UMEÅ UNIVERSITY

FINDING FITNESS

Empirical and theoretical explorations of inferring fitness effects from population level SNP data

Bea Angelica Andersson

Department of Ecology and Environmental Science (EMG)

Umeå 2024

Dissertation for PhD
ISBN: 978-91-8070-268-3
Cover design by Bea Angelica Andersson
Electronic version available at: http://umu.diva-portal.org/
Printed by: Tryckservice Umeå Universitet
Umeå, Sweden 2024

Evolution happens.

What remains open to dispute, especially among scientists, is how evolution happens. Scientific theories themselves evolve, adapting to fit new observations, new discoveries and new interpretations of old discoveries. Theories are not carved in tablets of stone. The greatest strength of science is that when faced with sufficient evidence scientists change their minds. Not all of them, for scientists are human and have the same failings as the rest of us, but enough of them to allow science to improve.

– Terry Pratchett, Ian Stewart & Jack Cohen

The Science of Discworld III – Darwin's watch

Contents

Abstract

The distribution of fitness effects (DFE) describes the likelihood that a new mutation has a specific effect on the fitness of an individual in a given population. The shape of the DFE is a result of several factors such as population size, mating system and selective environment, and can in turn influence the evolutionary potential of a species. The DFE has long been a field of intense research, but particularly since molecular methods enabled us to study of genetic variation in organisms empirically. This research has led to the development of several statistical methods that use population-level frequencies of single nucleotide polymorphisms (SNPs) to infer the DFE. However, these methods rely on assumptions about the data and the organism itself, which could potentially affect the accuracy of the inferences. In this thesis, I describe how two major factors – data quality and inbreeding – can affect the accuracy of DFE inferences. I also show how and when to (and when not to) use DFE inference methods based on SNP frequencies.

All genomic datasets contain inaccuracies and some level of uncertainty. The data sets are therefore often treated to remove the gaps or less reliable information, such as genotypes with low coverage. Some data sets need heavy filtering, which could reduce the amount of data available for analysis. We show that the choice of filter method affects the size of the final data set and the accuracy of the estimated DFE.

Many DFE estimation software assumes random mating within the study population. Unfortunately, this assumption induces some error when trying to estimate the DFE in inbred or selfing species. Some have assumed that this is a result of high rates of homozygosity in the data, and should only be a problem in populations with very high rates of selfing (>99%). We show that accuracy of the estimated DFE decreases already at relatively low rates of selfing (70%) and that removing homozygosity does not improve the accuracy, implying that another mechanism could be causing the error.

Sammanfattning på svenska

Nya mutationer kan ha olika effekt på fitness hos en individ; en mutation kan vara negativ, neutral eller positiv för överlevnad och/eller reproduktion. Sannolikheten att en mutation har en specifik påverkan på fitness kan beskrivas av fördelningen av fitnesseffekter, i vetenskaplig litteratur kallad "the distribution of fitness effects" eller DFE. Formen på DFEn hos en art eller population påverkas av faktorer såsom effektiv populationsstorlek, parningssystem och miljö, och kan i sin tur påverka artens/populationens evolutionspotential. Hur DFEn ser ut hos olika arter har länge varit ett aktivt forskningsfält, och fick ytterligare ett uppsving efter att molekylära metoder gjorde det möjligt att studera genetisk variation empiriskt. Denna utveckling ledde även till att en uppsjö av statistiska metoder utvecklades för att uppskatta DFEn från allelfrekvenser hos punktmutationer i en population. Dessa modeller bygger emellertid på ett antal antaganden om populationen och datan, där osanna antaganden kan orsaka feluppskattningar av DFEn. I denna avhandling undersöker jag hur två faktorer – datakvalitet och inavel – kan påverka hur väl dessa metoder uppskattar den korrekta DFEn. Jag beskriver även hur och när man bör (och inte bör) använda allelefrekvensbaserade metoder för att uppskatta DFE hos en art.

Alla genomiska dataset innehåller en viss grad av osäkerhet och kan sakna information för vissa individer och/eller platser i genomet. Denna sortens data brukar därför förbehandlas för att exkludera obefintliga, osäkra och eventuellt felaktiga data. Vissa behandlingsmetoder kan exkludera stora delar data, beroende på hur mönstret av osäker data ser ut. Jag visar att mängden och magnituden hos feluppskattningar av DFEn beror på både valet av filtreringsmetod och storleken på det slutgiltiga datasetet.

Många metoder för DFE-uppskattning utgår ifrån teoretiska modeller som antar slumpmässig parning inom studiepopulationen. Detta antagande kan dock introducera feluppskattningar när de används på inavlade och/eller självbefruktande arter. Vissa tidigare studier har påstått att denna effekt beror på inavlade arters höga homozygositet, och bara borde vara ett problem vid stark inavel eller nästan uteslutande (>99%) självbefruktning. Jag beskriver här att feluppskattningar av DFEn blir vanligare redan vid lägre förekomst av självbefruktning (70-80%), samt att homozygositet i sig inte verkar vara den ledande orsaken till feluppskattningar av DFEn. Detta tyder därmed på att någon annan mekanism orsakar de större felmarginaler vi ser vid DFE-analys av inavlade arter.

List of papers

I. **Andersson, B. A.**, Zhao, W., Haller, B., Brännström, Å., & Wang, X.-R. (2023. Inference of the distribution of fitness effects of mutations is affected by SNP filtering methods, sample size and population structure. *Molecular Ecology Resources*, 23(7): 1589-1603.
https://doi.org/10.1111/1755-0998.13825

II. **Andersson, B. A.**, Zhao, W., Haller, B.,. Wang, X.-R., & Brännström, Å. (2023. Effects of self-fertilization on DFE inference. [MANUSCRIPT]

III. Guo, J.-F., Zhao, W., **Andersson, B.**, Mao, J.-F., & Wang, X.-R. (2023. Genomic clines across the species boundary between a hybrid pine and its progenitor in the eastern Tibetan Plateau. *Plant Communications*, 4*(*4), 100574.
https://doi.org/10.1016/j.xplc.2023.100574

IV. Zhao, W., Gao, J., Hall, D**., Andersson, B. A.**, Bruxaux, J., Tomlinson, K. W., T., Drouzas, A. D., Suyama, Y., Wang, X.-R. Adaptive radiation of the Asian Pinus species under pervasive gene flow. [New Phytologist, IN REVISION]

Author contributions

Paper I: WZ and XRW designed the empirical study. All authors contributed to designing the simulation study. BH provided support for simulations in SLiM 4.0. BA and WZ performed empirical data analyses. ÅB provided statistical advice. BA, WZ and XRW wrote the manuscript draft. All authors contributed to the revision of the manuscript.

Paper II: BA, WZ, XRW and ÅB designed the study. BH provided support for simulations in SLiM 4.0.1. BA performed and analyzed simulated data. BA and WZ performed empirical data analyses. ÅB provided statistical advice. BA wrote the manuscript draft. All authors contributed to the revision of the manuscript.

Paper III: XRW and WZ designed the study. WZ and JFM conducted field sampling. JFG, WZ and BA analysed the data. JFG, WZ and XRW wrote the manuscript draft. All authors contributed to the revision of the manuscript.

Paper IV: XRW designed the study. WZ, JG and DH performed genotyping. JG, ADD, YS, KWT and XRW provided samples. WZ analysed data with input from DH, BA and JB. WZ and XRW wrote the manuscript draft. All authors contributed to the revision of the manuscript.

On the origin of theses

There are many ways of doing a PhD. Some projects start off with a detailed project description, a timetable, and a list of expected research outcomes. My PhD began with a blank slate and a promise of tackling some interesting, but mostly undecided, questions in plant speciation. More specifically, the project description included studying "the genomics of ecological selection and adaptation, and how genetic factors interact with ecology to facilitate speciation".

The prospects were vague but exciting. I was encouraged to explore any idea, concept or method that may lead to a greater understanding of how species adapt and diverge, converge or persist. But how do you, when you have a whole field to choose from, decide on one question to focus on? I could study the process of ecological niche differentiation, hybridization and introgression, the genomic structure of plants and/or questions related to the origins of genetic incompatibility, or just about any aspect of how species adapt to new environments. Wherever I started reading, the question I had in mind kept folding out into more and more complex systems of sub-questions. Each conclusion or knowledge gap I found came with lists of caveats and assumptions about the applicability of each method used to identify them, all building upon each other. And I kept coming back to the same underlying question, which always seemed too simplistic to be entertained: How can we ever say anything about speciation or adaptation without knowing what each gene and allele is doing to the fitness of the individual and, by extension, the population? Of course, knowing that is impossible without extensive research into the functions of each allele and selection experiments in every possible environment, in every type of species, population and genetic context. Right?

Enter the Distribution of Fitness Effects

If we want to describe the genetic basis of the evolutionary history of some natural species, we may think that we need to know how each and every relevant gene and allele behaves over time. But how do we even quantify the total fitness effect of a mutation? Of course, the fitness effect is dependent on what the mutation does, meaning which amino acid it produces and its function, but also when it is activated, its interaction with other alleles, and, finally, how it affects the individual's performance in the environment compared to other alleles. We can either choose to take the difficult path and try to survey all mutations, estimate their selective effects in the relevant environment(s) and calculate the total fitness of any individual with a given genome. Alternatively, we could try to find a way to measure the total distribution of alleles with specific effects on fitness. As it turns out, the concept of estimating the so-called *distribution of fitness effects* (DFE) instead of the fitness effects of individuals alleles is not new, and today it is apparently possible using only allele frequencies! For example, we can estimate the likelihood that a mutation will have a given selective effect by leveraging the fact that the species has already been subject to selection in the past, which should have affected the frequencies of genetic variants in the population. That way, we do not need to know exactly what each allele does in order to say something about how evolution has affected its genetic diversity. This is a step towards understanding how a species has evolved in the past, as well as a way to understand its future evolutionary potential. Why not start there?

History of the DFE

When a new mutation occurs, it may or may not affect the fitness of the individual in which it occurs. If it does not, it is considered selectively neutral. If it does, it can have a net advantageous (fitness increasing) or deleterious (fitness decreasing) effect. The probability distribution that describes how likely it is for a mutation to have a specific effect on fitness is, fittingly, known as the *distribution of fitness effects*, or DFE for short (Eyre-Walker & Keightley 2007). If the DFE has the highest density around deleterious fitness effects, it means that most mutations that occur are detrimental to the survival and/or reproductive output of the organism. Thus, the DFE of new mutations describes what "raw material" will be available for future selection and evolution in a given species or population. If we knew the shape of the DFE of that species or population, we should be able to predict some aspects of how and how quickly it could evolve. As such, the DFE has been a focus in evolutionary biology research practically since its inception…

Darwinism

Even with no knowledge of genetics, Darwin recognized that inherited differences among individuals is what enables evolution by natural selection (Darwin 1859). Thus, the DFE is built into the four tenets of Darwinian evolution. Briefly, for evolution to occur it is required that:

1. more individuals are produced than can survive each generation,
2. heritable phenotypic variation exists among individuals,
3. individuals with heritable phenotypic traits that are better suited to the environment have a higher rate of survival, and that
4. new species will form when reproductive isolation occurs (Darwin 1859).

