution of recombination events across chromosomes. More recombination events were identified on chromosome 6, and a greater number of events was observed for *C. parvum anthroponosum* sub-type IIcA5G3a than for any other strain. More than 90% of all recombination events occurred proximal to loci suspected to drive virulence or play a major role in host-parasite interactions in human cryptosporidiosis. Therefore it appears that in this pathogen too, recombination is an important force, generating novel gene combinations and driving the adaptive evolution of a pathogen to its host (Nader 0). The estimated divergence dates calculated in their study provide the first chronological description for genetic introgression between human-infective *Cryptosporidium spp.*. HybridCheck analyses revealed a chromosome-wide consensus that places a majority of introgression events between zoonotic (IIaA15G2R1 and IIcA5G3j) and anthroponotic (IIcA5G3a) *C. parvum* sub-type strains at approximately 10-15 thousand generations ago, while genetic introgression (or recombination) between the two more closely related zoonotic strains appears to be more

recent (between approximately 3 to 5 thousand generations ago) (Nader 0).

Based on infectivity studies in healthy adult volunteers, the average generation time within a host is 14.8 hours, and assuming a steady rate of transmission within host populations, they derived a minimum estimate of the recombination events of around 5.9 (zoonotic vs. zoonotic *C. parvum*), 17.6 (zoonotic vs. anthroponotic *C. parvum*), and 176.7 (*C. hominis* vs. *C. parvum*) years ago (Nader 0). In other words, they estimate that the evolutionary split between the two primary human-infective species appears to have occurred at the turn of the second industrial revolution, around 1840 (Nader 0).

Whilst this result is putative and needs validation with other dating methods before publication submission, it is a clear demonstration of the utility of HybridCheck for researchers in estimating coalescence times rapidly, across many recombination affected genomic regions.

Future directions for work involving HybridCheck include its continued use in other organisms. For example HybridCheck is already being used to generate preliminary results for population genomic data for mice (*Mus spp.*), being generated at the Earlham Institute, with the aim of confirming hypotheses of genetic isolation between species, and identifying potential introgressions between populations. Future work involving HybridCheck may also involve programmatic work. Bioinformatics methods and the detection of introgression is an active area of research, and more algorithms and methods will likely be created in the future. Therefore, HybridCheck would have greater utility as a provider of different methods for the detection and dating of recombinant and introgressed regions, that are able to work on multiple different data sources or formats. As a programming problem, such software code might be best implemented, using multiple dispatch, to make it more easily maintained, and more easily used. Multiple dispatch is

a feature of some programming languages in which a function (sometimes called a method) can be dynamically dispatched based on the type of more than one of its arguments. This thesis author has already co-founded, develops, and maintains a new bioinformatics infrastructure and community called BioJulia, based around a modern new programming language for scientists and technical programmers, called Julia. The language is high-level, implements a flexible type and multiple dispatch system, and can achieve speeds matching those of compiled software written in the C language, with less lines of code. These features make it ideal for the kind of rapid and flexible development that Bioinformaticians often do, and should development of HybridCheck continue towards this goal, the framework already has many high performance code modules and features that a BioJulia port of HybridCheck could take advantage of.

In the near future, approaches to recombination detection may also change. Currently, HybridCheck and other methods typically analyze DNA or protein sequences and identify regions that are phylogenetically incongruent i.e. where computed phylogenetic topologies change or there is a change-point in computed genetic distances. After the identification of these regions, it may be assumed they are recombination, or incomplete lineage sorting, and subsequent analyses, such as the dating method in HybridCheck, may be employed to try to distinguish whether the cause is recombination or incomplete lineage sorting. The cause may also be assumed based on rates of speciation or population size; incomplete lineage sorting is more likely when either of the two are high. However, as described in chapter 2, there are problems with this approach which leave room for future improvement.

For example, recombination blocks can become fragmented by accumulation of subsequent mutations following the recombination event. Consequently, older recombination blocks tend to be smaller, when they

are actually larger. Thus, not all mutations are accounted for, resulting in an underestimate of the divergence time particularly for old recombination events/regions of incomplete lineage sorting.

Furthermore, some methods of resolving introgression from incomplete lineage sorting require knowledge of branching orders, and sometimes these are unknown, and sometimes this is even because of the influence of introgression or incomplete lineage sorting. To solve this issue for the malaria parasite, Fontaine et al. 2015 obtained the correct species branching order of the *An. gambiae* complex and two Pyretophorus out-group species. To do this in the face of introgression and incomplete lineage sorting they used 50kb non-overlapping windows across a genome alignment and computed phylogeneies for each window. At least 85 tree topologies were observed. When these were sorted according to chromosome arm and their relative frequency, the most commonly observed topology for the X chromosome was highly discordant with the most commonly observed topology for the autosomes. They then grouped these phylogenetic toplogies, into three distinct topology categories based on the relative phylogenetic positions of two species: *An. arabiensis* and *An. quadriannulatus*, and they observed the topology category most commonly observed on the X chromosome, was not the same as for the autosomes. Dating the internal nodes of phylogeneies for each topology category allowed them to distinguish which category of topology best represented the true branching order, and which represented topologies that were caused by introgression. Given that almost all of the autosome was represented by a topology category that is affected by introgression and linkage disequilibrium, traditional phylogenetic methods for resolving a species level topology, which typically invoke some majority rule, would certainly have resulted in the incorrect answer.

The method utilized in their work will be of great benefit to researchers

studying complicated genomes where introgression, and incomplete lineage sorting, are prevalent. A likely future direction for the development of HybridCheck will be to take these methodological ideas and implement tools that make it trivial for researchers to decompose the gene trees computed across a genome, identify topological categories from those trees, and organize them, before analyzing the divergence times of the phylogenies in each topological category. In the future HybridCheck should make it simple to perform such an analysis along with other methods such as Patterson's D, f_d, and tests to distinguish introgression from incomplete lineage sorting. It should make it trivial to compile such multiple lines of evidence into a more complete picture of introgression, incomplete lineage sorting, and linkage, across genomes.

5.0.2.3 *F. cylindrus*

The study of *F. cylindrus* is in preparation to be submitted to the journal Nature this year. As such it is not possible to describe the impact in terms of a number of citations, or who has cited it and why at this time. However, as stated in discussion of chapter 4, reviewer comments led to further sequencing with PacBio technology, which resulted in confirmation that we had obtained strong evidence of diverged alleles. Furthermore, it is known that at time of writing, that unpublished data and correspondence from a colleague and co-author of the paper, Chris Bowler (perscom), that similar evidence of diverged alleles and differential expression has been found in another diatom species that his group study. Therefore, it could be that the data presented in this thesis and in the paper, are the first evidence of a common phenomenon and mechanism of adaptation in this group of organisms. Future work on this topic has already been described in the discussion of chapter 4: Imminent future work will show how the sequences of allelic pairs previously identified align to the new FALCON assembly.

This will reveal which pairs align to different haplotypes of a FALCON 'bubble' (true allelic pairs), and which pairs align to the same haplotype of a FALCON 'bubble' (potentially gene duplicates). Currently, multiple population samples of *F. cylindrus* are not available, and so analyses presented here used sequences from cultures, and so further population genetic analyses should be conducted in the future as more data becomes available, for example to assess the population structure of *F. cylindrus* and investigate if gene flow is occurring between subpopulations of *F. cylindrus*.

In conclusion, detecting and understanding how recombination is affecting the genomes is critical to understanding how species of interest evolve and adapt to dynamic environments, this thesis has demonstrated how recombination appears to have influenced the evolution and adaptation of two different eukaryotic micro-organisms. Future work will expand on the bioinformatics methodological techniques implemented in this thesis, as more and more data becomes available for these two species.

Bibliography

Abbott, R, D Albach, S Ansell, J. W. Arntzen, S. J. E. Baird, N Bierne, J Boughman, A Brelsford, C. A. Buerkle, R Buggs, R. K. Butlin, U Dieckmann, F Eroukhmanoff, A Grill, S. H. Cahan, J. S. Hermansen, G Hewitt, A. G. Hudson, C Jiggins, J Jones, B Keller, T Marczewski, J Mallet, P Martinez-Rodriguez, M Möst, S Mullen, R Nichols, A. W. Nolte, C Parisod, K Pfennig, A. M. Rice, M. G. Ritchie, B Seifert, C. M. Smadja, R Stelkens, J. M. Szymura, R Väinölä, J. B. W. Wolf, and D Zinner (2013). "Hybridization and speciation." In: *Journal of Evolutionary Biology* 26(2), pp. 229–46.

Adams, K. L. and J. F. Wendel (2005). "Allele-specific, bidirectional silencing of an alcohol dehydrogenase gene in different organs of interspecific diploid cotton hybrids". In: *Genetics* 171(4), pp. 2139–2142.

Agler, M. T., J. Ruhe, S. Kroll, C. Morhenn, S. T. Kim, D. Weigel, and E. M. Kemen (2016). "Microbial hub taxa link host and abiotic factors to plant microbiome variation". In: *PLoS Biology* 14(1), pp. 1–31.

Agrawal, A. F. (2001). "Sexual selection and the maintenance of sexual reproduction." In: *Nature* 411(6838), pp. 692–695.

Alberts, Johnson, and Lewis (2002). "Recombination". In: *Molecular Biology of the Cell*. 4th editio. New York: Garland Science. Chap. General Re.

Ardlie, K, S. N. Liu-Cordero, M. A. Eberle, M Daly, J Barrett, E Winchester, E. S. Lander, and L Kruglyak (2001). "Lower-than-expected linkage disequilibrium between tightly linked markers in humans suggests a role

for gene conversion." In: *American Journal of Human Genetics* 69(3), pp. 582–9.

Armbrust, E. V. (2009). "The life of diatoms in the world's oceans." In: *Nature* 459, pp. 185–192.

Arnheim, N., P. Calabrese, and M. Nordborg (2003). "Hot and cold spots of recombination in the human genome: the reason we should find them and how this can be achieved." In: *American Journal of Human Genetics* 73(1), pp. 5–16.

Arrigo, K. R., D. K. Perovich, R. S. Pickart, Z. W. Brown, G. L. V. Dijken, K. E. Lowry, M. M. Mills, M. A. Palmer, W. M. Balch, F. Bahr, N. R. Bates, C. Benitez-nelson, B. Bowler, E. Brownlee, J. K. Ehn, K. E. Frey, R. Garley, S. R. Laney, L. Lubelczyk, J. Mathis, A. Matsuoka, B. G. Mitchell, G. W. K. Moore, E. Ortega-retuerta, S. Pal, C. M. Polashenski, R. A. Reynolds, B. Schieber, H. M. Sosik, M. Stephens, and J. H. Swift (2012). "Massive phytoplankton blooms under arctic sea ice". In: *Science* 336(6087), p. 2012.

Assmy, P., V. Smetacek, M. Montresor, C. Klaas, J. Henjes, V. H. Strass, J. M. Arrieta, U. Bathmann, G. M. Berg, E. Breitbarth, B. Cisewski, L. Friedrichs, N. Fuchs, G. J. Herndl, S. Jansen, S. Kragefsky, M. Latasa, I. Peeken, R. Rottgers, R. Scharek, S. E. Schuller, S. Steigenberger, A. Webb, and D. Wolf-Gladrow (2013). "Thick-shelled, grazer-protected diatoms decouple ocean carbon and silicon cycles in the iron-limited Antarctic Circumpolar Current". In: *Proceedings of the National Academy of Sciences* 110(51), pp. 20633–20638.

Auton, A. and G. A. T. McVean (2007). "Recombination rate estimation in the presence of hotspots". In: *Genome Research* 17, pp. 1219–1227.

Baack, E. J. and L. H. Rieseberg (2007). "A genomic view of introgression and hybrid speciation." In: *Current Opinion in Genetics and Development* 17(6), pp. 513–8.

