Adaptive Evolution in Linked Genomes

Inaugural-Dissertation
zur
Erlangung des Doktorgrades
der Mathematisch-Naturwissenschaftlichen Fakultät
der Universität zu Köln
vorgelegt von

Stephan Schiffels

aus Aachen

Berichterstatter: Prof. Dr. Michael Lässig
(Gutachter)

Prof. Dr. Martin Lercher

Tag der mündlichen Prüfung: 27. Januar 2012

Zusammenfassung

Adaptive Evolution wird von verschiedenen Kräften beherrscht: *Mutationen* entstehen zufällig im Genom und erzeugen Unterschiede im reproduktiven Erfolg einzelner Individuen; *natürliche Selektion* verschiebt diese Variabilität zugunsten von Individuen mit hoher Fitness; *Gendrift* erzeugt Zufallsfluktuationen in der Zahl der Nachkommen und beeinflusst vor allem Mutationen mit schwachem Fitness-Effekt. Darüber hinaus stellt *Genkopplung* eine wichtige evolutionäre Kraft dar. Genkopplung erzeugt Interferenzen und Wechselwirkungen, durch die gleichzeitig entstehende Mutationen sich gegenseitig beeinflussen. In dieser Arbeit entwickeln wir ein umfassendes Modell für adaptive Evolution, welches Wechselwirkungen durch Genkopplung zwischen vorteilhaften und schädlichen Mutationen in einem einheitlichen System zusammenfasst. Unsere näherungsweise analytische Lösung beschreibt sowohl die Fixationsraten solcher Mutationen, als auch das Verhältnis zwischen vorteilhaften und schädlichen Allelen in der Sequenz. Unser Ergebnis zeigt, dass Wechselwirkungen durch Genkopplung ein Regime effektiver Neutralität erzeugen: Gene mit einem Fitness-Effekt, der kleiner ist als ein charakteristischer Wert, haben zufällig fixierte Allele, und sowohl vorteilhafte als auch schädliche Mutationen in diesen Regionen haben nahezu neutrale Fixationsraten. Diese Dynamik begrenzt nicht nur die Geschwindigkeit adaptiver Prozesse, sondern auch die Anpassung einer Population an ihre Umgebung. Wir wenden unser Modell auf zwei verschiedene Szenarien an: stationäre Adaptation in einer zeitabhängigen Umgebung, und Anpassung an eine konstante Umgebung. In beiden Fällen stimmen unsere analytischen Vorhersagen gut mit Simulationen überein. Unser Ergebnis zeigt, dass Genkopplung biologische Funktionen einer adaptierenden Population beeinträchtigen kann, wodurch natürlicher Anpassungsfähigkeit Grenzen gesetzt sind. Darüber hinaus entwickeln wir ein probabilistisches Modell, mit welchem Genom-Daten analysiert werden können, und welches Genkopplung explizit berücksichtigt. Tests anhand simulierter Daten zeigen, dass unsere Methode das Maß an positiver Selektion einer gekoppelten Sequenz korrekt vorhersagt. Im Gegensatz dazu interpretieren bisherige Methoden Genkopplungs-Effekte fälschlicherweise als positive Selektion. Wir wenden unsere Methode auf Genom-Daten der Spezies *Drosophila melanogaster* an und zeigen, dass ein substantieller Anteil der Sequenzunterschiede zweier Fliegen-Arten nicht auf natürliche Selektion, sondern auf Genkopplung zurückzuführen ist.

Abstract

Adaptive evolution is governed by various forces: *Mutations* occur randomly in the genome and generate variability in the individuals' reproductive success; *natural selection* shifts this variability in the population towards individuals with high fitness; *genetic drift* introduces random fluctuations in the number of offspring of an individual and affects weakly selected or neutral mutations. On top of these, *genetic linkage* can be an important evolutionary force. Linkage generates interference interac tions, by which simultaneously occurring mutations affect each other's chance of fixation. Here, we develop a comprehensive model of adaptive evolution in linked genomes, which integrates interference interactions between multiple beneficial and deleterious mutations into a unified framework. By an approximate analytical solution, we predict the fixation rates of these mutations, as well as the probabilities of beneficial and deleterious alleles at fixed genomic sites. We find that interference interactions generate a regime of emergent neutrality: all genomic sites with selection coefficients smaller in magnitude than a characteristic threshold have nearly random fixed alleles, and both beneficial and deleterious mutations at these sites have nearly neutral fixation rates. We show that this dynamics limits not only the speed of adaptation, but also a population's degree of adaptation in its current environment. We apply the model to different scenarios: stationary adaptation in a time-dependent environment, and approach to equilibrium in a fixed environment. In both cases, the analytical predictions are in good agreement with numerical simulations. Our results suggest that interference can severely compromise biological functions in an adapting population, which sets viability limits on adaptive evolution under linkage. We furthermore develop a likelihood-based inference method for genomic data, which explicitly takes into account genetic linkage. Tests with simulated datasets show that our method correctly predicts the amount of positive selection in linked sequence. In contrast, many existing tests falsely interpret traces from linkage as spurious positive selection. We apply our method to fruit fly genome data (*Drosophila melanogaster*), and find that a substantial fraction of sequence differences between two related fly species is in fact caused by linkage instead of natural selection.