If we assume that mutations create the heritable variation in 2), the difference in "suitability" to the environment mentioned in 3) is the difference in the net fitness effects of those mutations. In order to calculate the total fitness of an individual, we can define the fitness effect w_m of a mutation *m,* by comparing the survival rate and total reproductive output of individuals with the *m* mutation to the maximum survival rate and reproductive output of the "fittest" mutation as:

$$w_m = \frac{survival_m \times reproduction_m}{survival_{max} \times reproduction_{max}}$$

A mutation that does not affect survival but only produces 6 offspring where the most prolific phenotype produces 10 will thus have a fitness effect of 0.6, while the most prolific phenotype always has a fitness of 1. By multiplying the fitness effects (w) of all traits, or mutations, we can calculate the total fitness of an individual (Gillespie 1998). If we want to compare the fitness effects of two mutations in a diploid species, where each individual has two alleles at each site, we can also do that: If we set the fitness effect of one mutation to 1, we can write the relative fitness of each genotype (AA, BB and AB) as:

$$w_{AA} = 1, \qquad w_{AB} = 1 + hs, \qquad w_{BB} = 1 + s \qquad (-1 \leq s \leq 1)$$

In this context, *s* signifies the selection coefficient, describing the fitness of genotype BB compared to genotype AA; if *s* is negative, B is a deleterious mutation that selection will act against and if it is positive, it is an advantageous mutation that selection will favour. *h* is the heterozygous effect. Values of $0 < h < 1$ indicate incomplete dominance, where an AB individual will experience part of the fitness effect of BB. If h is 1 or 0, B is either fully dominant or recessive (Gillespie 1998). What is most relevant to us, however, is that the distribution of fitness effects describes the probability that a new mutation will have a specific selection coefficient *s* (Eyre-Walker & Keightley 2007).

The Modern Synthesis

While mendelian inheritance was largely integrated into evolutionary theory by the 1920s and 30s (Dobzhansky 1937, Fisher 1919, 1923, Punnett 1930), it was not until the 1960s that the development of molecular methods would enable quantitative analysis of genetic variation (Lewontin & Hubby 1966). In the meantime, there was considerable debate around whether genetic variation was common or rare in natural populations. The field was generally divided into supporters of one of two hypotheses; The *classical hypothesis* argued that advantageous mutations were expected to quickly reach fixation, while deleterious mutations would be purged, leading to relatively low levels of polymorphism and heterozygosity; The *balance hypothesis*, on the other hand, predicted that genetic variation would be common, and that multiple alleles were being maintained at intermediate frequencies by processes such as heterozygote advantage or frequency dependent selection (Casillas & Barbadilla 2017, Fisher 1923, Hey 1999, Kimura 1979).

At this time, the most common approach to calculate the expected level of genetic diversity in a population was with deterministic mathematical models, where population sizes are infinite (Casillas & Barbadilla 2017). Infinitely large populations will not experience random genetic drift, and the frequency of a mutation in the next generation will only depend on its current frequency and selection coefficient. However, both hypotheses could be supported under these conditions, and without data on empirical levels of genetic diversity the debate stalled (Casillas & Barbadilla 2017, Charlesworth 1992). Only one thing was clear to both sides – mutations are either deleterious or advantageous to some extent (Fig. 1), and natural selection is the main force governing the amount of genetic variation in populations (Ford 1965).

The Neutral Theory

In 1966, it was shown that genetic variation in proteins (allozymes) among individuals could be quantified with electrophoresis, which opened the door to put the predictions of the classical and balance hypotheses to the test (Lewontin & Hubby 1966). To the surprise of many, including supporters of the balance hypothesis, it was revealed that most species harboured vast amounts of genetic variation, and that more alleles segregated at intermediate frequencies than previously anticipated. To some, this was the ultimate evidence for the validity of the balance theory and that balancing selection was maintaining genetic variation in natural populations. For others, however, the amount of variation seemed too large to possibly be maintained by active selection on all alleles simultaneously (Kimura 1979, Kimura et al. 1963).

In the wake of this discovery, a new theory took hold; Kimura (1968), and not much later Jack Lester King and Thomas H. Jukes (King & Jukes 1969), argued that instead of all variation being actively maintained by selection, the vast majority of segregating mutations are selectively neutral (Fig. 1). Under Kimura's *neutral theory of molecular evolution*, alleles could occur at intermediate frequencies in the population, not as a result of balancing selection, but simply because of random, non-selective processes such as genetic drift (Kimura 1968). Accumulating vast amounts of neutral variation would not come with any selective constraint or benefit, and alleles could segregate freely. In fact, they argued, selected mutations were probably rare, most likely almost exclusively deleterious and would be purged or fixed in the population so quickly that most models of molecular evolution could assume neutrality of mutations (Kimura 1968, King & Jukes 1969).

Perhaps unsurprisingly, this position faced considerable backlash from the outset (Casillas & Barbadilla 2017, Gillespie 1994, Kimura 1979, Kreitman 1996). The

neutralist arguments were in complete contrast to the earlier consensus that, although the field was divided on exactly how, natural selection was the primary force maintaining variation within natural systems. And while neutral mutations were not unknown, it was assumed that only (or at least mostly) synonymous mutations were neutral (Ford 1965).

The "neutralist-selectionist debate" of the 1970s and 80s centred around these two opposing ideas of the shape of the DFE (Casillas & Barbadilla 2017, Kimura 1979). However, one critical difference in the neutral theory set it apart from most previous models – the assumption of finite population sizes. Unlike deterministic selectionist models, the stochastic neutral model assumed finite populations, where the frequency of a mutation can increase or decrease because of random sampling effects (Casillas & Barbadilla 2017, Kimura 1968, Kimura 1979, Kimura et al. 1963). Until this point, the effects of genetic drift had been mostly ignored as a nuisance parameter, with few exceptions (Dobzhansky & Pavlovsky 1957), but the neutral model exposed genetic drift as a critical process in molecular evolution. Using diffusion equations, normally used to describe the random movements of particles in gasses, Kimura even devised a way to predict the average time to fixation of neutral alleles (Casillas & Barbadilla 2017, Kimura 1968, Kimura et al. 1963). Importantly, it also provided many predictions about diversity in populations that were easily testable using new types of molecular data that were becoming available. Thus, the value of Kimura's neutral model was recognized even among many selectionists and soon gained widespread approval because it also provided a null model for testing the effects of selection under finite population sizes (Casillas & Barbadilla 2017, Gillespie 1994, Kern & Hahn 2018, Kimura 1968, Kimura 1979, Kreitman 1996).

The Nearly Neutral Theory

BOX 1

The effective population size, N_e, is the size of an idealized population, meaning a population experiencing constant size, random mating and non-overlapping generations) that would contain the same amount of genetic variation as the target population.

N_e is generally smaller than the census population size, for example because of inbreeding or loss of genetic variation through population size changes.

As data on genetic variation accumulated, it soon became clear that the neutral model provided a rather simplified explanation of how alleles behaved. For example, Kimura's infinite site model had shown that under complete neutrality, heterozygosity in populations should be directly correlated with effective population size (see Box 1) and mutation rate. However, heterozygosity surveys from very large populations, which under this model should be nearly 100% heterozygous, revealed this not to be the case; a study of the fruit fly *Drosophila willistoni*, with an effective population size in the order of 10^8 individuals and an assumed mutation rate around 10^{-7}, demonstrated heterozygosity in only 18% of the sampled loci (Ayala et al. 1972). Soon thereafter, Tomoko Ohta, who had been working as a PhD student for Kimura during the development of the neutral model, developed a crucial extension of the neutral model. In her research, she had found a correlation between the fitness effects of mutations and effective population size beyond what was applied under the neutral theory. The neutral theory assumed that most mutations had a selection coefficient far below $\frac{-1}{N_e}$ and behaved neutrally, and that the remainder of the mutations were strongly deleterious mutations and would be purged instantaneously (Casillas & Barbadilla 2017, Kimura 1968, Kimura 1979). Ohta proposed that instead of mutations being either neutral or strongly deleterious, there was also a class of nearly neutral mutations, whose selection coefficients were in the order of $\frac{-1}{N_e}$ (or $\frac{1}{N_e}$ for beneficial mutations) (Fig. 1). Focusing mainly on deleterious mutations, she predicted that these mutations would act almost

neutrally in that they could increase in frequency by genetic drift effects but would still have a lower chance of becoming fixed in the population due to selection against them. This would mean that slightly deleterious mutations could comprise a relatively large proportion of the segregating variation in the population.

Importantly, she noted that the relationship to effective population size meant that more mutations would fall within the range of $\frac{-1}{N_e} - \frac{1}{N_e}$ in small populations. This meant that deleterious mutations would more readily become fixed, while they would be more likely to be purged in large populations where selection would have a stronger relative effect. This would also mean that the observed levels of genetic variation within populations could be explained, while still explaining why the numbers of substitutions (fixed mutations) did not seem to increase with population size (Ohta 1973). At first, Ohta's model only included the effects of slightly deleterious mutations, as they were thought to be the most common and have the largest effect on population level genetic variation. This version of the model was known as the *slightly deleterious theory* (Ohta 1973). Later, however, slightly advantageous mutations were also included under the model, which then became the highly impactful *nearly neutral theory* (Ohta 1992), which is still the basis of most models of molecular evolution (Casillas & Barbadilla 2017, Chen et al. 2020, Kreitman 1996, Nei 2005, Ohta 1996). Today, most research assumes that the DFE is skewed towards effectively and nearly neutral mutations, mostly with deleterious effects, which closely aligns with Ohta's predictions (Bataillon & Bailey 2014, Chen et al. 2021, Kousathanas & Keightley 2013).

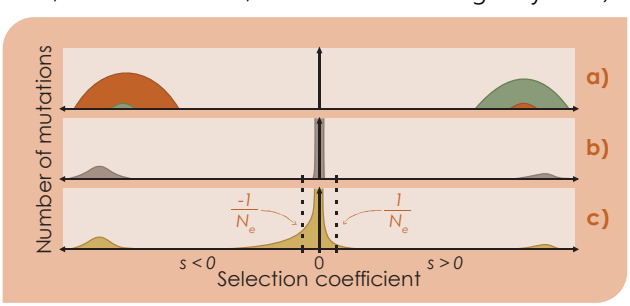

Figure 1
Assumed DFE of the a) classical VS balance hypothesis (orange VS green), b) neutral theory (grey), and c) nearly neutral theory (yellow).

Factors that affect the shape of the DFE

Of course, fitness is relative. A mutation that occurs in one individual may have a different effect in another due to genetic context (interaction between alleles), dominance pattern, physical environment, ecological niche, behaviour, etc.