3866 Baack, E. J., K. D. Whitney, and L. H. Rieseberg (2005). “Hybridization and
3867 genome size evolution: Timing and magnitude of nuclear DNA content
3868 increases in *Helianthus* homoploid hybrid species”. In: *New Phytologist*
3869 167(2), pp. 623–630.
3870 Baker, M. (2012). “De novo genome assembly: what every biologist should
3871 know”. In: *Nature Methods* 9(4), pp. 333–337.
3872 Baranwal, V. K., V. Mikkilineni, U. B. Zehr, A. K. Tyagi, and S. Kapoor
3873 (2012). “Heterosis: emerging ideas about hybrid vigour”. In: *Journal of*
3874 *Experimental Botany* 63(2), pp. 695–709.
3875 Barrick, J. E. and R. E. Lenski (2013). “Genome dynamics during experi-
3876 mental evolution.” In: *Nature Reviews Genetics* 14(12), pp. 827–39.
3877 Barrick, J. E., D. S. Yu, S. H. Yoon, H. Jeong, T. K. Oh, D. Schneider,
3878 R. E. Lenski, and J. F. Kim (2009). “Genome evolution and adaptation
3879 in a long-term experiment with *Escherichia coli*.” In: *Nature* 461(7268),
3880 pp. 1243–7.
3881 Barton, N. H., D. E. G Briggs, J. A. Eisen, D. B. Goldstein, and N. H. Patel
3882 (2007). *Evolution*. Cold Springs Harbor, NY: Cold Springs Harbor Press,
3883 p. 833.
3884 Bayer-Giraldi, M., I. Weikusat, H. Besir, and G. Dieckmann (2011). “Char-
3885 acterization of an antifreeze protein from the polar diatom *Fragilariopsis*
3886 *cylindrus* and its relevance in sea ice”. In: *Cryobiology* 63(3), pp. 210–
3887 219.
3888 Bazykin, A. D. (1969). “Hypothetical mechanism of speciaton”. In: *Evolution*
3889 23(4), pp. 685–687.
3890 Beck, S. and J. Trowsdale (2000). “The human Major Histocompatibility
3891 Complex: Lessons from the DNA sequence”. In: *Annual Review of*
3892 *Genomics and Human Genetics* 1, pp. 117–137.
3893 Belhaj, K, L. M. Cano, C Prince D, A Kemen, K Yoshida, Y. F. Dagdas, G. J.
3894 Ehterington, H Schoonbeek, H. P. van Esse, J. D. G. Jones, S Kamoun,

and S Schornack (2015). "*Arabidopsis* late blight: Infection of a nonhost plant by *Albugo laibachii* enables full colonization by *Phytophthora infestans*". In: *bioRxiv* (June), pp. 1–30.

Bernstein, H., C. Bernstein, and R. E. Michod (2011). "Meiosis as an evolutionary adaptation for DNA repair." In: *DNA repair*. Ed. by Inna Kruman. Intech. Chap. 19, pp. 357–382.

Bernstein, H, F. A. Hopf, and R. E. Michod (1987). "The molecular basis of the evolution of sex". In: *Advances in Genetics* 24, pp. 323–370.

Bernstein, H. (1985). "Genetic damage, mutation, and the evolution of sex". In: *Science* 229, pp. 1277–1281.

Bernstein, H., H. C. Byerly, F. A. Hopf, and R. E. Michod (1984). "Origin of sex". In: *Journal of Theoretical Biology* 110(3), pp. 323–351.

Biémont, C and C. Vieira (2006). "Junk DNA as an evolutionary force". In: *Nature Genetics* 43, pp. 521–524.

Biga, M. L. B. (1955). "Review of the species of the genus *Albugo* based on the morphology of the conidia." In: *Sydowia* 9, pp. 339–358.

Bodmer, W. F. and L. L. Cavalli-Sforza (1968). "A migration matrix model for the study of random genetic drift." In: *Genetics* 59(4), pp. 565–592.

Boni, M. F., D. Posada, and M. W. Feldman (2007). "An exact nonparametric method for inferring mosaic structure in sequence triplets". In: *Genetics* 176(2), pp. 1035–1047.

Bouckaert, R., J. Heled, D. Kühnert, T. Vaughan, C.-H. Wu, D. Xie, M. A. Suchard, A. Rambaut, and A. J. Drummond (2014). "BEAST 2: A software platform for bayesian evolutionary analysis". In: *PLoS Computational Biology* 10(4), e1003537.

Boutemy, L. S., S. R. F. King, J. Win, R. K. Hughes, T. a. Clarke, T. M. a. Blumenschein, S. Kamoun, and M. J. Banfield (2011). "Structures of Phytophthora RXLR effector proteins: a conserved but adaptable fold

underpins functional diversity." In: *The Journal of Biological Chemistry* 286(41), pp. 35834–42.

Bowler, C., A. Vardi, and A. E. Allen (2010). "Oceanographic and biogeochemical insights from diatom genomes." In: *Annual Review of Marine Science* 2, pp. 333–365.

Brent, R. (1973). *Algorithms for minimization without derivatives*. Mineola, New York: Dover Publications, p. 206.

Bromham, L. and D. Penny (2003). "The modern molecular clock." In: *Nature Reviews Genetics* 4, pp. 216–224.

Brown, J. K. M. (2003). "A cost of disease resistance: Paradigm or peculiarity?" In: *Trends in Genetics* 19(12), pp. 667–671.

Brown, R. P. and Z. Yang (2011). "Rate variation and estimation of divergence times using strict and relaxed clocks". In: *BMC Evolutionary Biology* 11(1), p. 271.

Bruen, T. C., H. Philippe, and D. Bryant (2006). "A simple and robust statistical test for detecting the presence of recombination". In: *Genetics* 172(April), pp. 2665–2681.

Bubb, K. L., D. Bovee, D. Buckley, E. Haugen, M. Kibukawa, M. Paddock, A. Palmieri, S. Subramanian, Y. Zhou, R. Kaul, P. Green, and M. V. Olson (2006). "Scan of human genome reveals no new loci under ancient balancing selection". In: *Genetics* 173(4), pp. 2165–2177.

Buckling, A. and P. B. Rainey (2002). "Antagonistic coevolution between a bacterium and a bacteriophage." In: *Proceedings. Biological sciences / The Royal Society* 269(1494), pp. 931–6.

Buerkle, C. A., R. J. Morris, M. A. Asmussen, and L. H. Rieseberg (2000). "The likelihood of homoploid hybrid speciation." In: *Heredity* 84, pp. 441–451.

Buerkle, C. A. and L. H. Rieseberg (2008). "The rate of genome stabilization in homoploid hybrid species". In: *Evolution* 62(2), pp. 266–275.

Burke, J. M., T. J. Voss, and M. L. Arnold (1998). “Genetic interactions and natural selection in Louisiana Iris hybrids”. In: *Evolution* 52(5), pp. 1304–1310.

Burke, J. M. and M. L. Arnold (2001). “Genetics and the fitness of hybrids”. In: *Annual Review of Genetics* 35, pp. 31–52.

Burton, R. S. (1990a). “Hybrid Breakdown in developmental time in the Copepod *Tigriopus californicus*”. In: *Evolution* 44(7), pp. 1814–1822.

— (1990b). “Hybrid breakdown in physiological response: a mechanistic approach”. In: *Evolution* 44(7), pp. 1806–1813.

Burton, R. S., P. D. Rawson, and S. Edmands (1999). “Genetic architecture of physiological phenotypes: Empirical evidence for coadapted gene complexes”. In: *American Zoologist* 39(2), pp. 451–462.

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”. In: *Genetics* 137(1), pp. 175–189.

Cambareri, E. B., M. J. Singer, and E. U. Selker (1991). “Recurrence of repeat-induced point mutation (RIP) in *Neurospora crassa*”. In: *Genetics* 127(4), pp. 699–710.

Casabianca, S., A. Penna, E. Pecchioli, A. Jordi, G. Basterretxea, and C. Vernesi (2012). “Population genetic structure and connectivity of the harmful dinoflagellate *Alexandrium minutum* in the Mediterranean Sea.” In: *Proceedings of the Royal Society B: Biological Sciences* 279(1726), pp. 129–38.

Cavalier-Smith, T (2002). “Origins of the machinery of recombination and sex.” In: *Heredity* 88, pp. 125–141.

Chakraborty, R. and M. Nei (1977). “Bottleneck effects on average heterozygosity and genetic distance with the stepwise mutation model”. In: *Evolution* 31(2), pp. 347–356.

3981 Chappell, T. M. and M. D. Rausher (2016). “Evolution of host range in
3982 *Coleosporium ipomoeae*, a plant pathogen with multiple hosts”. In: *Pro-*
3983 *ceedings of the National Academy of Sciences* 113(19), p. 201522997.
3984 Charlesworth, B. and D. Charlesworth (2010). *Elements of Evolutionary*
3985 *Genetics*. Greenwood Village, Colorado: Roberts and Company, p. 734.
3986 Charlesworth, D. (2006). “Balancing selection and its effects on sequences
3987 in nearby genome regions”. In: *PLoS Genetics* 2(4), pp. 379–384.
3988 Charlesworth, D. and J. H. Willis (2009). “The genetics of inbreeding de-
3989 pression.” In: *Nature Reviews Genetics* 10(11), pp. 783–96.
3990 Chepurnov, V. A., D. G. Mann, K. Sabbe, and W. Vyverman (2004). “Ex-
3991 perimental studies on sexual reproduction in diatoms”. In: *International*
3992 *Review of Cytology* 237, pp. 91–154.
3993 Choi, D and M. J. Priest (1955). “A key to the genus *Albugo*”. In: *Mycotaxon*
3994 53, pp. 261–272.
3995 Choi, Y.-J., H.-D. Shin, and M. Thines (2009). “The host range of *Albugo*
3996 *candida* extends from *Brassicaceae* through *Cleomaceae* to *Cappa-*
3997 *raceae*”. In: *Mycological Progress* 8(4), pp. 329–335.
3998 Choi, Y.-J., H.-D. Shin, S. Ploch, and M. Thines (2011). “Three new phy-
3999 logenetic lineages are the closest relatives of the widespread species
4000 Albugo candida.” In: *Fungal biology* 115(7), pp. 598–607.
4001 Clark Cockerham, C. and Z. B. Zeng (1996). “Design III with marker loci”.
4002 In: *Genetics* 143(3), pp. 1437–1456.
4003 Coop, G. and M. Przeworski (2007). “An evolutionary view of human recom-
4004 bination.” In: *Nature Reviews Genetics* 8(1), pp. 23–34.
4005 Cooper, A. J., A. O. Latunde-Dada, A Woods-Tör, J Lynn, J. A. Lucas, I. R.
4006 Crute, and E. B. Holub (2008). “Basic compatibility of *Albugo candida*
4007 in *Arabidopsis thaliana* and *Brassica juncea* causes broad-spectrum
4008 suppression of innate immunity.” In: *Molecular plant-microbe interactions*
4009 *: MPMI* 21(6), pp. 745–756.