Table of Contents

Introduction 6

1 Genomic model for adaptation 10

  1.1 Genome, mutations and fitness 10

  1.2 Adaptive dynamics 18

2 Fixation probability under linkage 23

  2.1 Pairwise interaction diagrams 23

  2.2 Single site propagator 32

  2.3 From pairwise to many locus interactions 34

  2.4 Selection regimes and emergent neutrality 39

3 Analysis of adaptive scenarios 42

  3.1 Computer simulation scheme 42

  3.2 Numerical solution 44

  3.3 Stationary adaptation 45

  3.4 Approach to equilibrium in a static fitness landscape 54

  3.5 Non-exponential and epistatic fitness landscapes 60

  3.6 Fluctuations and intermittency of the adaptive process 63

4 Inference in genomic data 67

  4.1 Substitution rate of passenger mutations 68

  4.2 Allele frequency of passenger mutations 75

  4.3 Single-locus computer simulations 80

  4.4 Multi-locus computer simulations 85

  4.5 Likelihood-model for genomic data 89

Discussion 96

Acknowledgments 98

References 99

Appendix 105

Erklärung 109

Lebenslauf 110

Introduction

Populations adapt to new or changing environments by increasing their overall fitness, that is, their survival probability and reproductive success. The “fuel” of this process are mutations in the individuals’ heritable information that occur at random times and with random effects. The fate of a mutation depends on its effect: Beneficial mutations convey a reproductive advantage and typically increase in frequency from generation to generation, and ultimately can become fixed in the whole population. In contrast, deleterious mutations will typically go extinct after only a few generations, because they cause a selective disadvantage. While these evolutionary processes are based on simple principles, their understanding on the molecular level has proven considerably difficult.

One reason for this difficulty is that the correspondence between changes in the genomic sequence and the resulting effect on the biological function is often opaque. While many encoded biological functions are known, most parts of the genome have unknown functional effects. But even *given* all this functional information, understanding the adaptive dynamics is still challenging. One feature that is particularly interesting in the light of statistical physics is genomic linkage. Linkage couples the fate of one mutation to neighboring mutations in the same sequence. The cause of this coupling is the common genomic background on which mutations occur. For example, if a new mutation occurs in a sequence that already carries an advantageous mutation, the new mutation will have an increased chance of fixation independent of its own fitness effect. We can therefore think of linkage as a dynamical coupling among mutations in a sequence, similar to interactive couplings in many-body systems known from statistical physics.

Genomic linkage comes in different strengths: While sexually reproducing organisms counteract linkage by recombination, asexual populations are strongly affected. One important consequence of linkage is its impact on the speed of adaptation. Several classical studies have shown that linkage interferences can substantially reduce the speed of adaptation in large asexual populations [25, 46, 24, 5, 28]. These results are supported by microbial evolution experiments that provide a growing amount of data on adaptive evolution under linkage [74, 66, 75, 70, 18, 60, 38, 4, 7, 42], and similar data are available for adaptive evolution in viral systems [10, 62, 52]. Even for higher organisms that reproduce sexually, the importance of linkage has been recognized [23, 36, 76, 67, 14]. Today, modern deep sequencing technologies open these systems to population genomic analysis, which provides unprecedented insights into the dynamical processes of adaptation under linkage. We can now ask new questions: How do the genomes of a population and their current fitness values evolve, and what are the rates of beneficial and deleterious changes observed in the process?

The model developed in chapter 1 establishes the conceptual framework to answer such questions. It describes the adaptive evolution of a finite asexual population, whose individuals have non-recombining genotypes of finite length. Evolution takes place by mutations, genetic drift, and selection, given by a genomic fitness landscape, which is specified by the distribution of selection coefficients between alleles at individual sequence sites. The evolving population is described by its genome state, i.e., by the fraction of fitter vs. less fit alleles in the genomic sequence. The genome state determines the rate of beneficial and deleterious mutations and the distribution of their fitness effects: if the population is well adapted, most sites are fixed at the fitter alleles and most novel mutations will be deleterious; if the population is poorly adapted, more mutations will be beneficial. Thus, the scope of our genomic model goes beyond that of previous studies, which analyze the statistics of substitutions given the rate and selection coefficients of mutations as fixed input parameters [28, 18, 57]. In particular, our model can describe non-stationary adaptation, i.e., processes in which the distribution of selection effects for mutations becomes itself time-dependent.