The classic example of the effect of environment and genetic context on the selective effect of mutations is that of sickle cell disease. Sickle cell disease is caused by a single nucleotide polymorphism (SNP) exchanging one A with a T in position 11p15.5 on chromosome 11 (Pauling et al. 1949, Stamatoyannopoulos 1972). The mutation creates a modified form of haemoglobin A, called haemoglobin S, which causes red blood cells to become thin and spiky in low oxygen conditions, rendering them more prone to breaking (Gordeuk et al. 2016, Herrick 1910, Pauling et al. 1949, Stamatoyannopoulos 1972, Zhang et al. 2016). In heterozygote individuals, symptoms are relatively mild since the default (A) allele can produce haemoglobin A to form regular blood cells, but being homozygous for the mutation often causes grave health problems: chronic pain, regular infections, ulcers, pulmonary hypertension and stroke are among the common symptoms (Gordeuk et al. 2016, Pauling et al. 1949, Zhang et al. 2016). However, the fragile cells have one very specific advantage – it disrupts the reproductive cycle of the *Plasmodium* parasite that causes malaria. The parasite invades the red blood cells as part of its life cycle, but the sickle-shaped cells break before the parasites have matured enough to be able to infect new cells, effectively stopping their reproduction (Allison 1954). Thus, in a context where malaria is prevalent, the sickle cell mutation can be adaptive. While homozygous individuals generally have low fitness, individuals carrying only one copy of the mutation show a 50-90% reduction in the parasite load compared to individuals without the sickle cell mutation while also having few symptoms of the disease (Allison 1954). Outside this environment, however, even heterozygotes are less fit as they derive no

benefit from malaria resistance, and the remaining symptoms are still detrimental to survival and reproduction (Haldane 1990). Thus, the fitness effects of alleles are contingent on the environment in which they are being evaluated.

The real-life effects of the fitness of an individual can, of course, only be construed in relation to the other individuals with which it competes. An individual without the sickle cell mutation would not have a relative selective disadvantage against other individuals even in a malaria-rich environment if no other individual had it either. The absolute fitness reduction that comes with susceptibility to malaria would affect the entire population equally, and the non-sickle cell allele would simply be selectively neutral. Only when the sickle-cell mutation is present is the other allele even under selection, and the direction of selection depends on the incidence of malaria. This goes to show that aspects of the genetic variation, or processes that affect it, within the population where selection is taking place are as relevant to consider as the mutations themselves when we want to understand the DFE. So, what are some of these aspects or processes we must consider?

Effective population size and the efficacy of selection

Consider a population of 100,000 individuals, where half of the population are homozygous for an allele A, and the other half homozygous for an allele B. A and B have no effect on fitness; they are selectively neutral against each other. If all individuals mate randomly, how many generations can we expect it to take before one allele, either A or B, has outcompeted the other, and all individuals in the population are homozygous for that allele? Each generation, each individual instance of each allele has a 50% chance – a coin flip – of being passed on to the next generation, and this happens each time the 100,000 individuals mate. For all individuals to become homozygous, one allele must win all coin flips.

Undoubtedly, in most cases, it would take a very long time before either allele was lost.

Now consider a population of two individuals, one homozygous for A and the other for B. Now, how many generations do we expect it to take before either allele is lost through chance? In the first generation we get two offspring, each with AB alleles; in the next, each offspring has a 50% chance of being AB, 25% of being AA and 25% of being BB. Thus, the chance of both offspring being either AA or BB already in the second generation is 2 * 0.25 * 0.25 = 0.125 or 12.5%. This illustrates the effect of population size on random genetic drift. When developing the neutral model, Kimura showed through diffusion models that the average time to fixation for a neutral allele was approximately $4N_e$ meaning that neutral alleles would go to fixation more quickly when effective population size is low. With few individuals available, a random death of an individual or the random inheritance of alleles will shift allele frequencies more than if the population was larger.

This is a simplified example, but it is the basis of the argument for the nearly neutral model described by Ohta (Kimura & Ohta 1971, 1973 1992). In smaller populations, natural selection will have a weaker effect relative to genetic drift. Thus, stronger selection effects are required to overcome the effects of random chance. This means that in a population with smaller effective population size, more mutations will behave like neutral mutations, and strongly selected mutations will behave as though they were under weaker selection. It also means that the population will accumulate deleterious mutations which could have a negative effect on absolute fitness.

Inbreeding

Inbreeding is when individuals in a population are more likely to breed with relatives than expected under random mating (Gillespie 1998). Just like small

population sizes, inbreeding can affect the efficacy of selection, which can in turn affect the DFE. One result of inbreeding is a reduction in effective population size (Charlesworth et al. 1993, Muller 1932), but not as a direct result of low *census* population sizes (meaning the actual number of individuals. Inbreeding decreases the effects of recombination, meaning that specific alleles at different sites will be inherited together more often. Imagine that, for example, two advantageous mutations arise in two unrelated individuals. If the two mutations occur in the same individual and they give an increased fitness than if an individual has only one mutation. However, if the offspring of the two individuals never inter-breed because of inbreeding, the two mutations will be in constant competition, stopping each other from going to fixation. This is called *Hill-Robertson interference* and is another example of how efficacy of selection is reduced in inbred populations (Hill & Robertson 1966). Similarly, deleterious mutations that are genetically linked to advantageous mutations can also "hitchhike" to fixation as the advantageous mutation is selected for (Haigh & Smith 1974).

Another effect of inbreeding is an increase in homozygosity, since rarer alleles become more likely to occur in the same individual. Thus, dominance effects are also more likely to affect inbred populations than outcrossing populations. For example, it can be argued that highly inbred or self-fertilizing species should be able to purge deleterious recessive alleles at a higher rate than an outcrossing population with the same effective population size due to the higher rate of homozygosity (Arunkumar et al. 2014, Mochales-Riaño et al. 2023). Some empirical evidence seems to support higher purging rates of deleterious alleles in inbreeding populations (though it may not be enough to reduce the genetic load induced by lower N_e) (Mochales-Riaño et al. 2023, Zeitler et al. 2023). Self-fertilization is the most extreme version of inbreeding, as an individual is mating with itself to create offspring. In this case, every site for which the parent is

homozygous will be homozygous in the offspring, and every site for which the parent is heterozygous has a 50% chance of being homozygous for either allele. Thus, selfing species will have high rates of homozygosity and be strongly affected by reduced recombination effects (Arunkumar et al. 2014, Heller & Smith 1978). It is currently unknown exactly how much each of these factors affect the survival of inbred species and populations. Estimating the DFE of inbred and selfing species could therefore be of considerable interest, for example in species threatened by extinction (Mochales-Riaño et al. 2023, Zhang et al. 2023).

Estimating the DFE from genomic data

There are several methods that can be used to estimate the DFE, but the most widely used are based on collecting genomic data from a target population and measuring and comparing the frequencies of neutral and selected mutations (Bataillon & Bailey 2014, Boyko et al. 2008, Johri et al. 2020, Keightley & Eyre-Walker 2007, Piganeau & Eyre-Walker 2003, Tataru & Bataillon 2019, Tataru et al. 2017). By collecting genomic data from natural populations, we are gathering data that contains information about the evolutionary history of that population, representing how mutations have behaved in response to selection pressures specific to that species, area, timepoint, etc. Importantly for us, if we can extract the information therein, we do not need to know the specific mechanisms of selection for each allele. Instead, we can look at patterns of variation to draw conclusions about the DFE among them.

Ultimately only two types of processes can affect the frequency of a genetic variant such as a single nucleotide polymorphism (SNP): *selectively neutral processes* such as random genetic drift, population size change, founder effects etc., or *selective processes* where the allele has a positive or negative effect on survival or reproduction. Neutral processes, by definition, affect the frequency of all alleles,

regardless of whether they have any effect on individual fitness, while selective processes act directly on mutations that alter fitness (Dobzhansky & Pavlovsky 1957). Thus, the combined effects of neutral processes and selection are what shapes which mutations are currently segregating (or have been fixed or lost) in the population. In theory, if neutral mutations are only affected by neutral forces such as genetic drift and population size changes, we should be able to use them as a control for demographic history in our model. By comparing them to the class of selected mutations, we could extract the patterns unique to the selected mutations to draw conclusions about the DFE. While the concept is simple, the underlying assumptions that must be made affect the accuracy of the estimates. The different methods of DFE estimation are described in more detail below, but the general procedure is the same across most of them, that is, collecting allele frequencies from a sample of individuals, dividing the mutations into neutral and selected mutations, and comparing their respective frequency patterns.

The site frequency spectrum (SFS)

The most widely used methods of DFE estimation are based on the frequencies of single nucleotide polymorphisms (*SNPs*) in a population (Bataillon & Bailey 2014, Boyko et al. 2008, Gutenkunst et al. 2009, Johri et al. 2020, Keightley & Eyre-Walker 2007, Tataru & Bataillon 2019). More specifically, these methods use the frequencies of allele frequencies, called the *site frequency spectrum* (SFS), of neutral and selected mutations to estimate the DFE. Briefly, most methods use the following steps to calculate the SFS from SNP data:

SNPs are first classified as neutral or selected. Exactly which mutations are truly "neutral" and "selected" is, of course, not known in most cases, so instead it is assumed that those that are *synonymous*, meaning that they do not affect which amino acid is being produced, are selectively neutral, while those that are *non-*

synonymous, i.e. change which amino acid is produced, can be affected by selection. However, there are cases where some mutations change the amino acid but not others (say a change from an A to a C, but not from an A to a T. For computational efficiency, only sites where all variants are synonymous are counted as putatively neutral. These are called fourfold degenerate sites – all four nucleotides that can occur at this site will produce the same amino acid. Similarly, only sites where none of the possible nucleotides produce the same amino acid are counted as putatively selected – these are called zerofold degenerate sites.

Secondly, allele frequencies are recorded for all fourfold (neutral) and zerofold (selected) degenerate sites. In a sample of 100 diploid individuals (N=100), the maximum frequency of any allele would be 200 (=2N), meaning that all individuals carry 2 copies of the allele. However, since polymorphic sites will contain more than one allele, only the frequency of one allele is recorded. Ideally, we wish to use the newer ("derived") alleles, while the frequencies of the older ("ancestral") alleles are excluded, to represent the frequency distribution of new mutations. This categorization of derived/ancestral mutations is generally done by comparisons with a reference genome (often from a species that shares a relatively recent common ancestor) where the allele that occurs in the reference genome is deemed ancestral and the alleles unique to the sample population are classified as derived. In more complicated cases, several reference genomes may be required to classify all alleles. If no suitable reference genome is available, however, the minor allele, i.e. the allele with the lowest frequency, is recorded.

Lastly, we calculate a *site frequency spectrum*, or SFS, for the fourfold (neutral) and zerofold degenerate sites, respectively. The SFS is the vector describing the number of sites that have an allele frequency of 0, 1, 2, [...], 2N. If the included alleles are classified as derived or ancestral using a reference genome, alleles may occur in frequencies all the way up to 2N. This is called an unfolded SFS. If we use

the simpler method of only including the minor alleles, the highest frequency any allele can take is N (representing 50% of all alleles at that site), and the rest of the SFS will consist of zeroes. This will create a folded SFS – any derived allele that occurs at a frequency f>N will be represented in the SFS as a frequency 2N-f instead, effectively folding the second half of the SFS onto the first. While this method may seem simplistic in comparison with the unfolded SFS, many DFE estimation software currently use the folded SFS with seemingly good results.

The empirical SFS of neutral and selected mutations are then compared to the expected SFS under different demographic scenarios, as well as under different DFEs of selected mutations. In general, we can use the assumption that mutations that are more deleterious will occur at lower frequencies than less deleterious alleles to estimate the number of mutations in the sample that can be assumed to have different fitness effects. The exact models used for calculating the expected SFS differ among software, but most use Fisher-Wright transition matrices (Keightley & Eyre-Walker 2007) or Poisson Random Field models (Boyko et al. 2008, Kim et al. 2017, Tataru & Bataillon 2019, Tataru et al. 2017) that incorporate some level of possible population size change together with the estimated DFE.