4010 Cousyn, C, L De Meester, J. K. Colbourne, L Brendonck, D Verschuren, and
4011 F Volckaert (2001). “Rapid, local adaptation of zooplankton behavior
4012 to changes in predation pressure in the absence of neutral genetic
4013 changes.” In: *Proceedings of the National Academy of Sciences of the*
4014 *United States of America* 98(11), pp. 6256–6260.
4015 Crow, J. F. (1994). “Advantages of sexual reproduction”. In: *Developmental*
4016 *Genetics* 15(3), pp. 205–213.
4017 Crow, J. F. and M. Kimura (1970). *An Introduction to Population Genetics*
4018 *Theory*. New York: Harper and Row, p. 656.
4019 Cruzan, M. B. and M. L. Arnold (1994). “Assortative Mating and Natural-
4020 Selection in an Iris Hybrid Zone”. In: *Evolution* 48(6), pp. 1946–1958.
4021 Darlington, C. D. and K. Mather (1950). *The Elements of Genetics*. New
4022 York: Macmillan, p. 462.
4023 Davidovich, N. A. and S. S. Bates (1998). “Patterns of sexual reproduction
4024 in diatoms”. In: *Hydrobiologia* 269-270, pp. 11–20.
4025 Davis, B and B. Columbia (2011). “Genetic load”. In: *eLS* July, pp. 1–4.
4026 Dawson, E., G. R. Abecasis, S. Bumpstead, Y. Chen, S. Hunt, D. M. Beare,
4027 J. Pabial, T. Dibling, E. Tinsley, S. Kirby, D. Carter, M. Papaspyridonos,
4028 S. Livingstone, R. Ganske, E. Lohmussaar, J. Zernant, N. Tonisson,
4029 M. Remm, R. Magi, T. Puurand, J. Vilo, A. Kurg, K. Rice, P. Deloukas,
4030 R. Mott, A. Metspalu, D. R. Bentley, L. R. Cardon, and I. Dunham (2002).
4031 “A first-generation linkage disequilibrium map of human chromosome
4032 22”. In: *Nature* 418, pp. 544–548.
4033 Decaestecker, E., S. Gaba, J. A. M. Raeymaekers, R. Stoks, L. Van Kerck-
4034 hoven, D. Ebert, and L. De Meester (2007). “Host-parasite ’Red Queen’
4035 dynamics archived in pond sediment.” In: *Nature* 450(7171), pp. 870–
4036 873.

4037 Defaveri, J. and J. Meril (2014). “Local adaptation to salinity in the three-
4038 spined stickleback?” In: *Journal of Evolutionary Biology* 27(2), pp. 290–
4039 302.
4040 Delaneau, O., J. Marchini, and J.-F. Zagury (2012). “A linear complexity
4041 phasing method for thousands of genomes.” In: *Nature Methods* 9(2),
4042 pp. 179–81.
4043 Delaneau, O., J.-F. Zagury, and J. Marchini (2013). “Improved whole-
4044 chromosome phasing for disease and population genetic studies.” In:
4045 *Nature Methods* 10(1), pp. 5–6.
4046 Delaneau, O., B. Howie, A. J. Cox, J.-F. Zagury, and J. Marchini (2013).
4047 “Haplotype estimation using sequencing reads”. In: *American Journal of*
4048 *Human Genetics* 93(4), pp. 687–696.
4049 Depotter, J. R. L., M. F. Seidl, T. A. Wood, and B. P.H. J. Thomma (2016).
4050 “Interspecific hybridization impacts host range and pathogenicity of fil-
4051 amentous microbes”. In: *Current Opinion in Microbiology* 32, pp. 7–
4052 13.
4053 Devlin, B. and N. Risch (1995). “Linkage disequilibrium measures for fine-
4054 scale mapping: a comparison”. In: *Genomics* 29, pp. 311–322.
4055 Diogo, D., C. Bouchier, C. D’Enfert, and M. E. Bougnoux (2009). “Loss
4056 of heterozygosity in commensal isolates of the asexual diploid yeast
4057 Candida albicans”. In: *Fungal Genetics and Biology* 46(2), pp. 159–168.
4058 Dobzhansky, T (1936). “Studies on hybrid sterility. II. Localization of steril-
4059 ity factors in *Drosophila Pseudoobscura* Hybrids.” In: *Genetics* 21(2),
4060 pp. 113–135.
4061 — (1970). *Genetics of the Evolutionary Process*. A Columbia Paperback.
4062 New York and London: Columbia University Press, p. 506.
4063 Dong, S., S. Raffaele, and S. Kamoun (2015). “The two-speed genomes
4064 of filamentous pathogens: Waltz with plants”. In: *Current Opinion in*
4065 *Genetics and Development* 35, pp. 57–65.

Dong, S., R. Stam, L. M. Cano, J. Song, J. Sklenar, K. Yoshida, T. O. Bozkurt, R. Oliva, Z. Liu, M. Tian, J. Win, M. J. Banfield, A. M. E. Jones, R. A. L. Van der Hoorn, and S. Kamoun (2014). "Effector Specialization in a Lineage of the Irish Potato Famine Pathogen". In: *Science* 343(6170), pp. 552–555.

Dres, M. and J. Mallet (2002). "Host races in plant-feeding insects and their importance in sympatric speciation". In: *Philosophical Transactions of the Royal Society B: Biological Sciences* 357(1420), pp. 471–492.

Drummond, A. and A. Rambaut (2007). "BEAST: Bayesian evolutionary analysis by sampling trees". In: *BMC Evolutionary Biology* 7(1), p. 214.

Drummond, A. J. and M. A. Suchard (2010). "Bayesian random local clocks, or one rate to rule them all." In: *BMC biology* 8, p. 114.

Drummond, A. J., M. A. Suchard, D. Xie, and A. Rambaut (2012). "Bayesian phylogenetics with BEAUti and the BEAST 1.7". In: *Molecular Biology and Evolution* 29(8), pp. 1969–1973.

Dujon, B. (2010). "Yeast evolutionary genomics". In: *Nature Reviews Genetics* 11(7), pp. 512–524.

Durand, E. Y., N. Patterson, D. Reich, and M. Slatkin (2011). "Testing for ancient admixture between closely related populations." In: *Molecular biology and evolution* 28(8), pp. 2239–52.

Edmands, S. (1999). "Heterosis and outbreeding depression in interpopulation crosses spanning a wide range of divergence". In: *Evolution* 53(6), pp. 1757–1768.

Edwards, S. V. and P. W. Hedrick (1998). "Evolution and ecology of MHC molecules: From genomics to sexual selection". In: *Trends in Ecology and Evolution* 13(8), pp. 305–311.

Eisen, J. A. (2000). "Horizontal gene transfer among microbial genomes: new insights from complete genome analysis." In: *Current Opinion in Genetics and Development* 10(6), pp. 606–11.

4095 Endler, J. A. (1977). “Geographic variation, speciation, and clines.” In:
4096 *Monographs in Population Biology* 10, pp. 1–246.
4097 Endler, J. A. (1982). “Problems in distinguishing historical from ecological
4098 factors in biogeography”. In: *American Zoologist* 22(2), pp. 441–452.
4099 Ersek, T, J. T. English, and J. E. Schoelz (1995). “Creation of species
4100 hybrids of *Phytophthora* with modified host ranges by zoospore fusion”.
4101 In: *Phytopathology* 85(11), pp. 1343–1347.
4102 Felsenstein, J. and S. Yokoyama (1976). “The evolutionary advantage of
4103 recombination. II. Individual selection for recombination”. In: *Genetics*
4104 83(4), pp. 845–859.
4105 Felsenstein, J. (1981). “Evolutionary trees from DNA sequences: A maxi-
4106 mum likelihood approach”. English. In: *Journal of Molecular Evolution*
4107 17(6), pp. 368–376.
4108 Fisher, R. A. (1930). *The Genetical Theory of Natural Selection*. Oxford: At
4109 The Clarendon Press, p. 308.
4110 Fontaine, M. C., J. B. Pease, A. Steele, R. M. Waterhouse, D. E. Neafsey,
4111 I. V. Sharakhov, X. Jiang, A. B. Hall, F. Catteruccia, E. Kakani, S. N.
4112 Mitchell, Y. Wu, H. A. Smith, R. R. Love, M. K. N. Lawniczak, M. A.
4113 Slotman, S. J. Emrich, M. W. Hahn, and N. J. Besansky (2015). “Ex-
4114 tensive introgression in a malaria vector species complex revealed by
4115 phylogenomics”. In: *Science* 347(6217), pp. 12585241–12585246.
4116 Ford Doolittle, W. (1999). “Lateral genomics”. In: *Trends in Biochemical*
4117 *Sciences* 24(12), pp. 5–8.
4118 Frank, S (1993). “Coevolutionary genetics of plant pathogens”. In: *Evolu-*
4119 *tionay Ecology* 7, pp. 45–75.
4120 Frankham, R (1996). “Relationship of genetic variation to population size in
4121 wildlife”. In: *Conservation Biology* 10(6), pp. 1500–1508.
4122 Gallagher, J. (1983). “Cell Enlargement in *Skeletonema Costatum* (*Bacillar-*
4123 *iophyceae*)”. In: *Journal of phycology* 19(4), pp. 539–542.

Galtier, N. and V. Daubin (2008). “Dealing with incongruence in phylogenomic analyses.” In: *Philosophical Transactions of the Royal Society of London. Series B, Biological Sciences* 363(1512), pp. 4023–4029.

Gardner, K., C. A. Buerkle, J. Whitton, and L. H. Rieseberg (2000). “Inferring Epistasis in Wild Sunflower Hybrid Zones”. In: *Epistasis and the Evolutionary Process*. Ed. by J. B. Wolf, E. D. Brodie, and M. J. Wade. Oxford: Oxford University Press. Chap. 16, pp. 264–279.

Gerrish, P. J. and R. E. Lenski (1998). “The fate of competing beneficial mutations in an asexual population.” In: *Genetica* 102-103(1-6), pp. 127–44.

Gijzen, M., C. Ishmael, and S. D. Shrestha (2014). “Epigenetic control of effectors in plant pathogens”. In: *Frontiers in Plant Science* 5, pp. 1–4.

Giraud, A, M Radman, I Matic, and F Taddei (2001). “The rise and fall of mutator bacteria.” In: *Current Opinion in Microbiology* 4(5), pp. 582–5.

Goodwin, S, L. S. Sujkowski, A. T. Dyer, B. A. Fry, and W. E. Fry (1995). “Direct detection of gene flow and probably sexual reproduction of *Phytophthora infestans* in northern North America”. In: *Phytopathology* 85, pp. 473–479.

Goodwin, S. B., B. A. Cohen, and W. E. Fry (1994). “Panglobal distribution of a single clonal lineage of the Irish potato famine fungus.” In: *Proceedings of the National Academy of Sciences of the United States of America* 91(24), pp. 11591–5.

Goodwin, S. B., L. J. Spielman, J. M. Matuszak, S. M. Bergeron, and W. E. Fry (1992). “Clonal diversity and genetic differentiation of *Phytophthora infestans* populations in northern and central Mexico”. In: *Phytopathology* 82, pp. 955–961.

Goodwin, S., J. D. McPherson, and W. R. McCombie (2016). “Coming of age: ten years of next-generation sequencing technologies”. In: *Nature Reviews Genetics* 17(6), pp. 333–351.

Grauer, D. and W.-H. Li (2000). *Fundamentals of Molecular Evolution*. Second Edi. Sunderland: Sinauer Associates. Chap. Genes, gen, p. 481.

Greaves, J. H., R Redfern, P. B. Ayres, and J. E. Gill (1977). "Warfarin resistance: a balanced polymorphism in the Norway rat". In: *Genetical Research* 30(3), pp. 257–63.

Green, R. E. et al. (2010). "A draft sequence of the Neandertal genome". In: *Science* 328(5979), pp. 710–722.

Hacquard, S., B. Kracher, T. Maekawa, S. Vernaldi, P. Schulze-Lefert, and E. Ver Loren van Themaat (2013). "Mosaic genome structure of the barley powdery mildew pathogen and conservation of transcriptional programs in divergent hosts." In: *Proceedings of the National Academy of Sciences of the United States of America* 110(24), pp. 2219–28.

Haldane, J. B. S. (1948). "The theory of a cline". In: *Journal of Genetics* 48(3), pp. 277–284.

Haldane, J. B. S. and S. D. Jayakar (1963). "Polymorphism due to selection of varying direction". In: *Journal of Genetics* 58(2), pp. 237–242.

Hanski, I. (1998). "Metapopulation dynamics". In: *Nature* 396, pp. 41–49.

Hasegawa, M., H. Kishino, and T.-a. Yano (1985). "Dating of the human-ape splitting by a molecular clock of mitochondrial DNA". In: *Journal of Molecular Evolution* 22(2), pp. 160–174.