Shapes of the estimated DFE

The exact shape of the DFE is arguably unique to each species, population and evolutionary context. Yet, some assumptions about the possible or most common DFE shapes are required for statistical inference from genomic data. Software for DFE estimation will therefore produce different possible distributions depending on the underlying model assumptions. For example, many software will assume that the DFE follows a gamma distribution (Boyko et al. 2008, Keightley & Eyre-Walker 2007, Kim et al. 2017, Piganeau & Eyre-Walker 2003, Tataru & Bataillon 2019). The gamma distribution is described by two parameters, commonly the

shape and scale parameters, or possibly reparametrized as the shape and mean ($scale^{-1}$) parameters. Depending on the value of these parameters, the gamma distribution can take many different forms – from nearly exponential, to unimodal, to almost flat – which makes it versatile for describing the DFE. For example, the Nearly Neutral Theory suggests that most mutations are selectively nearly neutral or weakly deleterious, with a decreasing frequency of mutations with stronger and stronger deleterious effects. This could easily be described as a gamma distribution with mean and shape parameter values near 0. However, there is no theoretical guarantee that the DFE cannot take other shapes, with bi- or multimodal distributions. In these cases, assuming a gamma distribution would impede the accuracy of the estimated DFE (Kousathanas & Keightley 2013). Some models use discretized distributions that can take any shape, but these models must often make other concessions, such as not being able to control for population size changes due to computational complexity (Wilson et al. 2011).

Most software currently only estimate the neutral-to-deleterious part of the DFE, and opt to estimate any effects of positive selection by other means (Bataillon & Bailey 2014, Eyre-Walker & Keightley 2009, Kim et al. 2017, Tataru & Bataillon 2019, Tataru et al. 2017). This, again, is due to the assumption that advantageous alleles are rare, especially among segregating alleles as they are more likely to fix more rapidly than neutral or deleterious alleles. By focusing only on the deleterious half of the DFE, we can use a gamma distribution to describe it, and avoid the problem of having to distinguish between mutations that are positively and negatively selected among the non-synonymous sites. Instead, the proportion of advantageous substitutions (mutations that have become fixed), α, is often estimated by calculating the expected number of deleterious substitutions under the estimated DFE and population size, and comparing it to the number of fixed sites in the neutral and selected SFS (Eyre-Walker & Keightley 2009).

Our studies

As a first step, I was to apply the method of DFE estimation from site frequency spectra on a real data set where the DFE would give valuable insights into the process of speciation. The Tibetan pine, *Pinus densata*, represented an excellent candidate; *P. densata* is a hybrid species that shows strong signs of having undergone recent selective adaptation to the harsh environment atop the Tibetan plateau (Mao & Wang 2011). It is also, as mentioned, a hybrid species with ancestry from both *P. yunnanensis*, and *P. tabuliformis* (Wang & Szmidt 1994, Wang et al. 2001). The data set was derived using exome capture methods, thus not including intronic regions (regions outside of genes. Although the data would not be suitable to analyses requiring very long gene regions, we assumed that a SNP frequency-based approach would be appropriate. We selected the software DFE-alpha (Eyre-Walker & Keightley 2009), a well-used software which uses the frequencies of the least common alleles at each site to estimate the DFE and got to work.

Problem 1: Data quality

One issue presented itself when my colleague Wei Zhao and I had both performed the DFE analysis on five populations of *P. densata*. His data sets included a few more individuals than mine, but preparation was otherwise identical. My results showed that the DFE seemed to include mostly mutations with relatively weakly deleterious effects, and only ~3% of mutations were classified as strongly deleterious. Wei's results, however, indicated that over 30% of all mutations were strongly deleterious in some of the same populations. This was worrying. If the inclusion of just a couple of extra individuals in a population could change the estimated DFE to such an extent, how would we possibly know which – if any – results to trust? We could only assume that this was a result of some shortcomings of the data or the method. Either the data sets were flawed (biased, too small, etc)

and one or both of our samples did not represent the true variation in the population, or the method was misinterpreting some pattern within it. Trial tests with different sample sizes and pre-processing methods further confirmed that the variations we could get with the same starting data set were – regrettably – rather large. Maybe more worryingly, our data set (while not massive or the most advanced) did not seem to be of obviously worse quality than many others where DFE-alpha or similar methods had been applied, implying that other studies may unknowingly have encountered this effect (see for example Kutschera et al. (2020). I could find no guidelines or recommendations about either the number of SNPs or the number of individuals needed to obtain reliable results from DFE-alpha. Instead, we designed an experiment to make them.

Our hypotheses were that either: 1) The data set contained too little information (too few individuals or sites) for DFE-alpha to be able to draw any conclusions, or that 2) including or excluding some individuals or sites affected the pattern of SNP frequencies in final data set after filtering. By downloading a larger data set and creating smaller sub-data sets by sampling from it, we should be able to show which factors could affect the DFE estimates from DFE-alpha. Using data from *Arabidopsis thaliana*, we created data sets of different sizes, using different filtering methods and settings, and estimated the DFE from each of them. A year after the first discovery, the results came in: estimates of the DFE from the same populations were wildly different depending on sample sizes, and even the method used to filter the data prior to analysis had a large effect. But the feeling of retribution did not last long…

Problem 2: Inbreeding

Unfortunately, we had glossed over an important biological factor – inbreeding also violates an assumption of DFE-alpha, and *Arabidopsis thaliana* is an almost

exclusively self-fertilizing species (Abbott & Gomes 1989, Alonso-Blanco et al. 2016, Bechsgaard et al. 2006). The few studies that have attempted to estimate the error in high selfing populations show steep drop-offs in accuracy (Gilbert et al. 2021). We also identified an interesting pattern; in the *A. thaliana* data set, almost all alleles occurred in an even number of copies in our data set. This, in itself, is not necessarily surprising because inbred species will have a high rate of homozygosity (individuals carrying two copies of the same allele instead of two different alleles), but it could cause problems since DFE-alpha uses allele frequencies for estimating the DFE. A diploid species where all individuals are homozygous for an allele will contain no sites with an odd allele frequency, meaning that the SFS will have a frequency of 0 for all odd frequencies. So instead of solving a problem, we had swiftly found another!

Problem 3: The truth?

To test the effects of sample size and filtering without the confounding effect of inbreeding in *A. thaliana*, we instead downloaded a similar data set from a related outbreeding species, *A. lyrata*, and performed the same analyses again. This time, we could see the effects of sample size clearly, as well as how different filter methods could alter the estimated DFE. Yet, there was a nagging question that we, and soon enough a couple of reviewers, had identified. When we analyse the different DFE estimates, we can see that they vary – but how do we know which estimate is *better*? Unless we know what the real DFE is, we can never truly know which of two estimated DFE's are more accurate. At best, we could see which result is most common, and perhaps assume that the true DFE is somewhere around there, or an average of several different trials. However, there is no guarantee that that would be the case, as DFE-alpha's interpretation of data sets with certain characteristics could be biased in one direction instead of spreading equally around the true value. We needed a data set where we knew what the true DFE

was, so that we could measure the accuracy of the estimates, and not just the variance between them…

Solution: Simulation

So, what do you do if the perfect data set does not exist? You create it. By simulating populations where mutations were drawn from a known DFE, and being able to modify aspects such as mutation and recombination rate, demographic history and the rate of selfing, we could simulate populations that followed all of the assumptions of DFE-alpha save for the exact variable we wanted to study. Further, we were able to create data sets that mimicked an empirical data set. This made it possible to estimate a baseline accuracy estimate for DFE alpha. In this way, we could show the effects of both sample sizes, filtering missing data, population structure and level of inbreeding. Together with the results from the empirical data sets, we could finally quantify the effects of each factor and show which methods should be used in the most common scenarios.

Finishing at the start (spoiler warning)

While I aimed to study the speciation process more directly, the problems we found along the way turned out to be both interesting and very valuable. The conclusion drawn from our explorations of the accuracy of DFE-alpha in different contexts informed our choices in future studies. Firstly, one conclusion we could draw from our results was that, since DFE-alpha does not use information such as linkage disequilibrium or relatedness among individuals, it requires very large numbers of alleles to draw conclusions from allele frequencies. Thus, we deemed that the data set from *Pinus densata*, a study which had been put on ice during this exploration phase, was probably best suited for other types of analyses. In its stead, Jing-Fang Guo and others (2023) were able to describe extensive population structure, identify potentially introgressed alleles involved in adaptation to

elevation and reproductive isolation using methods that leveraged information such as associations between ancestry information, environmental variables and the relationship between genetic variation in the exome and intergenic regions. Secondly, we showed that if we can identify the general pattern of low quality/missing genotypes in a data set, this can inform our choice of filtering method for pre-processing. Thus, in the latest study on adaptive radiation in the *Pinus* complex, we chose to change the filter method used from subsampling to downsampling, and made sure that all samples included >8 individuals and >1 million sites (in this case ≥12 individuals and >5 million sites) to ensure that the results were reliable. We will also continue to explore methods of estimating the DFE in inbred species.

Aims & objectives

This thesis aims to explain and explore one of the most common methods of quantifying the distribution of fitness effects from whole genome data sets. The main part of the thesis will focus on assessing the accuracy and robustness to deviation from core assumptions of the underlying models of one specific method of estimating the deleterious DFE (DFE-alpha. The second part will show how DFE-alpha and/or other methods can be used in context with other analyses for assessing fitness effects of mutations to inform conclusions about species history. Specifically, I focused on the following questions:

1. How does variation in data quality and/or quantity affect the accuracy of DFE estimation? (**Paper I**)
2. How does variation in life-history traits, specifically inbreeding, affect the accuracy of DFE estimation? (**Paper II**)
3. When and how can the estimation of the deleterious DFE be used to draw conclusions about evolution in natural populations? (**Paper I-IV**)

Data and methods

Simulated data sets (Papers I, II)

To test the accuracy of DFE-alpha, we created artificial genomic data sets with a known DFE using SLiM 4.0.1. SLiM produces genomic data sets by simulating populations of individuals with separate genomes for which mutations and recombination occur at given rates. In what I will hereby refer to as our default model, we simulated uniform populations of 10,000 individuals with 50Mb genomes and recombination rates of 4×10^{-8}. The DFE of selected mutations was a gamma distribution including only deleterious mutations, which follows the assumption made by DFE-alpha. Generally, selected mutations are generated within the simulation, while neutral mutations are added later by tracking the ancestry of the individuals at the point of sampling, since neutral mutations do not alter the likelihood of survival or reproduction. We used a ratio of 4:1 of selected to neutral mutations (based on the ratio of non-synonymous to synonymous sites in *Arabidopsis lyrata*), giving a mutation rate of 7×10^{-8}. Running this type of simulation with an outcrossing population should produce close to ideal data sets for DFE-alpha with minimal error, save for the assumption of complete recombination (no linkage.

To best replicate the procedure used in analysis of empirical data, we created VCF files based on VCF files from 100 randomly sampled individuals from each simulated population. This method preserves diploid genotypes unlike the built in SFS function in SLiM which generates SFS from random haploid genomes in the population. This becomes relevant where homozygosity can influence the results (Paper II.

Paper I: We simulated one population of 10K individuals with a strongly deleterious gamma DFE (mean selection coefficient *Es* of -100, shape *b* of 0.1. We sampled 10×100 individuals for which we created VCFs and SFS, for 4 of which we also randomly masked 20% of the genotypes DFE estimate accuracy after filtering.

Paper II: We simulated 16 separate populations of 10K individuals with self-fertilization rates of either 0%, 50%; 60%, 70%, 80%, 90%, 95%, 99%, or 99,9%, and either a weakly deleterious (mean -0.001, shape 0.1) or slightly stronger deleterious (mean -0.01, shape 0.1) DFE. We sampled 100 individuals per population to make the SFS.

Missing data filtering (Papers I, IV)

Empirical genomic data sets invariably contain genotype reads of lower quality and/or those that are completely missing. These reads are often removed to ensure accuracy of the data. Calculating an SFS from SNP data, however, requires the number of available genotypes to be the same for all sites. Since DFE-alpha uses SFS for its analyses, the missing genotypes must then be either filtered out or somehow reintroduced before we can estimate the DFE from our data sets.

Three popular methods of handling missing genotype data are downsampling, imputation and subsampling (Note: that the names down- and subsampling are not consistent across studies but are used here to enable discussion about them without confusion) (Fig. 2). Downsampling and subsampling both work by removing a portion of the data to match the sites with fewer available genotypes, while imputation "fills in" missing genotypes based on the allele frequencies in other genotypes. For each method, we also tried several different settings to get an overview of the variability in outcomes within each method.

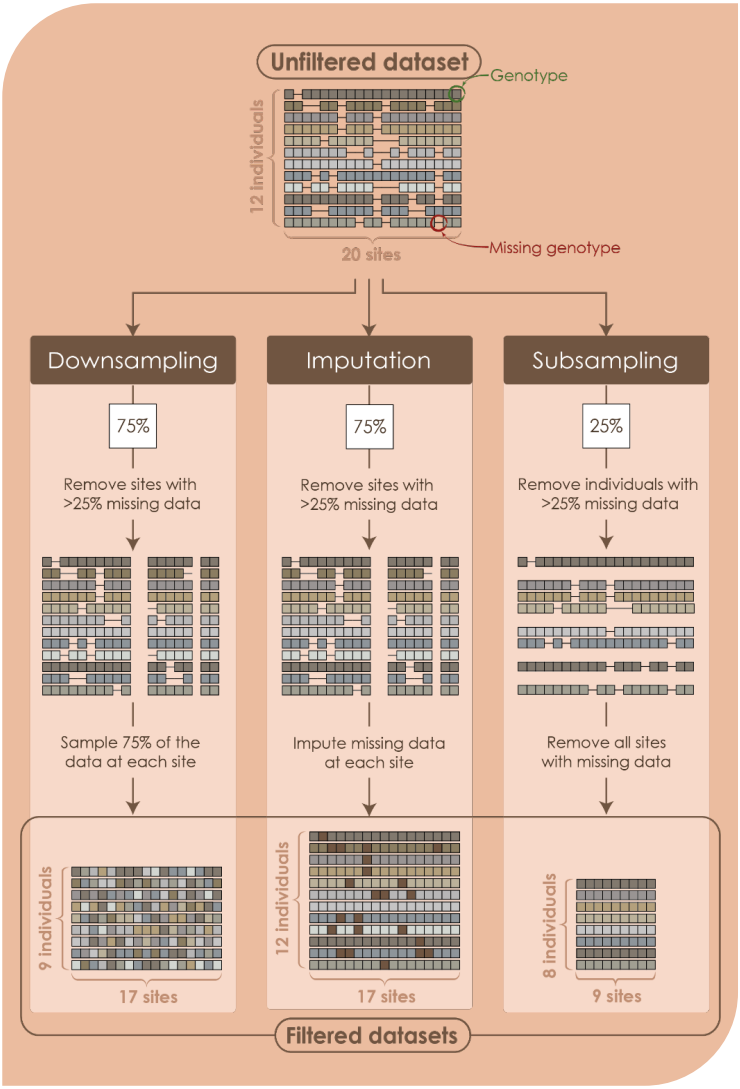

Figure 2
Three methods for treating missing data were examined in paper 1: *Downsampling*, *Imputation* and *Subsampling*. The figure presents an example of the resulting sizes and characteristics of the filtered datasets under each method. For example, downsampling does not preserve individual identities, and the dark brown squares in the imputed data set indicate genotypes that have been added by imputation.

Downsampling creates a data set by sampling a set number, say *n*, of genotypes for all sites. Sites for which there are not enough genotypes available to sample are removed. In this way, downsampling creates *n* genomes with no missing data, comprised of genotypes from different individuals, but excluding sites where the data quality was low. As such, just like in imputation, a threshold of 90% in downsampling will remove all sites for which more than 10% of the genotypes are

missing, but it will also mean that the number of genotypes sampled in the remaining sites will be 90% of the total number of individuals sequenced.

Imputation works by comparing our set of genomes with a reference genome and identifying IBD (isolation by distance) blocks that have been inherited more or less identically from a common ancestor, together with a hidden Markov model that accounts for the uncertainty of the inferences (Browning et al. 2018). This gives rise to a data set where missing genotypes are filled in with the most likely allele for each individual and site. The threshold in imputation is simply a limit of the lowest required data quality for sites to be included in the analysis, or reversely, how high the proportion of missing data can be at site before we consider the possibility of incorrect imputation to be too high and so choose to remove the sites from the analysis. Thus, a threshold of 90% will remove all sites for which more than 10% of the genotypes are missing and impute the genotypes for all other sites.

Subsampling is the simplest method of removing missing genotype data, but it also removes the largest amount of data in the process. In subsampling, the order of filtering is reversed, by first filtering out individuals with a high missing rate. Thus, a threshold of 10% would mean removing all individuals with a missing genotype frequency above 10%. As a next step, all sites where at least one of the remaining individuals have a missing genotype are removed.

Paper I: Empirical datasets from two populations of *Arabidopsis lyrata* were filtered with imputation at thresholds 90%, 80% and 70%, downsampling at thresholds 75%, 66% and 50%, and subsampling at thresholds 10%, 15%, 20% and 25%.

Paper IV: Missing data was filtered using downsampling at an 80% threshold.

DFE estimation with DFE-alpha (Papers I, II, IV)

The DFE was estimated with the software DFE-alpha (Keightley & Eyre-Walker 2007). DFE-alpha uses the SFS of putatively neutral sites to estimate a demographic model, and the SFS of putatively selected sites to estimate a gamma distribution that describes the distribution of selection coefficients of new mutations. The demographic model assumes one population size change from a size N_1 to N_2 individuals at a time t_2. The population sizes are relative; N_1 always starts at 100 individuals, and N_2 can take any value between 0 to 1,000, where 1,000 thus indicates a 10x population size increase. Together, these are used to create a distribution where the proportions of mutations with selective effects scaled by the effective population size, $N_e s$, are estimated. Mutations are categorised as either effectively neutral ($0 \leq -N_e s < 1$), slightly deleterious ($1 \leq -N_e s < 10$), moderately deleterious ($10 \leq -N_e s < 100$) or strongly deleterious ($-N_e s \geq 100$. Under this categorisation, mutations with an absolute $N_e s$ value below 1 will behave neutrally since the effect of random genetic drift is stronger than that of selection. Yet, the information derived from this distribution is contingent on the accuracy of both the estimated gamma distribution and the assumed demography.

Paper I & II: DFE-alpha was run with default parameters under the two-epoch demographic model. For simulated datasets the accuracy of the estimated gamma distribution (describing the distribution of selection coefficients) were evaluated using Earth Mover's Distance, comparing the overlap of the different estimated DFEs in each dataset with the true DFE given in SLiM. In all datasets, 95% confidence intervals were calculated using bootstrapping (99 iterations), and the $N_e s$-scaled DFEs (mutations divided into four categories, see above) were plotted for comparison.

Paper IV: DFE-alpha was run separately on each population, and 95% confidence intervals (999 iterations) were calculated and plotted.

Our results and their implications

Data quality

In any empirical science, the quality of your data unavoidably influences the quality of the results of any analysis you perform on it. Larger sample sizes are better – but often considerably more expensive – and avoiding bias is an implicit struggle in population genomics where data sets are so large that patterns may not emerge until after several rounds of processing. Better yet, genomic data sets need several steps of pruning, filtering and interpretation before it is ready for most statistical analyses. How then do we ensure that these steps themselves do not induce bias, or even reduce the quality of our data? Among the myriad statistical methods, independently developed software packages and filtering pipelines being made, we aimed to quantify the effects of one step in the processing of genomic data, specifically needed before DFE analysis with SFS-based methods: filtering of missing data.

Since the SFS describes the population-level frequencies of mutations across a number of genomic sites, it requires that all individuals in the sample contain complete genotype information for all included sites. Say, for example, that a mutation A that is, in reality, present in 10 copies in the population and allele B in 12, but that the genotype of one individual homozygous for allele A had a low coverage (its genotype uncertain) in the sequencing step and was excluded from the final data set. In an unfolded SFS, this site would count towards the mutation with a frequency of 8 instead of 10, while the other allele B at the site would retain its true frequency of 12. Any missing information would thus shift the allele frequencies downwards in sites with missing genotypes, inducing an artificial bias towards low-frequency alleles.

To avoid this, SFS must include only sites for which we have data for an equal number of genotypes. How to arrive at which sites and which genotypes to preserve from the whole data set can, however, be difficult. In our study, we chose to highlight three methods which we called subsampling, downsampling and imputation. In essence, subsampling removes individuals with high missing rates first, and then removes all sites with missing data; downsampling samples a set number of (random) genotypes at each site, removing sites for which there is not enough data; imputation removes sites with high rates of missing data first, and then infers the most likely genotype for each missing genotype at the remaining sites using linkage information. In order to show which filtering method effects were due to the reduced size of the data set, we also performed several tests on data sets with different numbers of individuals and sites in simulated data sets where there was no missing data to start. The filter methods were performed both on simulated data sets with random genotypes removed to simulate missing data, and on an empirical data set (*Arabidopsis lyrata*.

In our experiments, the accuracy of the DFE estimate was correlated with the size of the data matrix, quantified as the total number of SNPs included in the 0-fold and 4-fold SFS, both in complete data sets and after filtering. While the general result of "more data gives better results" may be intuitive, the extent of these effects and the relative importance of the different dimensions of sample sizes have not been quantified before. The SFS is made up of a vector of frequencies, where the length of the vector represents the number of haploid genomes in the population (= the maximum frequency), and the frequencies listed in the vector depend on the number of sites included in the analysis. Both factors (number of individuals and sites included in the analysis) are important but seem to affect the accuracy of the results in subtly different ways.