Heath, T. A., J. P. Huelsenbeck, and T. Stadler (2014). "The fossilized birth-death process for coherent calibration of divergence-time estimates". In: *Proceedings of the National Academy of Sciences* 111, pp. 2957–2966.

Hedrick, P. W. (1980). "Hitchhiking: a comparison of linkage and partial selfing." In: *Genetics* 94(3), pp. 791–808.

— (1987). "Gametic disequilibrium measures: proceed with caution." In: *Genetics* 117(2), pp. 331–341.

Hedrick, P. W. and M. E. Gilpin (1997). “Genetic effective size of a metapopulation”. In: *Metapopulation Dynamics: Ecology, Genetics, and Evolution*. Ed. by I. Hanski and M. Gilpin. New York: Academic Press, pp. 166–182.

Hedrick, P. W. (2010). *The genetics of populations*. 4th. Sudbury: Jones and Bartlett Publishers, p. 675.

— (2013). “Adaptive introgression in animals: examples and comparison to new mutation and standing variation as sources of adaptive variation”. In: *Molecular Ecology* 22(18), pp. 4606–4618.

Hein, J., M. Schierup, and C. Wiuf (2004). *Gene Genealogies, Variation and Evolution: A primer in coalescent theory*. Oxford and New York: Oxford University Press, p. 296.

Hewitt, G. M. (1985). “Analysis of hybrid zones”. In: *Annual Review of Ecology, Evolution, and Systematics* 16(1985), pp. 113–148.

Hill, W. G. and A. Robertson (1966). “The effect of linkage on artificial selection.” In: *Genetics Research* 8, pp. 269–294.

Hill, W. G. and A Robertson (1968). “Linkage disequilibrium in finite populations.” In: *Theoretical and applied genetics.* 38(6), pp. 226–31.

Hoekstra, H. E. and J. A. Coyne (2007). “The locus of evolution: Evo devo and the genetics of adaptation”. In: *Evolution* 61(5), pp. 995–1016.

Hollocher, H. and C. I. Wu (1996). “The genetics of reproductive isolation in the *Drosophila simulans* clade: X vs. autosomal effects and male vs. female effects”. In: *Genetics* 143(3), pp. 1243–1255.

Holub, E. B., E Brose, M Tör, C Clay, I. R. Crute, J. L. Beynon, M Tor, C Clay, I. R. Crute, and J. L. Beynon (1995). “Phenotypic and genotypic variation in the interaction between *Arabidopsis thaliana* and *Albugo candida*.” In: *Molecular Plant-Microbe Interactions : MPMI* 8(6), pp. 916–928.

Howie, B., J. Marchini, and M. Stephens (2011). “Genotype imputation with thousands of genomes”. In: *G3* 1(6), pp. 457–470.

Hudson, R. R. and N. L. Kaplan (1988). “The coalescent process in models with selection and recombination.” In: *Genetics* 120(3), pp. 831–40.

Hudson, R. R. (2001). “Two-locus sampling distributions and their application”. In: *Genetics* 159, pp. 1805–1817.

Huson, D. H. (1998). “SplitsTree: analyzing and visualizing evolutionary data.” In: *Bioinformatics* 14(1), pp. 68–73.

Huson, D. H. and D. Bryant (2006). “Application of phylogenetic networks in evolutionary studies.” In: *Molecular Biology and Evolution* 23(2), pp. 254–67.

Ingvarsson, P. K. and M. C. Whitlock (2000). “Heterosis increases the effective migration rate.” In: *Proceedings. Biological sciences / The Royal Society* 267(1450), pp. 1321–1326.

Jayaraman, R (2011). “Phase variation and adaptation in bacteria: A Red Queen arms race”. In: *Current Science* 100(8), pp. 1163–1171.

Jex, A. R., M. A. Schneider, and T. H. Cribb (2006). “The importance of host ecology in thelastomatoid (Nematoda: Oxyurida) host specificity.” In: *Parasitology International* 55(3), pp. 169–74.

Johansen-Morris, A. D. and R. G. Latta (2006). “Fitness consequences of hybridization between ecotypes of Avena barbata: hybrid breakdown, hybrid vigor, and transgressive segregation.” In: *Evolution* 60(8), pp. 1585–1595.

Jouet, A., M. McMullan, and C. van Oosterhout (2015). “The effects of recombination, mutation and selection on the evolution of the Rp1 resistance genes in grasses”. In: *Molecular Ecology* 24(12), pp. 3077–3092.

Judson, O. P. and B. B. Normark (1996). “Ancient asexual scandals”. In: *Trends in Ecology and Evolution* 11(2), pp. 41–46.

4236 Jukes, T. and C. Cantor (1969). “Evolution of Protein Molecules”. In: *Mam-*
4237 *malian Protein Metabolism III*. Ed. by H. N. Munro. New York: Academic
4238 Press, pp. 21–132.
4239 Kang, S. H. and G. A. Fryxell (1992). “*Fragilariopsis cylindrus* (Grunow)
4240 Krieger: The most abundant diatom in water column assemblages of
4241 Antarctic marginal ice-edge zones”. In: *Polar Biology* 12(6-7), pp. 609–
4242 627.
4243 Kaplan, N. L., R. R. Hudson, and C. H. Langley (1989). “The ’hitchhiking
4244 effect’ revisited”. In: *Genetics* 123(4), pp. 887–899.
4245 Karrenberg, S., C. Lexer, and L. H. Rieseberg (2007). “Reconstructing
4246 the history of selection during homoploid hybrid speciation.” In: *The*
4247 *American Naturalist* 169(6), pp. 725–37.
4248 Kauppi, L., A. J. Jeffreys, and S. Keeney (2004). “Where the crossovers are:
4249 recombination distributions in mammals.” In: *Nature Reviews Genetics*
4250 5(6), pp. 413–424.
4251 Kemen, E. and J. D. G. Jones (2012). “Obligate biotroph parasitism: can
4252 we link genomes to lifestyles?” In: *Trends in Plant Science*, pp. 1–10.
4253 Kemen, E., A. Gardiner, T. Schultz-Larsen, A. C. Kemen, A. L. Balmuth,
4254 A. Robert-Seilaniantz, K. Bailey, E. Holub, D. J. Studholme, D. MacLean,
4255 and J. D. G. Jones (2011). “Gene gain and loss during evolution of
4256 obligate parasitism in the white rust pathogen of *Arabidopsis thaliana*”.
4257 In: *PLoS Biology* 9(7), e1001094.
4258 Kimura, M. (1962). “On the probability of fixation of mutant genes in a
4259 population.” In: *Genetics* 47(391), pp. 713–719.
4260 Kimura, M (1968). “Evolutionary rate at the molecular level.” In: *Nature*
4261 217(5129), pp. 624–626.
4262 Kimura, M. and J. F. Crow (1964). “The number of alleles that can be
4263 maintained in a finite population”. In: *Genetics* 49, pp. 725–738.

Kimura, M and T Ohta (1971). *Theoretical Aspects of Population Genetics*. Monographs in Population Biology. Princeton, New Jersy: Princeton University Press.

Kimura, M. (1980). “A simple method for estimating evolutionary rates of base substitutions through comparative studies of nucleotide sequences”. In: *Journal of Molecular Evolution* 16(2), pp. 111–120.

— (1983). *The Neutral Theory of Molecular Evolution*. Cambridge: Cambridge University Press, p. 367.

Kimura, M. and J. F. Crow (1963). “The measurement of effective population number”. In: *Evolution* 17(3), pp. 279–288.

Koenig, R. L., R. C. Ploetz, and H. C. Kistler (1997). “*Fusarium oxysporum* f. sp. *cubense* consists of a small number of divergent and globally distributed clonal lineages.” In: *Phytopathology* 87(9), pp. 915–923.

Kovach, R. P., B. K. Hand, P. A. Hohenlohe, T. F. Cosart, M. C. Boyer, H. H. Neville, C. C. Muhlfeld, S. J. Amish, K. Carim, S. R. Narum, W. H. Lowe, F. W. Allendorf, and G. Luikart (2016). “Vive la résistance: genome-wide selection against introduced alleles in invasive hybrid zones”. In: *Proceedings of the Royal Society B: Biological Sciences* 283(1843).

Kozmin, S, G Slezak, A Reynaud-Angelin, C Elie, Y de Rycke, S Boiteux, and E Sage (2005). “UVA radiation is highly mutagenic in cells that are unable to repair 7,8-dihydro-8-oxoguanine in *Saccharomyces cerevisiae*”. In: *Proceedings of the National Academy of Sciences of the United States of America* 102(38), pp. 13538–43.

Krell, A. (2006). “Salt stress tolerance in the psychrophilic diatom *Fragilariopsis cylindrus*”. PhD thesis. Bremen, p. 124.

Kuhner, M. K. (2006). “LAMARC 2.0: maximum likelihood and Bayesian estimation of population parameters.” In: *Bioinformatics* 22(6), pp. 768–770.

Lai, Z., T. Nakazato, M. Salmaso, J. M. Burke, S. Tang, S. J. Knapp, and L. H. Rieseberg (2005). “Extensive chromosomal repatterning and the evolution of sterility barriers in hybrid sunflower species”. In: *Genetics* 171(1), pp. 291–303.

Lamichhaney, S., J. Berglund, M. S. Almén, K. Maqbool, M. Grabherr, A. Martinez-Barrio, M. Promerová, C.-J. Rubin, C. Wang, N. Zamani, B. R. Grant, P. R. Grant, M. T. Webster, and L. Andersson (2015). “Evolution of Darwins finches and their beaks revealed by genome sequencing”. In: *Nature* 518(7539), pp. 371–375.

Lamour, A. K. H., J. Mudge, D. Gobena, O. P. Hurtado-gonzales, K Bharti, R. S. Donahoo, S. Finley, E. Huitema, J. Hulvey, and D. Platt (2010). “Genome sequencing and mapping reveal loss of heterozygosity as a mechanism for rapid adaptation in the vegetable pathogen *Phytophthora capsici*.” In: *Molecular Plant-Microbe Interactions* 25(10), pp. 1350–60.

Lamour, K. H. and S. Kamoun (2009). *Oomycete Genetics and Genomics: Diversity, Interactions and Research Tools*. Chichester: Wiley Blackwell, p. 582.

Lamour, K. H., R. Stam, J. Jupe, E. Huitema, E. Road, and D. Dd (2012). “Pathogen profile The oomycete broad-host-range pathogen Phytophthora capsici”. In: *Molecular Plant Pathology* 13(4), pp. 329–337.

Lande, R. (1994). “Risk of population extinction from fixation of new deleterious mutations”. In: *Evolution* 48(5), pp. 1460–1469.

Le Quesne, W. J. (1969). “A method of selection of characters in numerical taxonomy”. In: *Systematic Biology* 18(2), pp. 201–205.

Lemey, P, M Salemi, and A. Vamdamme (2009). *The Phylogenetic Handbook*. Cambridge: Cambridge University Press, p. 723.

Lemey, P., M. Lott, D. P. Martin, and V. Moulton (2009). “Identifying recombinants in human and primate immunodeficiency virus sequence alignments using quartet scanning.” In: *BMC bioinformatics* 10, p. 126.

Lenormand, T. (2002). “Gene flow and the limits to natural selection”. In: *Trends in Ecology and Evolution* 17(4), pp. 183–189.

Levins, R. (1969). “Some demographic and genetic consequences of environmental heterogeneity for biological control”. In: *Bull. Entomol. Soc. Am.* 15, pp. 237–240.

Lewontin, R. C. (1988). “On measures of gametic disequilibrium”. In: *Genetics* 120(3), pp. 849–852.

Lewontin, R. C. and K Kojima (1960). “The evolutionary dynamics of complex polymorphisms”. In: *Evolution* 14(4), pp. 458–472.

Li, C. C. (1976). *First Course in Population Genetics*. Pacific Grove, CA: Boxwood Press, p. 631.