Firstly, our results indicate that while DFE estimates generally become more accurate in samples with more individuals, the accuracy is disproportionately low in data sets with 8 or fewer individuals. This effect is strongest – and clearly visible in the difference between the estimated proportions of fitness effect and the size of the 95% confidence intervals – in samples with only 4 simulated individuals (Fig. 3a-c), and is suggested by the deviation in estimated DFE in the empirical sample (Fig. 3e. However, the variation in the estimated mean of the DFE for data sets with 8 simulated individuals (from -4 to -6.36×10^6, expected -100; Table S2, Paper I) indicate that these estimates may also be more sensitive to the choice of individuals than larger samples. It should be noted that the error induced by the small sample size is not always accompanied by an increase in the 95% CI's in our study. Instead, we may get a result indicating relatively high confidence in the estimate. Keightley and Eyre-Walker (2007) note in their original test of DFE-alpha's accuracy that the estimates of the distribution mean are more variable than estimates of the shape parameter.

Secondly, data sets with few sites show a clear increase in the size of 95% CI's (Fig. 3b in paper I. We tested the effect of extracting 1,000, 10,000, 100,000, 1 million, 10 million and using all 55 million sites in the data set. Of course, most of the information in the SFS comes from the number of segregating sites, and in these data sets the number of SNPs corresponded to roughly 1.1% of the number of sites included. Although the magnitudes of differences between the levels in this trial are much larger than when comparing the number of individuals in the sample, this range of data set sizes is represented in the literature (Eyre-Walker & Keightley 2007, Hämälä & Tiffin 2020, Keightley & Eyre-Walker 2007). Thus, knowing that comparing data sets with 15,000 SNPs to those with 5,000 SNPs showed the highest measured error increasing threefold should be of some value when reading and designing studies on the DFE. Data sets with fewer than a total of

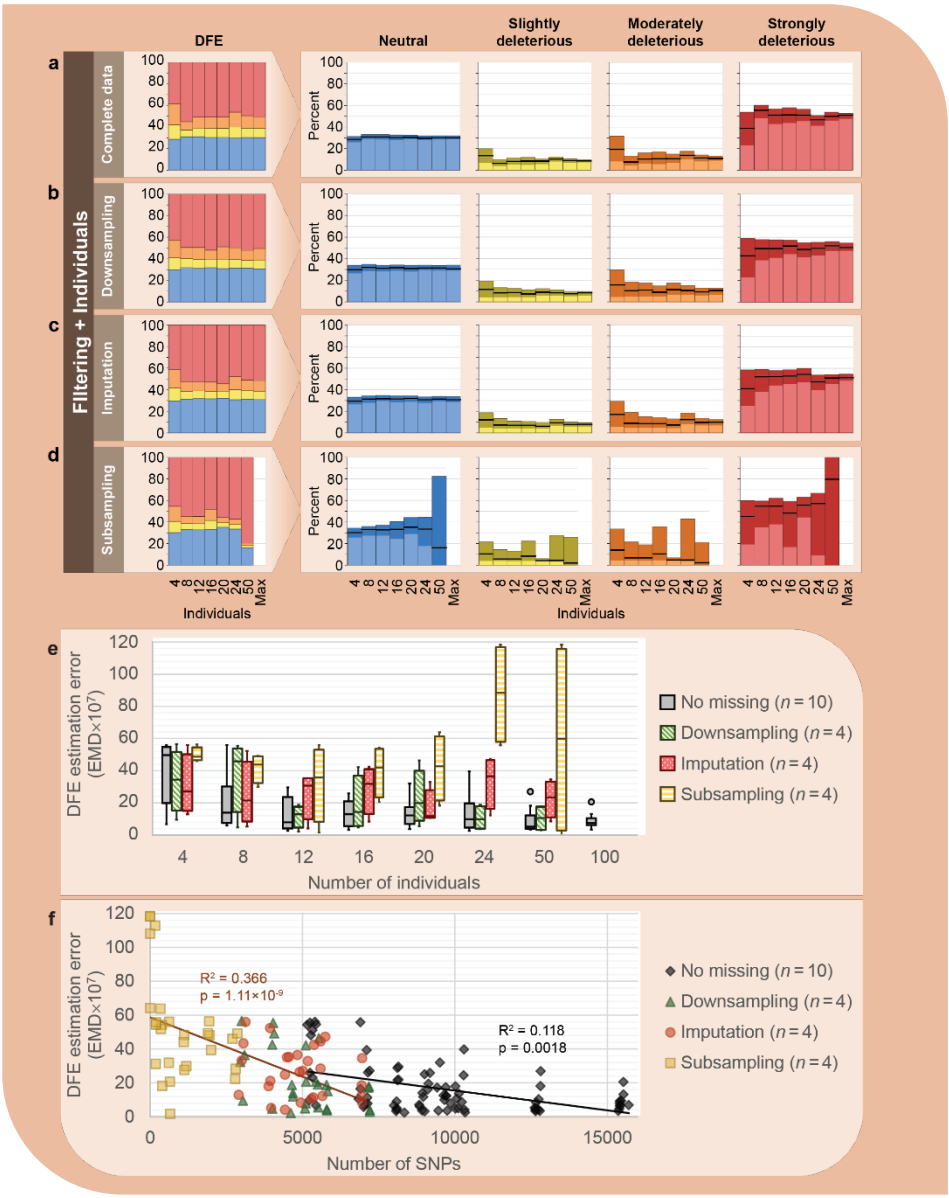

Figure 3 Results from paper I demonstrate the variability (a-d) and accuracy of estimated DFE (e) as a result of number of individuals in data sets without missing data, as well as in data sets with missing data filtered in three different ways. The correlation between accuracy in DFE estimates and number of SNPs in each data set is also presented, with regression lines shown separately for the group of complete data sets (black line) and filtered data sets (brown line) (f).

1,000 SNPs increased the inaccuracy and uncertainty of the results. Further, the resulting size of a data set filtered with subsampling could be up to 90% smaller per individual/loci.

Taken together, these results show that choosing the appropriate filter method for your data set is important when performing SFS-based DFE analysis. The size of the resulting data matrix after filtration may differ significantly depending on the method you use, and different methods are sensitive to different kinds of missing data patterns. For example, subsampling will remove a higher percentage of the data if the missing data is evenly dispersed across individuals and loci, because all sites and individuals with missing data will be removed in one of the two steps, while downsampling retains the same amount of data as long as the missing rate per locus is lower than the threshold. If, however, the missing data is concentrated to specific loci or individuals, subsampling can remove those individuals or loci and continue to use all available data for the remaining individuals and sites. In this case, subsampling can retain more data than downsampling, which will only use a fraction of the available data even for sites where all individuals have complete data. Thus, mapping the pattern of missing data could potentially help in choosing an appropriate filtering method and threshold.

Inbreeding/selfing

While low data quality or small sample sizes can cause reduced accuracy in many empirical analyses, these issues are potentially remediable. Additional data collection and/or using newer or other methods of sequencing (such as combining methods of short and long-read sequencing) can complement the existing data and improve the quality of the data set for more accurate conclusions. It is another matter when the source of the inaccuracy in analyses is part of the life-history of your study species. Inbreeding and self-fertilization mating strategies have long

been seen as evolutionary dead-ends. Inbred species, by definition, should have lower genetic diversity than their outcrossing counterparts leading to lower evolutionary potential in changing environments and lower effective population sizes, meaning lower efficacy of selection for both purging deleterious variants and selecting for beneficial mutations (Charlesworth et al. 1993, Heller & Smith 1978, Muller 1932). According to *Muller's ratchet*, high rates of inbreeding should thus lead to accumulation of deleterious variation, reducing overall fitness until extinction (Muller 1932). Yet, many species of highly prolific organisms such as protists, fungi, nematodes and plants use self-fertilization as their main mode of reproduction (Abu Awad & Roze 2020, Alonso-Blanco et al. 2016, Gilbert et al. 2021, Gossmann et al. 2010). As we are entering times of great climatic change, the viability of selfing species is again a hot topic in molecular biology research. With today's molecular method and analysis tools, we can finally study genetic variation at scale, and the results are intriguing, if somewhat contradictory. Some studies suggest that purging of deleterious mutations might be higher in inbred species, in some cases, reducing their genetic load (Abu Awad & Roze 2020, Leon-Apodaca et al. 2023, Mochales-Riaño et al. 2023) – for example, the increased rates of homozygosity in inbred species might be exposing recessive deleterious alleles to selection (Mochales-Riaño et al. 2023). However, others estimate that this effect is negligible (Zeitler et al. 2023). Some studies suggest that the levels of genetic diversity necessary for population persistence might be lower than previously expected, while yet others suggest that genetic variability is more important for population survival than the accumulation of deleterious variation. To top it off, even basic assumptions about recombination or mutation rates might be affecting how we interpret the risk of extinction due to inbreeding (Sianta et al. 2022).

With these unique challenges faced by inbred species, it seems only logical to want to quantify the distribution of fitness effects to understand their evolutionary strategies and, potentially, their futures. Unfortunately, high rates (≥99%) of self-fertilization have already been shown to severely reduce the accuracy of DFE estimates (Gilbert et al. 2021). In our data sets, selfing increases homozygosity in the population, which in turn increases the relative proportions of alleles that occur at even-numbered frequencies in the SFS (Fig. 4). If high homozygosity affects the accuracy of the estimated DFE, sampling only one allele per site as if only one haplotype per individual was sampled could mitigate this effect. Indeed, this is a method that has been used in selfing species (see e.g. Hämälä & Tiffin (2020). Whether this adjustment actually makes the estimate more accurate has, however, not yet been confirmed.

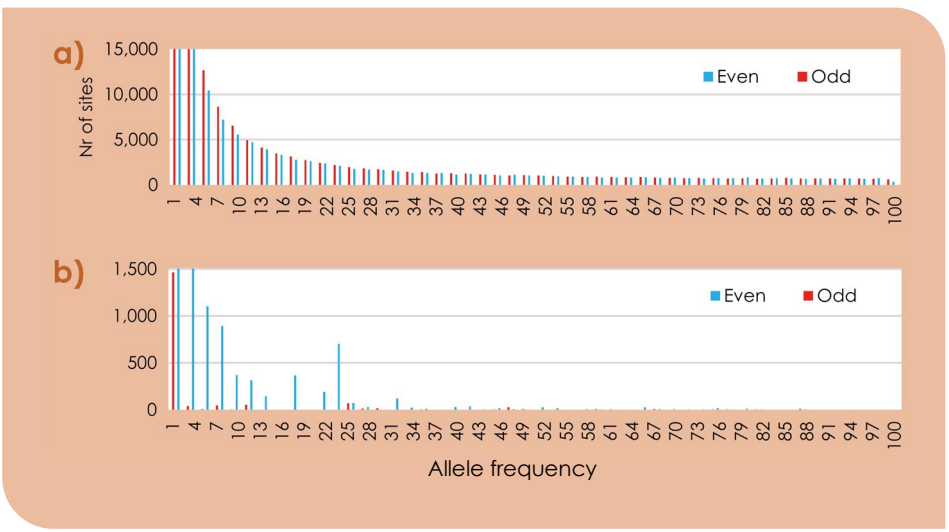

Figure 4 Site frequency spectra for an (a) outcrossing and (b) 99.9% self-fertilizing population with DFE -0.01 and shape 0.1. The blue bars mark the frequency of alleles that are found in an even number of copies within the sample, while the red bars show the frequency of mutations that occur at each of the odd frequencies. Y-axis cut off at 15,000 and 1,500 for visibility.