Li, H. (2011). “Improving SNP discovery by base alignment quality”. In: *Bioinformatics* 27(8), pp. 1157–1158.

Li, H. and R. Durbin (2009). “Fast and accurate short read alignment with Burrows-Wheeler transform”. In: *Bioinformatics* 25(14), pp. 1754–1760.

Li, H., B. Handsaker, A. Wysoker, T. Fennell, J. Ruan, N. Homer, G. Marth, G. Abecasis, and R. Durbin (2009). “The sequence alignment/map format and SAMtools”. In: *Bioinformatics* 25(16), pp. 2078–2079.

Li, W.-h. (2008). “Molecular Clocks”. In: *eLS*.

Lippman, Z. B. and D. Zamir (2007). “Heterosis: revisiting the magic”. In: *Trends in Genetics* 23(2), pp. 60–66.

Little, T. J. and P. D. Hebert (1996). “Ancient asexuals: scandal or artifact?” In: *Trends in Ecology & Evolution* 11(7), p. 296.

Löytynoja, A. and N. Goldman (2005). “An algorithm for progressive multiple alignment of sequences with insertions.” In: *Proceedings of the National Academy of Sciences of the United States of America* 102(30), pp. 10557–62.

4348 Lunt, D. H. (2008). “Genetic tests of ancient asexuality in root knot ne-
4349 matodes reveal recent hybrid origins.” In: *BMC evolutionary biology* 8,
4350 p. 194.
4351 Lynch, M., J. Conery, and R. Burger (1995). “Mutational meltdowns in
4352 sexual populations”. In: *Evolution* 49(6), pp. 1067–1080.
4353 Lynch, M. and W. Gabriel (1990). “Mutation load and the survival of small
4354 populations”. In: *Evolution* 44(7), pp. 1725–1737.
4355 Macholán, M., P. Munclinger, M. Šugerková, P. Dufková, B. Bímová, E.
4356 Božíková, J. Zima, and J. Piálek (2007). “Genetic analysis of autosomal
4357 and X-linked markers across a mouse hybrid zone”. In: *Evolution* 61(4),
4358 pp. 746–771.
4359 Madigan, M., J. Martinko, D. Stahl, and D. Clark (2012). “Genetics of
4360 Bacteria and Archaea”. In: *Biology of Microorganisms*. Thirteenth. New
4361 York: Pearson. Chap. 10, pp. 291–316.
4362 Mäkinen, V, D Belazzougui, F Cunial, and A. I. Tomescu (2015). *Genome-*
4363 *Scale Algorithm Design: Biological Sequence Analysis in the Era of*
4364 *High-Throughput Sequencing*. Cambridge: Cambridge University Press.
4365 Mallet, J. (2005). “Hybridization as an invasion of the genome”. In: *Trends*
4366 *in Ecology and Evolution* 20(5), pp. 229–237.
4367 Mann, D. G., S. M. McDonald, M. M. Bayer, S. J. M. Droop, V. A. Chepurnov,
4368 R. E. Loke, A Ciobanu, and J. M. H. Du Buf (2004). “The *Sellaphora*
4369 *pupula* species complex (Bacillariophyceae): morphometric analysis,
4370 ultrastructure and mating data provide evidence for five new species”.
4371 In: *Phycologia* 43(4), pp. 459–482.
4372 Mann, D. G. (1989). “The species concept in diatoms: Evidence for mor-
4373 phologically distinct, sympatric gamodemes in four epipelic species”. In:
4374 *Plant Systematics and Evolution* 164(1-4), pp. 215–237.
4375 Marchini, J. and B. Howie (2010). “Genotype imputation for genome-wide
4376 association studies”. In: *Nature Reviews Genetics* 11(7), pp. 499–511.

4377 Marchini, J., B. Howie, S. Myers, G. McVean, and P. Donnelly (2007). “A new
4378 multipoint method for genome-wide association studies by imputation of
4379 genotypes.” In: *Nature Genetics* 39(7), pp. 906–13.
4380 Martin, D and E Rybicki (2000). “RDP: detection of recombination amongst
4381 aligned sequences.” In: *Bioinformatics* 16(6), pp. 562–563.
4382 Martin, D. P., B. Murrell, M. Golden, A. Khoosal, and B. Muhire (2015).
4383 “RDP4: Detection and analysis of recombination patterns in virus genomes”.
4384 In: *Virus Evolution* 1(1), pp. 1–5.
4385 Martin, F. and S. Kamoun (2012). *Effectors in plant-microbe interactions*.
4386 Chichester: Wiley Blackwell, p. 444.
4387 Martin, S. H., J. W. Davey, and C. D. Jiggins (2014). “Evaluating the use of
4388 ABBA-BABA statistics to locate introgressed loci.” In: *Molecular Biology
4389 and Evolution* 32(1), pp. 244–257.
4390 Martin, S. H., K. K. Dasmahapatra, N. J. Nadeau, C. Salazar, J. R. Walters,
4391 F. Simpson, M. Blaxter, A. Manica, J. Mallet, and C. D. Jiggins (2013).
4392 “Genome-wide evidence for speciation with gene flow in *Heliconius*
4393 butterflies”. In: *Genome Research* 23, pp. 1817–1828.
4394 Martinsen, G. D., T. G. Whitham, R. J. Turek, and P. Keim (2001). “Hybrid
4395 populations selectively filter gene introgression between species.” In:
4396 *Evolution* 55(7), pp. 1325–1335.
4397 Maruyama, T. (1970). “On the rate of decrease of heterozygosity in circu-
4398 lar stepping stone models of populations”. In: *Theoretical Population
4399 Biology* 1(1), pp. 101–119.
4400 May, R. M. and R. M. Anderson (1983). “Epidemiology and genetics in
4401 the coevolution of parasites and hosts.” In: *Proceedings of the Royal
4402 Society of London. Series B* 219(1216), pp. 281–313.
4403 Maynard-Smith, J and J Haigh (1974). “Hitch-hiking effect of a favorable
4404 gene”. In: *Genetics Research* 23(1), pp. 23–35.

Maynard Smith, J and N. H. Smith (1998). “Detecting recombination from gene trees.” In: *Molecular Biology and Evolution* 15(2), pp. 590–599.

Mayr, E (1942). *Systematics and the Origin of Species from the Viewpoint of a Zoologist*. New York: Coloumbia University Press, p. 334.

McMahill, M. S., C. W. Sham, and D. K. Bishop (2007). “Synthesis-dependent strand annealing in meiosis.” In: *PLoS biology* 5(11), e299.

McMullan, M., A. Gardiner, K. Bailey, E. Kemen, B. B. J. Ward, V. Cevik, A. Robert-Seilaniantz, T. Schultz-Larsen, A. Balmuth, E. Holub, C. van Oosterhout, and J. D. J. Jones (2015). “Evidence for suppression of immunity as a driver for genomic introgressions and host range expansion in races of *Albugo candida*, a generalist parasite.” In: *eLife* 4(4), pp. 1–24.

McVean, G., P. Awadalla, and P. Fearnhead (2002). “A coalescent-based method for detecting and estimating recombination from gene sequences”. In: *Genetics* 160(March), pp. 1231–1241.

Meena, P., C Chattopadhyay, F. Singh, B. Singh, and A. Gupta (2002). “Yield loss in indian mustard due to white rust and effect of some cultural practices on Alternaria blight and white rust severity”. In: *Brassica* 4(1 & 2), pp. 18–24.

Mehrabi, R., A. H. Bahkali, K. A. Abd-Elsalam, M. Moslem, S. Ben M’Barek, A. M. Gohari, M. K. Jashni, I. Stergiopoulos, G. H. J. Kema, and P. J.G. M. De Wit (2011). “Horizontal gene and chromosome transfer in plant pathogenic fungi affecting host range”. In: *FEMS Microbiology Reviews* 35(3), pp. 542–554.

Meselson, M. and D. M. Welch (2007). “Stable heterozygosity?” In: *Science* 318(October), pp. 202–204.

Metzgar, D., T. Scripps, and L. Jolla (2007). “Mutation Rates: Evolution”. In: *eLS*, pp. 1–3.

4433 Mielczarek, M. and J. Szyda (2016). “Review of alignment and SNP calling
4434 algorithms for next-generation sequencing data”. In: *Journal of Applied*
4435 *Genetics* 57(1), pp. 71–79.
4436 Mimitou, E. P. and L. S. Symington (2009). “Nucleases and helicases take
4437 center stage in homologous recombination”. In: *Trends in Biochemical*
4438 *Sciences* 34(5), pp. 264–272.
4439 Mock, T., R. P. Otillar, J. Strauss, A. E. Allen, C. L. Dupont, S. Frickenhaus,
4440 F. Maumus, M. McMullan, A. Salamov, R. Sanges, S. Schmutz, A.
4441 Toseland, A. Veluchamy, B. J. Ward, T. Wu, K. W. Barry, A. Falciatore,
4442 M. I. Ferrante, A. E. Fortunato, G. Glockner, A. Gruber, R. Hipkin, M. G.
4443 Janech, P. Kroth, F. Leese, E. A. Lindquist, B. R. Lyon, J. Martin, P.
4444 Mayer, M. Parker, H. Quesneville, J. A. Raymond, C. Uhlig, R. E. Valas,
4445 K. U. Valentin, A. Z. Worden, E. V. Armbrust, C. Bowler, B. R. Green,
4446 C. van Oosterhout, and I. V. Grigoriev (0). “Extensive genetic diversity
4447 and differential bi-allelic expression in a Southern Ocean diatom”. In:
4448 *Nature*.
4449 Mock, T., R. P. Otillar, J. Strauss, M. McMullan, P. Paajanen, J. Schmutz,
4450 A. Salamov, R. Sanges, A. Toseland, B. J. Ward, A. E. Allen, C. L.
4451 Dupont, S. Frickenhaus, F. Maumus, A. Veluchamy, T. Wu, K. W. Barry,
4452 A. Falciatore, M. I. Ferrante, A. E. Fortunato, G. Glöckner, A. Gruber,
4453 R. Hipkin, M. G. Janech, P. G. Kroth, F. Leese, E. A. Lindquist, B. R.
4454 Lyon, J. Martin, C. Mayer, M. Parker, H. Quesneville, J. A. Raymond,
4455 C. Uhlig, R. E. Valas, K. U. Valentin, A. Z. Worden, E. V. Armbrust, M. D.
4456 Clark, C. Bowler, B. R. Green, V. Moulton, C. van Oosterhout, and I. V.
4457 Grigoriev (2017). “Evolutionary genomics of the cold-adapted diatom
4458 *Fragilariopsis cylindrus*”. In: *Nature* 541(7638), pp. 536–540.
4459 Morgan, W. and S. Kamoun (2007). “RXLR effectors of plant pathogenic
4460 oomycetes.” In: *Current Opinion in Microbiology* 10(4), pp. 332–8.

Morrell, P. L., T. D. Williams-Coplin, A. L. Lattu, J. E. Bowers, J. M. Chandler, and A. H. Paterson (2005). “Crop-to-weed introgression has impacted allelic composition of johnsongrass populations with and without recent exposure to cultivated sorghum”. In: *Molecular Ecology* 14(7), pp. 2143–2154.

Muller, H. J. (1932). “Some genetic aspects of sex”. In: *American Naturalist* 66, pp. 118–138.

Muller, H. J. (1942). “Isolating mechanisms, evolution, and temperature”. In: *Biology Symposium* 6, pp. 71–125.

Myerowitz, R. (1997). “Tay-Sachs disease-causing mutations and neutral polymorphisms in the Hex A gene”. In: *Human Mutation* 9(3), pp. 195–208.

Nadeau, N. J., A Whibley, R. T. Jones, J. W. Davey, K. K. Dasmahapatra, S. W. Baxter, M. A. Quail, M Joron, R. H. Ffrench-Constant, M. L. Blaxter, J Mallet, and C. D. Jiggins (2012). “Genomic islands of divergence in hybridizing *Heliconius butterflies* identified by large-scale targeted sequencing.” In: *Philosophical Transactions of the Royal Society B: Biological Sciences*, pp. 343–353.

Nader, J (0). “A Genetic Basis for Anthroponosis by Cryptosporidium”. In: *Unpublished*.

Nagai, S., Y. Hori, T. Manabe, and I. Imai (1995). “Restoration of cell size by vegetative cell enlargement in *Coscinodiscus wailesii* (Bacillariophyceae)”. In: *Phycologia* 34(6), pp. 533–535.

Needleman, S. B. and C. D. Wunsch (1970). “A general method applicable to the search for similarities in the amino acid sequence of two proteins”. In: *Journal of Molecular Biology* 48(3), pp. 443–453.

Nei, M (1987). *Molecular Evolutionary Genetics*. New York: Columbia University Press, p. 512.

4489 Nei, M and W. H. Li (1973). “Linkage disequilibrium in subdivided popula-
4490 tions.” In: *Genetics* 75(1), pp. 213–219.
4491 Nei, M. (2005). “Selectionism and neutralism in molecular evolution”. In:
4492 *Molecular Biology and Evolution* 22(12), pp. 2318–2342.
4493 Nei, M., T. Maruyama, and R. Chakraborty (1975). “The bottleneck effect
4494 and genetic variability in populations”. In: *Evolution* 29(1), pp. 1–10.
4495 Nielsen, R., J. S. Paul, A. Albrechtsen, and Y. S. Song (2011). “Genotype
4496 and SNP calling from next-generation sequencing data.” In: *Nature*
4497 *Reviews Genetics* 12(6), pp. 443–51.
4498 Nunney, L. (1999). “The effective size of a hierarchically structured popula-
4499 tion”. In: *Evolution* 53(1), pp. 1–10.
4500 Ochman, H, J. G. Lawrence, and E. A. Groisman (2000). “Lateral gene
4501 transfer and the nature of bacterial innovation.” In: *Nature* 405(6784),
4502 pp. 299–304.
4503 O’Connell, J., D. Gurdasani, O. Delaneau, N. Pirastu, S. Ulivi, M. Cocca, M.
4504 Traglia, J. Huang, J. E. Huffman, I. Rudan, R. McQuillan, R. M. Fraser,
4505 H. Campbell, O. Polasek, G. Asiki, K. Ekoru, C. Hayward, A. F. Wright,
4506 V. Vitart, P. Navarro, J. F. Zagury, J. F. Wilson, D. Toniolo, P. Gasparini,
4507 N. Soranzo, M. S. Sandhu, and J. Marchini (2014). “A general approach
4508 for haplotype phasing across the full spectrum of relatedness”. In: *PLoS*
4509 *Genetics* 10(4), e1004234.
4510 Ohta, T (1973). “Slightly deleterious mutant substitutions in evolution.” In:
4511 *Nature* 246(5428), pp. 96–98.
4512 Ohta, T. (1971). “Associative overdominance caused by linked detrimental
4513 mutations”. In: *Genetics Research* 18, pp. 277–286.
4514 Ohta, T. and M. Kimura (1969). “Linkage disequilibrium due to random
4515 genetic drift”. In: *Genetics Research* 13(691), pp. 47–55.

Oosterhout, C. van (2009). “A new theory of MHC evolution: beyond selection on the immune genes.” In: *Proceedings. Biological Sciences / The Royal Society* 276(1657), pp. 657–65.

Padhukasahasram, B., J. D. Wall, P. Marjoram, and M. Nordborg (2006). “Estimating recombination rates from single-nucleotide polymorphisms using summary statistics”. In: *Genetics* 174(3), pp. 1517–1528.

Padidam, M, S Sawyer, and C. M. Fauquet (1999). “Possible emergence of new geminiviruses by frequent recombination.” In: *Virology* 265(2), pp. 218–225.

Palopoli, M. F. and C. I. Wu (1994). “Genetics of hybrid male sterility between *Drosophila* sibling species: A complex web of epistasis is revealed in interspecific studies”. In: *Genetics* 138, pp. 329–341.

Paraskevis, D., K. Deforche, P. Lemey, G. Magiorkinis, A. Hatzakis, and A. M. Vandamme (2005). “SlidingBayes: Exploring recombination using a sliding window approach based on Bayesian phylogenetic inference”. In: *Bioinformatics* 21(7), pp. 1274–1275.

Paten, B., D. Earl, and N. Nguyen (2011). “Cactus : Algorithms for genome multiple sequence alignment”. In: *Genome Research* 21, pp. 1512–1528.

Paterson, S., T. Vogwill, A. Buckling, R. Benmayor, A. J. Spiers, N. R. Thomson, M. Quail, F. Smith, D. Walker, B. Libberton, A. Fenton, N. Hall, and M. a. Brockhurst (2010). “Antagonistic coevolution accelerates molecular evolution.” In: *Nature* 464(7286), pp. 275–278.

Payseur, B. A., M. Place, and J. L. Weber (2008). “Linkage Disequilibrium between STRPs and SNPs across the Human Genome”. In: *American Journal of Human Genetics* 82(5), pp. 1039–1050.

Pedersen, A. B., S. Altizer, M. Poss, A. A. Cunningham, and C. L. Nunn (2005). “Patterns of host specificity and transmission among parasites of

wild primates". In: *International Journal for Parasitology* 35(6), pp. 647–657.

Peng, B. and M. Kimmel (2005). "simuPOP: A forward-time population genetics simulation environment". In: *Bioinformatics* 21(18), pp. 3686–3687.

Ploch, S., Y. J. Choi, C. Rost, H. D. Shin, E. Schilling, and M. Thines (2010). "Evolution of diversity in Albugo is driven by high host specificity and multiple speciation events on closely related Brassicaceae". In: *Molecular Phylogenetics and Evolution* 57(2), pp. 812–820.

Policy, S. D. and D. J. Conway (2001). "Strong diversifying selection on domains of the *Plasmodium falciparum* apical membrane antigen 1 gene". In: *Genetics* 158(4), pp. 1505–1512.

Pondaven, P., O. Ragueneau, P. Tréguer, A. Hauvespre, L. Dezileau, and J. L. Reyss (2000). "Resolving the opal paradoxin the Southern Ocean". In: *Nature* 405(6783), 168172.

Posada, D and K. A. Crandall (2001). "Evaluation of methods for detecting recombination from DNA sequences: computer simulations." In: *Proceedings of the National Academy of Sciences of the United States of America* 98(24), pp. 13757–62.

Posada, D., K. A. Crandall, and E. C. Holmes (2002). "Recombination in evolutionary genomics." In: *Annual Review of Genetics* 36, pp. 75–97.

Pouchkina-Stantcheva, N. N., B. M. McGee, C. Boschetti, D. Tolleter, S. Chakrabortee, A. V. Popova, F. Meersman, D. Macherel, D. K. Hincha, and A. Tunnacliffe (2007). "Functional divergence of former alleles in an ancient asexual invertebrate." In: *Science* 318(5848), pp. 268–71.

Poulickova, A. (2008). "Morphology, cytology and sexual reproduction in the aerophytic cave diatom *Luticola dismutica* (Bacillariophyceae)". In: *Preslia* 80(1), pp. 87–99.

4572 Poulin, R and D Mouillot (2003). “Parasite specialization from a phyloge-
4573 netic perspective: a new index of host specificity”. In: *Parasitology* 126,
4574 pp. 473–480.
4575 — (2005). “Combining phylogenetic and ecological information into a new
4576 index of host specificity.” In: *The Journal of Parasitology* 91(3), pp. 511–
4577 514.
4578 Poulin, R. (2011). *Evolutionary Ecology of Parasites*. Second Edi. New
4579 Jersey: Princeton University Press, p. 360.
4580 Poulin, R. and D. B. Keeney (2008). “Host specificity under molecular and
4581 experimental scrutiny”. In: *Trends in Parasitology* 24(1), pp. 24–28.
4582 Pritchard, J. K. and M Przeworski (2001). “Linkage disequilibrium in humans:
4583 models and data.” In: *American Journal of Human Genetics* 69(1), pp. 1–
4584 14.
4585 Proulx, S. R. and P. C. Phillips (2006). “Allelic divergence precedes and
4586 promotes gene duplication.” In: *Evolution* 60(2003), pp. 881–892.
4587 Ptak, S. E., K. Voelpel, and M. Przeworski (2004). “Insights into recombi-
4588 nation from patterns of linkage disequilibrium in humans”. In: *Genetics*
4589 167(1), pp. 387–397.
4590 Raffaele, S. and S. Kamoun (2012). “Genome evolution in filamentous plant
4591 pathogens: why bigger can be better”. In: *Nature Reviews Microbiology*
4592 (May), pp. 1–14.
4593 Raffaele, S., J. Win, L. M. Cano, and S. Kamoun (2010). “Analyses of
4594 genome architecture and gene expression reveal novel candidate vir-
4595 ulence factors in the secretome of *Phytophthora infestans*”. In: *BMC*
4596 *Genomics* 11(1), p. 637.
4597 Rainey, P. B., A Buckling, R Kassen, and M Travisano (2000). “The emer-
4598 gence and maintenance of diversity: insights from experimental bacterial
4599 populations.” In: *Trends in Ecology & Evolution* 15(6), pp. 243–247.

Raymond, J. A. and H. J. Kim (2012). “Possible role of horizontal gene transfer in the colonization of sea ice by Algae”. In: *PLoS ONE* 7(5).

Ridley, M. (2004). *Evolution*. 3rd. Malden: Blackwell Publishing, p. 751.

Rieseberg, L. H., M. A. Archer, and R. K. Wayne (1999). “Transgressive segregation, adaptation and speciation.” In: *Heredity* 83, pp. 363–372.

Rieseberg, L. H., B. Sinervo, C. R. Linder, M. C. Ungerer, and D. M. Arias (1996). “Role of gene interactions in hybrid speciation: Evidence from ancient and experimental hybrids”. In: *Science* 272(5262), pp. 741–745.

Rieseberg, L. H. and S. E. Carney (1998). “Plant hybridization”. In: *New Phytologist* 140(4), pp. 599–624.

Rieseberg, L. H., J. Whitton, and K. Gardner (1999). “Hybrid zones and the genetic architecture of a barrier to gene flow between two sunflower species”. In: *Genetics* 152(2), pp. 713–727.

Robbins, T. R., L. E. Walker, K. D. Gorospe, S. A. Karl, A. W. Schrey, E. D. McCoy, and H. R. Mushinsky (2014). “Rise and fall of a hybrid zone: Implications for the roles of aggression, mate choice, and secondary succession”. In: *Journal of Heredity* 105(2), pp. 226–236.

Robertson, A. (1962). “Selection for heterozygotes in small populations.” In: *Genetics* 47(1953), pp. 1291–1300.

Rogers, J. and R. A. Gibbs (2014). “Comparative primate genomics:emerging patterns of genome content and dynamics”. In: *Nature Reviews Genetics* 15(5), pp. 347–359.

Rose, L. E., P. D. Bittner-Eddy, C. H. Langley, E. B. Holub, R. W. Michelmore, and J. L. Beynon (2004). “The maintenance of extreme amino acid diversity at the disease resistance gene, RPP13, in *Arabidopsis thaliana*”. In: *Genetics* 166(3), pp. 1517–1527.

Rousset, F. (1997). “Genetic differentiation and estimation of gene flow from f statistics under isolation by distance”. In: *Genetics* 145, pp. 1219–1228.

Ruderfer, D. M., S. C. Pratt, H. S. Seidel, and L. Kruglyak (2006). “Population genomic analysis of outcrossing and recombination in yeast.” In: *Nature Genetics* 38(9), pp. 1077–1081.