Our results suggest that while populations with rates of >99% selfing do indeed show a low accuracy in DFE estimates as predicted by Gilbert et al. (2021), the effect seems to start at rates around 70% and gradually increase until around 90% selfing (Fig. 5). At 90% selfing and above, the accuracies of the DFE estimates are almost equally low. Thus, even relatively low rates of selfing affect the accuracy of DFE estimates with DFE-alpha. What is more, using the "haploidized" data set (using one allele per site per individual instead of diploid genotypes) did not necessarily improve the accuracy of the DFE estimate, and the average reduction in the error was generally small. This suggests that some factor other than homozygosity affects the accuracy of DFE estimates in inbred populations, such as the effects of reduced recombination (Sianta et al. 2022, Soni et al. 2023). It is possible that another underlying model of calculating the expected SFS is needed if the assumption of random mating is violated, where the rate of inbreeding should be incorporated explicitly before the DFE is estimated (Blischak et al. 2020).

Figure 5 DFE estimation error (EMD) in simulated populations with different selfing rates when using a haploid (orange) VS diploid SFS (green), estimated in two different DFEs (shape 0.1, mean -0.001 or -0.01.

Concluding remarks

In my quest for understanding the DFE, I ended up studying how we misunderstand it in various ways. This may sound like a failure but to know the pitfalls and difficulties of any topic or procedure is, in my opinion, a vital part of learning how to do it correctly. We have shown that the effects of low sample size and filtering methods, as well as inbreeding, can "trick" DFE-alpha into giving us estimates that deviate from the real DFE, by a small or large margin. However, this discovery is ultimately a step towards increased understanding. At this stage, we know more about the required sample sizes (≥12 individuals and >5 million sites) of DFE-alpha (Andersson et al. 2023). We also know when effects of selfing start to decrease the accuracy of estimates (detectable above 50% selfing, but strongest effects above 80%), and that it tends to skew results towards overestimation of slightly deleterious effects. We were also able to use these guidelines in our own research – both by adjusting the method we use for filtering and in the choice of sample sizes to ensure the best quality results (Zhao et al. Unpublished), and in knowing when not to apply DFE-alpha at all (Guo et al. 2023).

Today, the technology is rapidly evolving and the quality of both data and statistical methods are likely to improve. As such, I do not expect our results to be applicable to all research in the future. However, its relevance to current and past research is enough to justify its value. The guidelines we have developed, and that I hope to continue to develop going forward, I hope will help others obtain more reliable DFE estimates, as well as provide a way of judging the reliability of such analyses.

Acknowledgements

Thank you, Xiao-Ru Wang and Åke Brännström, for being great supervisors and letting me pursue my interests through all the issues and changes. With your encouragement I've been able to learn many new skills and develop my understanding evolutionary theory while of trying to reel in a project that took us in a direction none of us had planned – and it worked out!

Thank you to the staff at HPC2N for the help at all stages of computation, data analysis and script writing. Your assistance and resources made this work possible, as it has for so many other important scientific works.

Thank you, Wei Zhao, for all the painstaking work you've done to fix any and every mistake and detail to do with every data set, and for all of the discussions that were needed to disentangle the complicated methodology questions we had along the way.

Thank you, Ben Haller, for the introduction to and support in using SLiM for our simulations, and for agreeing to host the first ever SLiM workshop at Umeå University! I am lucky to have been able to witness this software start to take over the world of population genetics, and I hope to see much more. Thanks to you, Peter Ralph and the rest of the community, the potential research question that can now be answered are innumerable – and it is my goal to tackle at least a few of them in the future.

Thank you, Alisa, for being a solid friend and PhD companion from the day we met; whether either of us was excited, frustrated or confused about work or life, it was always good to know there was someone who would understand and listen to any rant or deep discussion that needed to come out! One day we will publish a paper together, however convoluted the premise might need to become…

Thank you to Jade, Julia, Jenny and Alex for being stellar roommates at different time-points, and for making the office a nice place to be, even though I was M.I.A. for a lot of the time! Thank you to the entirety of Xiao-Ru's research group as well as the Journal Club gang for being welcoming and including me in great scientific discussions on everything from pine genetics to the impending AI takeover. And thanks to all of the other staff and PhD students at EMG for creating a great environment during teaching, courses and the occasional skiing trip!

Thank you, Barbara, for being a great mentor and friend throughout my educational journey, and for the encouragement to pursue my interests at all junctions. Every cup of tea has come with great life advice that I will carry with me always.

Thank you, Bent, for inspiring 14-year-old me to go into biology in the first place, and for first awakening the idea that I could pursue a PhD. One day I *will* steal your job!

Thank you, mom, for being my sounding board, test audience, life coach and biggest cheerleader at all times. No self-help book or course can ever rival your ability to get right down to the issues and rebuild my motivation when I get stuck. Thank you also to my siblings, Marc, Cornelia and David for always cheering me on and believing in me. Love you all.

Thank you, Nils, Johannes and Ludvig, for getting me out of work-mode to have fun once in a while and giving this anxious hermit a social life again! One day this biologist will get a birdie…

Thank you, Oskar, for your love and patience, for always being there to help me get back on the wagon whenever I fell off – and for kicking my butt when I needed it. I promise to do the same for you when it is your turn, wherever I am.

Thank you to Alicia, Maria, Ida, Mathilda, Rikard and Malin, for listening to my complaints and garbled infodumps at different points throughout my PhD. Some of you may be far away, but the effort to not tell me to just shut up is ever appreciated.

References

Abbott, R. J., & Gomes, M. F. 1989. Population genetic structure and outcrossing rate of *Arabidopsis thaliana* (L.) Heynh. *Heredity*, *62*(3), 411-418. https://doi.org/10.1038/hdy.1989.56

Abu Awad, D., & Roze, D. 2020. Epistasis, inbreeding depression, and the evolution of self-fertilization. *Evolution*, *74*(7), 1301-1320. https://doi.org/10.1111/evo.13961

Allison, A. C. 1954. Protection afforded by sickle-cell trait against subtertian malarial infection. *The British Medical Journal*, *1*(4857), 290-294.

Alonso-Blanco, C., Andrade, J., Becker, C., Bemm, F., Bergelson, J., Borgwardt, K. M., . . . Genomes, C. 2016. 1,135 Genomes Reveal the Global Pattern of Polymorphism in *Arabidopsis thaliana*. *Cell*, *166*(2), 481-491. https://doi.org/10.1016/j.cell.2016.05.063

Andersson, B. A., Zhao, W., Haller, B., Brännström, Å., & Wang, X.-R. 2023. Inference of the distribution of fitness effects of mutations is affected by SNP filtering methods, sample size and population structure. https://doi.org/10.22541/au.168311072.23828759/v1

Arunkumar, R., Ness, R. W., Wright, S. I., & Barrett, S. C. H. 2014. The evolution of selfing is accompanied by reduced efficacy of selection and purging of deleterious mutations. *Genetics*, *199*(3), 817-829. https://doi.org/10.1534/genetics.114.172809

Ayala, F. J., Powell, J. R., Tracey, M. L., Mourão, C. A., & Pérez-Salas, S. 1972. Enzyme variability in the *Drosophila willistoni* group. IV. Genic variation in natural populations of *Drosophila willistoni*. *Genetics*, *70*(1), 113-139. https://doi.org/10.1093/genetics/70.1.113

Bataillon, T., & Bailey, S. F. 2014. Effects of new mutations on fitness: insights from models and data. In C. W. Fox & T. A. Mousseau (Eds.), *Year in Evolutionary Biology* (Vol. 1320, pp. 76-92. Blackwell Science Publ. https://doi.org/10.1111/nyas.12460

Bechsgaard, J. S., Castric, V., Charlesworth, D., Vekemans, X., & Schierup, M. H. 2006. The Transition to Self-Compatibility in *Arabidopsis thaliana* and Evolution within S-Haplotypes over 10 Myr. *Molecular Biology and Evolution*, *23*(9), 1741-1750. https://doi.org/10.1093/molbev/msl042

Blischak, P. D., Barker, M. S., & Gutenkunst, R. N. 2020. Inferring the demographic history of inbred species from genome-wide SNP frequency data. *Molecular Biology and Evolution*, *37*(7), 2124-2136. https://doi.org/10.1093/molbev/msaa042

Boyko, A. R., Williamson, S. H., Indap, A. R., Degenhardt, J. D., Hernandez, R. D., Lohmueller, K. E., . . . Bustamante, C. D. 2008. Assessing the evolutionary impact of amino acid

mutations in the human genome. *Plos Genetics*, *4*(5), e1000083, Article e1000083. https://doi.org/10.1371/journal.pgen.1000083

Browning, B. L., Zhou, Y., & Browning, S. R. 2018. A one-penny imputed genome from next-generation reference panels. *The American Journal of Human Genetics*, *103*(3), 338-348. https://doi.org/10.1016/j.ajhg.2018.07.015

Casillas, S., & Barbadilla, A. 2017. Molecular population genetics. *Genetics*, *205*(3), 1003-1035. https://doi.org/10.1534/genetics.116.196493

Charlesworth, B. 1992. Molecular panselectionism. *Science*, *257*(5068), 420-421. https://doi.org/10.1126/science.257.5068.420-a

Charlesworth, D., Charlesworth, B., & Morgan, M. T. 1993. Mutation accumulation in finite outbreeding and inbreeding populations. *Genetical Research*, *61*(1), 39-56. https://doi.org/10.1017/S0016672300031086

Chen, J., Bataillon, T., Glémin, S., & Lascoux, M. 2021. What does the distribution of fitness effects of new mutations reflect? Insights from plants. *New Phytologist*. https://doi.org/10.1111/nph.17826

Chen, J., Glemin, S., & Lascoux, M. 2020. From drift to draft: how much do beneficial mutations actually contribute to predictions of Ohta's slightly deleterious model of molecular evolution? *Genetics*, *214*(4), 1005-1018. https://doi.org/10.1534/genetics.119.302869

Darwin, C. 1859. *On the origin of species* (G. Beer, Ed. Revised edition 2008 ed.. Oxford University Press.

Dobzhansky, T. 1937. *Genetics and the origin of species*. Columbia university press.

Dobzhansky, T., & Pavlovsky, O. 1957. An experimental study of interaction between genetic drift and natural selection. *Evolution*, *11*(3), 311-319. https://doi.org/10.2307/2405795

Eyre-Walker, A., & Keightley, P. D. 2007. The distribution of fitness effects of new mutations. *Nature Reviews Genetics*, *8*(8), 610-618. https://doi.org/10.1038/nrg2146

Eyre-Walker, A., & Keightley, P. D. 2009. Estimating the rate of adaptive molecular evolution in the presence of slightly deleterious mutations and population size change. *Molecular Biology and Evolution*, *26*(9), 2097-2108. https://doi.org/10.1093/molbev/msp119

Fisher, R. A. 1919. XV.—The correlation between relatives on the supposition of mendelian inheritance. *Transactions of the Royal Society of Edinburgh*, *52*(2), 399-433. https://doi.org/10.1017/S0080456800012163

Fisher, R. A. 1923. XXI.—On the dominance ratio. *Proceedings of the Royal Society of Edinburgh*, *42*, 321-341. https://doi.org/10.1017/S0370164600023993

Ford, E. B. 1965. *Ecological genetics* (2. Ed). Methuen.