Ruhe, J., M. Agler, A. Placzek, K. Kramer, I. Finkemeier, and E. Kemen (2016). “Obligate biotroph pathogens of the genus *Albugo* are better adapted to active host defense compared to niche competitors”. In: *Frontiers in Plant Science* 7(820), pp. 1–17.

Russell, N (1988). “Oswald Avery and the origin of molecular biology.” In: *British Journal for the History of Science* 21, pp. 393–400.

Sabbe, K., V. A. Chepurnov, W. Vyverman, and D. G. Mann (2004). “Apomixis in Achnanthes (Bacillariophyceae); development of a model system for diatom reproductive biology”. In: *European Journal of Phycology* 39(3), pp. 327–341.

Saharan, G. S. and P. R. Verma (1992). *White rusts: a review of economically important species*. Tech. rep. Ottawa, Ontario: International Development Research Centre.

Saharan, G. S., P. R. Verma, P. D. Meena, and A Kumar (2014). *White Rust of Crucifers: Biology, Ecology and Management*. New Delhi: Springer India.

Sanford, E and M. W. Kelly (2011). “Local adaptation in marine invertebrates”. In: *Annual Review of Marine Science* 3, pp. 509–535.

Sanger, F. and A. R. Coulson (1975). “A rapid method for determining sequences in DNA by primed synthesis with DNA polymerase”. In: *Journal of Molecular Biology* 94(3), pp. 441–448.

Sanger, F, S Nicklen, and A. R. Coulson (1977). “DNA sequencing with chain-terminating inhibitors.” In: *Proceedings of the National Academy of Sciences of the United States of America* 74(12), pp. 5463–7.

Scally, A., J. Y. Dutheil, L. W. Hillier, G. E. Jordan, I. Goodhead, J. Herrero, A. Hobolth, T. Lappalainen, T. Mailund, T. Marques, S. Mccarthy, S. H.

Montgomery, P. C. Schwalie, Y. A. Tang, M. C. Ward, Y. Xue, B. Yngvadottir, C. Alkan, L. N. Andersen, S. Fitzgerald, T. a. Graves, Y. Gu, P. Heath, and A. Heger (2012). "Insights into hominid evolution from the gorilla genome sequence". In: *Nature* 483(7388), pp. 169–175.

Schilthuizen, M., R. F. Hoekstra, and E. Gittenberger (1999). "Selective increase of a rare haplotype in a land snail hybrid zone". In: *Proceedings of the Royal Society B: Biological Sciences* 266(1434), p. 2181.

Schlupp, I., A. Taebel-Hellwig, and M. Tobler (2010). "Equal fecundity in asexual and sexual mollies (*Poecilia*)". In: *Environmental Biology of Fishes* 88(2), pp. 201–206.

Schurko, A. M., M. Neiman, and J. M. Logsdon (2009). "Signs of sex: what we know and how we know it". In: *Trends in Ecology and Evolution* 24, pp. 208–217.

Scotti-Saintagne, C., S. Mariette, I. Porth, P. G. Goicoechea, T. Barreneche, C. Bodénès, K. Burg, and A. Kremer (2004). "Genome scanning for interspecific differentiation between two closely related oak species (Quercus robur L. and Q. petraea (Matt.) Liebl.)" In: *Genetics* 168(3), pp. 1615–1626.

Seger, J and J Antonovics (1988). "Dynamics of some simple host-parasite models with more than two in each species genotypes". In: *Philosophical Transactions of the Royal Society of London. Series B , Biological* 319(1196), pp. 541–555.

Shee, C., J. L. Gibson, and S. M. Rosenberg (2012). "Two mechanisms produce mutation hotspots at DNA breaks in *Escherichia coli*". In: *Cell Reports* 2(4), pp. 714–721.

Shriner, D., D. Nickle, M. Jensen, and J. Mullins (2003). "Potential impact of recombination on sitewise approaches for detecting positive natural selection". In: *Genetical Research* 81(2), pp. 115–121.

4686 Sildever, S., J. Sefbom, I. Lips, and A. Godhe (2016). “Competetive advan-
4687 tage and higher fitness in native populations of genetically structured
4688 planktonic diatoms.” In: *Environmental Microbiology* 18(12), pp. 4403–
4689 4411.
4690 Slatkin, M. (1995). “A measure of population subdivision based on mi-
4691 crosatellite allele frequencies”. In: *Genetics* 139(1), pp. 457–462.
4692 Slatkin, M. (1977). “Gene flow and genetic drift in a species subject to
4693 frequent local extinctions”. In: *Theoretical Population Biology* 12(3),
4694 pp. 253–262.
4695 Slattery, M., H. N. Kamel, S. Ankisetty, D. J. Gochfeld, C. A. Hoover, and
4696 R. W. Thacker (2008). “Hybrid vigor in a tropical pacific soft-coral com-
4697 munity”. In: *Ecological Monographs* 78(3), pp. 423–443.
4698 Smith, G. R. (2012). “How RecBCD enzyme and Chi promote DNA break
4699 repair and recombination: A molecular biologist’s view”. In: *Microbiology
4700 and Molecular Biology Reviews* 76(2), pp. 217–228.
4701 Smith, J. M. (1992). “Analyzing the mosaic structure of genes.” In: *Journal
4702 of Molecular Evolution* 34(2), pp. 126–129.
4703 Smith, T. and M. Waterman (1981). “Identification of common molecular
4704 subsequences”. In: *Journal of Molecular Biology* 147(1), pp. 195–197.
4705 Spurgin, L. G. and D. S. Richardson (2010). “How pathogens drive genetic
4706 diversity: MHC, mechanisms and misunderstandings.” In: *Proceedings.
4707 Biological Sciences / The Royal Society* 277, pp. 979–988.
4708 Spurgin, L. G., C. van Oosterhout, J. C. Illera, S. Bridgett, K. Gharbi, B. C.
4709 Emerson, and D. S. Richardson (2011). “Gene conversion rapidly gener-
4710 ates major histocompatibility complex diversity in recently founded bird
4711 populations”. In: *Molecular Ecology* 20, pp. 5213–5225.

4712 Steele, K. A., E. Humphreys, C. R. Wellings, and M. J. Dickinson (2001).
4713 "Support for a stepwise mutation model for pathogen evolution in aus-
4714 tralasian *Puccinia striiformis* f.sp. *tritici* by use of molecular markers". In:
4715 *Plant Pathology* 50(2), pp. 174–180.
4716 Stefansson, T. S., B. A. McDonald, and Y. Willi (2014). "The influence of
4717 genetic drift and selection on quantitative traits in a plant pathogenic
4718 fungus". In: *PLoS ONE* 9(11).
4719 Stephens, M. and P. Donnelly (2003). "A comparison of Bayesian methods
4720 for haplotype reconstruction from population genotype data". In: *Am. J.*
4721 *Hum. Genet* 73(2002), pp. 1162–1169.
4722 Stukenbrock, E. H. (2013). "Evolution, selection and isolation: a genomic
4723 view of speciation in fungal plant pathogens". In: *New Phytologist* 199(4),
4724 pp. 895–907.
4725 Stukenbrock, E. H. (2016). "The role of hybridization in the evolution and
4726 emergence of new fungal plant pathogens." In: *Phytopathology* 106(2),
4727 pp. 104–112.
4728 Stukenbrock, E. H. and T. Bataillon (2012). "A population genomics per-
4729 spective on the emergence and adaptation of new plant pathogens in
4730 agro-ecosystems". In: *PLoS Pathogens* 8(9), pp. 1–4.
4731 Stukenbrock, E. H., F. B. Christiansen, T. T. Hansen, J. Y. Dutheil, and
4732 M. H. Schierup (2012). "Fusion of two divergent fungal individuals led to
4733 the recent emergence of a unique widespread pathogen species." In:
4734 *Proceedings of the National Academy of Sciences of the United States*
4735 *of America* 109(27), pp. 1–6.
4736 Stumpf, M. P. H. and G. A. T. McVean (2003). "Estimating recombination
4737 rates from population-genetic data." In: *Nature Reviews Genetics* 4(12),
4738 pp. 959–68.

4739 Suchard, M. A., R. E. Weiss, K. S. Dorman, and J. S. Sinsheimer (2002).
4740 “Oh brother, where art thou? A Bayes factor test for recombination with
4741 uncertain heritage.” In: *Systematic Biology* 51(5), pp. 715–728.
4742 Sung, P. and H. Klein (2006). “Mechanism of homologous recombination:
4743 mediators and helicases take on regulatory functions”. In: *Nat Rev Mol*
4744 *Cell Biol* 7(10), pp. 739–750.
4745 Svensson, O., A. Smith, J. García-alonso, and C. van Oosterhout (2016).
4746 “Hybridization generates a hopeful monster : a hermaphroditic selfing
4747 cichlid”. In: *Royal Society Open Science* 3, p. 150684.
4748 Swanson-Wagner, R. A., Y. Jia, R. DeCook, L. a. Borsuk, D. Nettleton, and
4749 P. S. Schnable (2006). “All possible modes of gene action are observed
4750 in a global comparison of gene expression in a maize F1 hybrid and its
4751 inbred parents.” In: *Proceedings of the National Academy of Sciences*
4752 *of the United States of America* 103(18), pp. 6805–6810.
4753 Taddei, F and M Radman (1997). “Role of mutator alleles in adaptive
4754 evolution”. In: *Nature* 387, pp. 700–702.
4755 Tanaka, M. M., C. T. Bergstrom, and B. R. Levin (2003). “The evolution
4756 of mutator genes in bacterial populations: the roles of environmental
4757 change and timing.” In: *Genetics* 164(3), pp. 843–54.
4758 Tavare, S (1986). “Some probabilistic and statistical problems in the analysis
4759 of DNA sequences”. In: *American Mathematical Society: Lectures on*
4760 *Mathematics in the Life Sciences* 17, pp. 57–86.
4761 Tenaillon, O., B. Toupance, L. Nagard, and B. Godelle (1999). “Mutators,
4762 Population Size, Adaptive Landscape and the Adaptation of Asexual
4763 Populations of Bacteria”. In: *Genetics* 152, pp. 485–493.
4764 Thines, M., Y. J. Choi, E. Kemen, S. Ploch, E. Holub, H.-D. Shin, and J.
4765 Jones (2009). “A new species of *Albugo* parasitic to *Arabidopsis thaliana*
4766 reveals new evolutionary patterns in white blister rusts (*Albuginaceae*)”.

In: *Persoonia - Molecular Phylogeny and Evolution of Fungi* 22(1), pp. 123–128.

Thines, M. (2014). “Phylogeny and evolution of plant pathogenic oomycetes - a global overview”. In: *European Journal of Plant Pathology* 138(3), pp. 431–447.

Thomas, D. N. and G. S. Dieckmann (2002). “Antarctic Sea ice - a habitat for extremophiles.” In: *Science* 295(5555), pp. 641–644.

Thomson, G. (1977). “The effect of a selected locus on linked neutral loci”. In: *Genetics* 85(4), pp. 753–788.

Thorne, J. L., H Kishino, and I. S. Painter (1998). “Estimating the rate of evolution of the rate of molecular evolution.” In: *Molecular Biology and Evolution* 15(12), pp. 1647–57.

True, J. R., B. S. Weir, and C. C. Laurie (1996). “A genome-wide survey of hybrid incompatibility factors by the introgression of marked segments of *Drosophila mauritiana* chromosomes into *Drosophila simulans*”. In: *Genetics* 142(3), pp. 819–837.

Turner, T. L., M. W. Hahn, and S. V. Nuzhdin (2005). “Genomic islands of speciation in *Anopheles gambiae*”. In: *PLoS Biology* 3(9), pp. 1572–1578.

Ungerer, M. C., S. J. Baird, J Pan, and L. H. Rieseberg (1998). “Rapid hybrid speciation in wild sunflowers.” In: *Proceedings of the National Academy of Sciences of the United States of America* 95(20), pp. 11757–11762.