Gilbert, K. J., Zdraljevic, S., Cook, D. E., Cutter, A. D., Andersen, E. C., & Baer, C. F. 2021. The distribution of mutational effects on fitness in *Caenorhabditis elegans* inferred from standing genetic variation. *Genetics*. https://doi.org/10.1093/genetics/iyab166

Gillespie, J. H. 1994. *The causes of molecular evolution*. Oxford University Press.

Gillespie, J. H. 1998. *Population genetics: A concise guide*. The Johns Hopkins University Press.

Gordeuk, V. R., Castro, O. L., & Machado, R. F. 2016. Pathophysiology and treatment of pulmonary hypertension in sickle cell disease. *Blood*, *127*(7), 820-828. https://doi.org/10.1182/blood-2015-08-618561

Gossmann, T. I., Song, B. H., Windsor, A. J., Mitchell-Olds, T., Dixon, C. J., Kapralov, M. V., . . . Eyre-Walker, A. 2010. Genome wide analyses reveal little evidence for adaptive evolution in many plant species. *Molecular Biology and Evolution*, *27*(8), 1822-1832. https://doi.org/10.1093/molbev/msq079

Guo, J.-F., Zhao, W., Andersson, B., Mao, J.-F., & Wang, X.-R. 2023. Genomic clines across the species boundary between a hybrid pine and its progenitor in the eastern Tibetan Plateau. *Plant Communications*, *4*(4), 100574. https://doi.org/10.1016/j.xplc.2023.100574

Gutenkunst, R. N., Hernandez, R. D., Williamson, S. H., & Bustamante, C. D. 2009. Inferring the joint demographic history of multiple populations from multidimensional SNP frequency data. *Plos Genetics*, *5*(10), e1000695, Article e1000695. https://doi.org/10.1371/journal.pgen.1000695

Haigh, J., & Smith, J. M. 1974. The hitch-hiking effect of a favourable gene. *Genetical Research*, *23*(1), 23-35. https://doi.org/10.1017/S0016672300014634

Haldane, J. B. 1990. *The causes of evolution* (Vol. 5). Princeton University Press.

Heller, R., & Smith, J. M. 1978. Does Muller's ratchet work with selfing? *Genetics Research*, *32*(3), 289-293. https://doi.org/10.1017/S0016672300018784

Herrick, J. B. 1910. Peculiar elongated and sickle-shaped red blood corpuscles in a case of severe anemia. *Archives of Internal Medicine*, *VI*(5), 517-521. https://doi.org/10.1001/archinte.1910.00050330050003

Hey, J. 1999. The neutralist, the fly and the selectionist. *Trends in Ecology & Evolution*, *14*(1), 35-38. https://doi.org/10.1016/S0169-5347(98)01497-9

Hill, W. G., & Robertson, A. 1966. The effect of linkage on limits to artificial selection. *Genetical Research*, *8*(3), 269-294. https://doi.org/10.1017/S0016672300010156

Hämälä, T., & Tiffin, P. 2020. Biased gene conversion constrains adaptation in *Arabidopsis thaliana*. *Genetics*, *215*(3), 831-846. https://doi.org/10.1534/genetics.120.303335

Johri, P., Charlesworth, B., & Jensen, J. D. 2020. Toward an evolutionarily appropriate null model: jointly inferring demography and purifying selection. *Genetics*, *215*(1), 173-192. https://doi.org/10.1534/genetics.119.303002

Keightley, P. D., & Eyre-Walker, A. 2007. Joint inference of the distribution of fitness effects of deleterious mutations and population demography based on nucleotide polymorphism frequencies. *Genetics*, *177*(4), 2251-2261. https://doi.org/10.1534/genetics.107.080663

Kern, A. D., & Hahn, M. W. 2018. The neutral theory in light of natural selection. *Molecular Biology and Evolution*, *35*(6), 1366-1371. https://doi.org/10.1093/molbev/msy092

Kim, B. Y., Huber, C. D., & Lohmueller, K. E. 2017. Inference of the distribution of selection coefficients for new nonsynonymous mutations using large samples. *Genetics*, *206*(1), 345-361. https://doi.org/10.1534/genetics.116.197145

Kimura, M. 1968. Evolutionary rate at molecular level. *Nature*, *217*(217), 624-626. https://doi.org/10.1038/217624a0

Kimura, M. 1979. The neutral theory of molecular evolution. *Scientific American*, *241*(5), 98-129.

Kimura, M., Maruyama, T., & Crow, J. F. 1963. The mutation load in small populations. *Genetics*, *48*(10), 1303-1312.

Kimura, M., & Ohta, T. 1971. Protein polymorphism as a phase of molecular evolution. *Nature (London)*, *229*(5285), 467-469. https://doi.org/10.1038/229467a0

King, J. L., & Jukes, T. H. 1969. Non-Darwinian evolution. *Science*, *164*(3881), 788-798. https://doi.org/doi:10.1126/science.164.3881.788

Kousathanas, A., & Keightley, P. D. 2013. A comparison of models to infer the distribution of fitness effects of new mutations. *Genetics*, *193*(4), 1197–1208. https://doi.org/10.1534/genetics.112.148023

Kreitman, M. 1996. The neutral theory is dead. Long live the neutral theory. *Bioessays*, *18*(8), 678-683. https://doi.org/10.1002/bies.950180812

Kutschera, V. E., Poelstra, J. W., Botero-Castro, F., Dussex, N., Gennnnell, N. J., Hunt, G. R., . . . Wolf, J. B. W. 2020. Purifying selection in corvids is less efficient on islands. *Molecular Biology and Evolution*, *37*(2), 469-474. https://doi.org/10.1093/molbev/msz233

Leon-Apodaca, A. V., Kumar, M., Castillo, A. d., Conroy, G. C., Lamont, R. W., Ogbourne, S., . . . Szpiech, Z. A. 2023. Genomic consequences of isolation and inbreeding in an

island dingo population. *bioRxiv* 2023.2009.2015.557950. https://doi.org/10.1101/2023.09.15.557950

Lewontin, R. C., & Hubby, J. L. 1966. A molecular approach to study of genic heterozygosity in natural populations. 2. Amount of variation and degree of heterozygosity in natural populations of *Drosophila pseudoobscura*. *Genetics*, *54*(2), 595-609.

Mao, J.-F., & Wang, X.-R. 2011. Distinct niche divergence characterizes the homoploid hybrid speciation of *Pinus densata* on the tibetan plateau. *The American Naturalist*, *177*(4), 424-439. https://doi.org/10.1086/658905

Mochales-Riaño, G., Fontsere, C., de Manuel, M., Talavera, A., Burriel-Carranza, B., Tejero-Cicuéndez, H., . . . Carranza, S. 2023. Genomics reveals introgression and purging of deleterious mutations in the Arabian leopard (*Panthera pardus nimr*). *iScience*, *26*(9), 107481. https://doi.org/10.1016/j.isci.2023.107481

Muller, H. J. 1932. Some genetic aspects of sex. *The American Naturalist*, *66*, 118 - 138.

Nei, M. 2005. Selectionism and neutralism in molecular evolution. *Molecular Biology and Evolution*, *22*(12), 2318-2342. https://doi.org/10.1093/molbev/msi242

Ohta, T. 1973. Slightly deleterious mutant substitutions in evolution. *Nature*, *246*(5428), 96-98. https://doi.org/10.1038/246096a0

Ohta, T. 1992. The nearly neutral theory of molecular evolution. *Annual Review of Ecology and Systematics*, *23*, 263-286. https://doi.org/10.1146/annurev.es.23.110192.001403

Ohta, T. 1996. The current significance and standing of neutral and nearly neutral theories. *Bioessays*, *18*(8), 673-677. https://doi.org/10.1002/bies.950180811

Pauling, L., Itano, H. A., & et al. 1949. Sickle cell anemia a molecular disease. *Science*, *110*(2865), 543-548. https://doi.org/10.1126/science.110.2865.543

Piganeau, G., & Eyre-Walker, A. 2003. Estimating the distribution of fitness effects from DNA sequence data: Implications for the molecular clock. *Proceedings of the National Academy of Sciences of the United States of America*, *100*(18), 10335-10340. https://doi.org/10.1073/pnas.1833064100

Punnett, R. C. 1930. The genetical theory of natural selection. *Nature*, *126*(3181), 595-597. https://doi.org/10.1038/126595a0

Sianta, S. A., Peischl, S., Moeller, D. A., & Brandvain, Y. 2022. The efficacy of selection may increase or decrease with selfing depending upon the recombination environment. *Evolution*, *77*(2), 394-408. https://doi.org/10.1093/evolut/qpac013

Soni, V., Pfeifer, S. P., & Jensen, J. D. 2023. The effects of mutation and recombination rate heterogeneity on the inference of demography and the distribution of fitness effects. [Preprint]. *bioRxiv* 2023.*11.11.566703*. https://doi.org/10.1101/2023.11.11.566703

Stamatoyannopoulos, G. 1972. The molecular basis of hemoglobin disease. *Annu Rev Genet*, *6*, 47-70. https://doi.org/10.1146/annurev.ge.06.120172.000403

Tataru, P., & Bataillon, T. 2019. polyDFEv2.0: testing for invariance of the distribution of fitness effects within and across species. *Bioinformatics*, *35*(16), 2868-2869. https://doi.org/10.1093/bioinformatics/bty1060

Tataru, P., Mollion, M., Glémin, S., & Bataillon, T. 2017. Inference of distribution of fitness effects and proportion of adaptive substitutions from polymorphism data. *Genetics*, *207*(3), 1103-1119. https://doi.org/10.1534/genetics.117.300323

Wang, X.-R., & Szmidt, A. 1994. Hybridization and Chloroplast DNA Variation in a Pinus Species Complex from Asia. *Evolution*, *48*, 1020-1031. https://doi.org/10.2307/2410363

Wang, X.-R., Szmidt, A., & Savolainen, O. 2001. Genetic composition and diploid hybrid speciation of a high mountain pine, *Pinus densata*, native to the Tibetan Plateau. *Genetics*, *159*, 337-346.

Wilson, D. J., Hernandez, R. D., Andolfatto, P., & Przeworski, M. 2011. A population genetics-phylogenetics approach to inferring natural selection in coding sequences. *Plos Genetics*, *7*(12), e1002395. https://doi.org/10.1371/journal.pgen.1002395

Zeitler, L., Parisod, C., & Gilbert, K. J. 2023. Purging due to self-fertilization does not prevent accumulation of expansion load. *Plos Genetics*, *19*(9), e1010883. https://doi.org/10.1371/journal.pgen.1010883

Zhang, D., Xu, C., Manwani, D., & Frenette, P. S. 2016. Neutrophils, platelets, and inflammatory pathways at the nexus of sickle cell disease pathophysiology. *Blood*, *127*(7), 801-809. https://doi.org/10.1182/blood-2015-09-618538

Zhang, Y., Stern, A. J., & Nielsen, R. 2023. The evolutionary dynamics of local adaptations under genetic rescue is determined by mutational load and polygenicity. *The Journal of heredity*.