Ungerer, M. C., S. C. Strakosh, and Y. Zhen (2006). “Genome expansion in three hybrid sunflower species is associated with retrotransposon proliferation”. In: *Current Biology* 16(20), R872–R873.

Vancoppenolle, M., K. M. Meiners, C. Michel, L. Bopp, F. Brabant, G. Carnat, B. Delille, D. Lannuzel, G. Madec, S. Moreau, J. L. Tison, and P. van der Merwe (2013). “Role of sea ice in global biogeochemical cycles:

4795 Emerging views and challenges". In: *Quaternary Science Reviews* 79,
4796 pp. 207–230.
4797 Vilenchik, M. M. and A. G. Knudson (2003). "Endogenous DNA double-
4798 strand breaks: production, fidelity of repair, and induction of cancer." In:
4799 *Proceedings of the National Academy of Sciences* 100(22), pp. 12871–
4800 12876.
4801 Visser, J. A.G. M. de (2002). "The fate of microbial mutators." In: *Microbiol-*
4802 *ogy* 148, pp. 1247–52.
4803 Visser, J. A.G. M. de and S. F. Elena (2007). "The evolution of sex: empirical
4804 insights into the roles of epistasis and drift." In: *Nature Reviews Genetics*
4805 8(2), pp. 139–149.
4806 Voordeckers, K., C. A. Brown, K. Vanneste, E. van der Zande, A. Voet,
4807 S. Maere, and K. J. Verstrepen (2012). "Reconstruction of ancestral
4808 metabolic enzymes reveals molecular mechanisms underlying evolu-
4809 tionary innovation through gene duplication". In: *PLoS Biology* 10(12),
4810 e1001446.
4811 Wahlund, S. (1928). "Zusammensetzung von populationen und korrelation-
4812 serscheinungen vom standpunkt der vererbungslehre aus betrachtet".
4813 In: *Hereditas* 11, pp. 65–106.
4814 Wakeley, J (2009). *Coalescent Theory: An Introduction*. Greenwood Village,
4815 Colorado: Roberts and Company, p. 328.
4816 Walker, J and M. J. Priest (2007). "A new species of Albugo on *Pterostylis*
4817 (*Orchidaceae*) from Australia: confirmation of the genus *Albugo* on a
4818 monocotyledonous host." In: *Australian Plant Pathology* 36, pp. 181–
4819 185.
4820 Wall, J. D. (2000). "A comparison of estimators of the population recombi-
4821 nation rate". In: *Molecular Biology and Evolution* 17(1), pp. 156–163.

4822 Wang, S., D. Bailey, K. Lindsay, J. K. Moore, and M. Holland (2014). “Impact
4823 of sea ice on the marine iron cycle and phytoplankton productivity”. In:
4824 *Biogeosciences* 11(17), pp. 4713–4731.
4825 Waples (2002). “Definition and estimation of the effective population size
4826 in the conservation of endangered species”. In: *Population Viability*
4827 *Analysis*. Ed. by S. R. Beissinger and D. R. McCullough. Chicago:
4828 Chicago University Press. Chap. 8, pp. 147–168.
4829 Ward, B. and C. van Oosterhout (2016). “Hybridcheck: Software for the rapid
4830 detection, visualization and dating of recombinant regions in genome
4831 sequence data”. In: *Molecular Ecology Resources* 16(2), pp. 534–539.
4832 Watson, J. D. and F. H. C. Crick (1953). “Molecular structure of nucleic
4833 acids”. In: *Nature* 171(4356), pp. 737–738.
4834 Watterson, G. A. (1975). “On the number of segregating sites in genetical
4835 models without recombination.” In: *Theoretical population biology* 7,
4836 pp. 256–276.
4837 Weisse, T., U. Scheffel, P. Stadler, and W. Foissner (2007). “Local adapta-
4838 tion among geographically distant clones of the cosmopolitan freshwater
4839 ciliate *Meseres corlissi*. II. Response to pH”. In: *Aquatic Microbial Ecol-*
4840 *ogy* 47(3), pp. 289–297.
4841 Welch, M. D. and M Meselson (2000). “Evidence for the evolution of bdelloid
4842 rotifers without sexual reproduction or genetic exchange.” In: *Science*
4843 288(5469), pp. 1211–1215.
4844 Werner, D (1977). *The Biology of Diatoms*. Ed. by D. Werner. Oakland,
4845 California: University of California Press, p. 498.
4846 Whibley, A. C., N. B. Langlade, C. Andalo, A. I. Hanna, A. Bangham, C.
4847 Thébaud, and E. Coen (2006). “Evolutionary paths underlying flower
4848 color variation in *Antirrhinum*”. In: *Science* 313(5789), pp. 963–966.
4849 Whisson, S., R. Vetukuri, A. Avrova, and C. Dixelius (2012). “Can silencing
4850 of transposons contribute to variation in effector gene expression in

Phytophthora infestans?" In: *Mobile Genetic Elements* 2(2), pp. 110–114.

White, C., K. A. Selkoe, J. Watson, D. A. Siegel, D. C. Zacherl, and R. J. Toonen (2010). "Ocean currents help explain population genetic structure". In: *Proceedings The Royal Society* 277(1688), pp. 1685–94.

White, M. A. C., R. E. Blackith, R. M. Blackith, and J. Cheney (1966). "Cytogenetics of the viatica group morabine grasshoppers. I. the coastal species". In: *Australian Journal of Zoology* 15(2), pp. 263–302.

Whitlock, M. C. and N. H. Barton (1997). "The effective size of subdivided population". In: *Genetics* 146(1), pp. 427–441.

Wicker, T., F. Sabot, A. Hua-Van, J. L. Bennetzen, P. Capy, B. Chalhoub, A. Flavell, P. Leroy, M. Morgante, O. Panaud, E. Paux, P. SanMiguel, and A. H. Schulman (2007). "A unified classification system for eukaryotic transposable elements". In: *Nature Reviews Genetics* 8(12), pp. 973–982.

Win, J., K. V. Krasileva, S. Kamoun, K. Shirasu, B. J. Staskawicz, and M. J. Banfield (2012). "Sequence divergent RXLR effectors share a structural fold conserved across plant pathogenic oomycete species". In: *PLoS Pathogens* 8(1).

Worobey, M. (2001). "A novel approach to detecting and measuring recombination: New insights into evolution in viruses, bacteria, and mitochondria". In: *Molecular Biology and Evolution* 18(8), pp. 1425–1434.

Wright, S. (1932). "The roles of mutations, inbreeding, crossbreeding and selection in evolution". In: *Proceedings of the 11th International Congress of Genetics* 1, pp. 356–366.

Wright, S. (1931). "Evolution in Mendelian populations". In: *Genetics* 16, pp. 97–159.

— (1937). "The distribution of gene frequencies in populations." In: *Science* 23, pp. 307–320.

— (1940). “Breeding structure of populations in relation to speciation”. In: *The American Naturalist* 74(752), pp. 232–248.

— (1943). “Isolation by distance”. In: *Genetics* 28(2), pp. 114–138.

Yampolsky, L. Y., T. M. M. Schaer, and D. Ebert (2014). “Adaptive phenotypic plasticity and local adaptation for temperature tolerance in freshwater zooplankton.” In: *Proceedings. Biological sciences / The Royal Society* 281(1776), p. 20132744.

Yatabe, Y., N. C. Kane, C. Scotti-Saintagne, and L. H. Rieseberg (2007). “Rampant gene exchange across a strong reproductive barrier between the annual sunflowers, *Helianthus annuus* and *H. petiolaris*.” In: *Genetics* 175(4), pp. 1883–93.

Zhan, J., R. E. Pettway, and B. A. McDonald (2003). “The global genetic structure of the wheat pathogen *Mycosphaerella graminicola* is characterized by high nuclear diversity, low mitochondrial diversity, regular recombination, and gene flow”. In: *Fungal Genetics and Biology* 38(3), pp. 286–297.

Appendices

APPENDIX A

FALCON assembly haplotype divergence

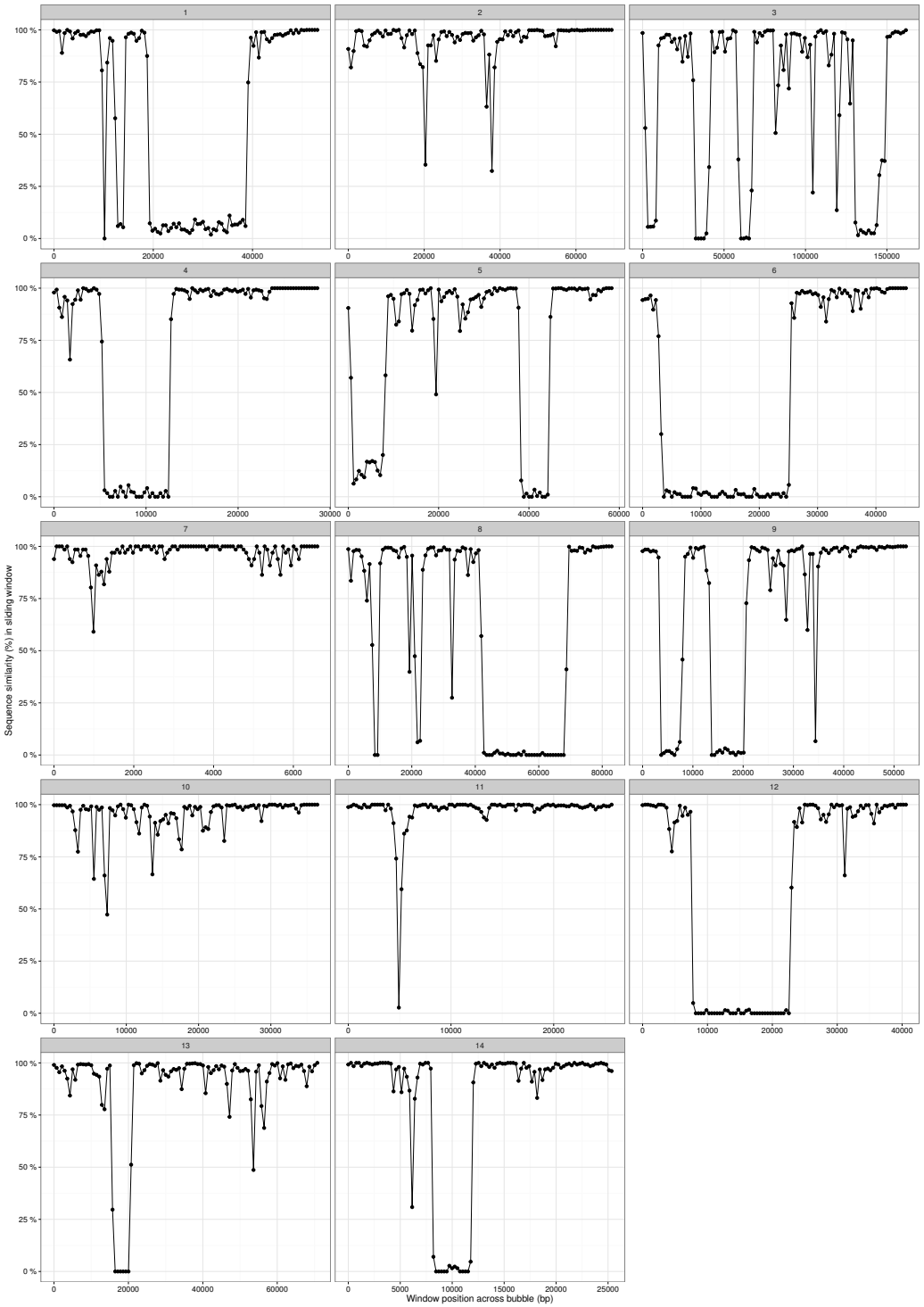


Figure A.1: Sequence similarity calculated with sliding windows across each haplotype 'bubble' in chromosome 000002F, from the *F. cylindrus* FALCON genome assembly. Regions of divergence and indels are apparent.