Linkage enters our model by affecting the *efficacy* of the adaptive process: Because other mutations influence the fate of a given mutation, its chance of fixation is strongly affected. In chapter 2, we develop an approximate calculus for the chance of fixation under multiple simultaneous interacting mutations. Since any mutation is both the *target* of interference effects from other mutations, and is itself *interfering* with yet other target mutations, we obtain an approximate, self-consistent summation of interference interactions between all co-occurring mutations. We show that these interactions partition the adaptive dynamics into strongly beneficial *driver* mutations, which fix without substantial interference, and beneficial or deleterious *passenger* mutations, which suffer from strong interference. Our analytical approach differs from the two classes of models analyzed in previous work. The clonal interference calculus [28] focuses on the dynamics of driver mutations, but it does not consider passenger mutations and neglects the effects of multiple co-occurring mutations. On the other hand, the traveling-wave approach assumes an ensemble of many co-occurring mutations, which have the same or similar selective effect [64, 18]. The adaptive dynamics studied in this thesis, which takes place in a linked genome with a broader distribution of selection coefficients, follows neither of these models: it is governed by interference interactions between *few* strongly beneficial substitutions and their effect on more weakly selected alleles.

In chapter 3, we analyze the biological implications of our model for two specific scenarios of adaptive evolution, using computer simulations. The first is a stationary adaptive process maintained by an explicitly time-dependent fitness “seascape”, in which selection coefficients at individual genomic sites change direction at a constant rate [48, 49, 50]. Such time-dependence of selection describes changing environments, which can be generated by external conditions, migration or co-evolution. An example is the ongoing antigen-antibody co-evolution of the human influenza virus [10]. Our model predicts a selection regime of *effective neutrality*. Mutations in this regime have effectively neutral fixation rates, independent of their selection coefficients. The model furthermore predicts the speed of adaptation, and the population’s degree of adaptation in its current environment. The second adaptive scenario is the approach to evolutionary equilibrium in a static fitness landscape, starting from a poorly adapted initial state. This case describes, for example, the long-term laboratory evolution of bacterial populations in a constant environment [50]. The predictions of our model are now time-dependent: the regime of effectively neutral sites and the speed of adaptation decrease over time, while the degree of adaptation increases.

Finally, in chapter 4, we develop an application scheme of our theory to genomic data. Many existing methods that infer adaptive evolution from genomic data treat linkage at best heuristically, but mostly ignore it entirely [23]. Here we develop a new method that explicitly incorporates linkage between neighboring sites and the resulting hitchhiking and interference effects. This method is based on a simpler model than introduced in chapter 1, but with two important extensions: First, in order to apply our theory to higher organisms we need to incorporate *genetic recombination*, which is a distinctive feature of sexual reproduction; second, we need to explicitly describe within-population diversity (*polymorphisms*), which was not covered in the analyses of chapters 2 and 3. We develop a likelihood-framework which can be used for inference from genomic data of within-species and between-species diversity. Because of recombination, observables now become position-dependent, due to the distance-dependence of driver-passenger effects. We first apply our likelihood-framework to simulated datasets and demonstrate its statistical power. We then apply this method to real genomic data from the fruit fly *Drosophila melanogaster* and find that many genes have a substantial number of deleterious passenger substitutions and hence exhibit strong linkage effects.

1 Genomic model for adaptation

Here we develop a genomic model of adaptation under linkage. The model describes the evolution of a finite asexual population, whose individuals have non-recombining genotypes of finite length. Evolution takes place by mutation, selection and genetic drift. We show how the *adaptive process*, which changes the frequencies of beneficial and deleterious alleles at polymorphic sites, is linked to the *genome state*, which includes the distribution of beneficial and deleterious alleles at fixed sites.

1.1 Genome, mutations and fitness

We consider a population as a set of N individuals, each described by one linear chromosome. Each chromosome consists of L sites with two possible alleles 1 and -1. As notation, we use $a_{ij} = \{-1, 1\}$ with $i = 1 \ldots N$ and$j = 1 \ldots L$ to describe allele j of individual i, and the vector notation $\boldsymbol{a} \in \{-1, 1\}^L$ to denote a whole chromosome. The population is kept fixed throughout evolution and the reproduction scheme follows the familiar Wright-Fisher process [21]: Each next generation of N individuals is sampled with replacement from the previous generation, where the individual i is sampled with probability

$$p_i = \frac{e^{F(\boldsymbol{a}_i)}}{\sum_{k=1}^{N} e^{F(\boldsymbol{a}_k)}}. \tag{1}$$

The fitness landscape $F(\boldsymbol{a}) \mapsto \mathbb{R}$ assigns a fitness value to each chromosome. Note that with a trivial fitness landscape $F(\boldsymbol{a}) \equiv F$, the sampling probability of a given individual is simply $p_i \equiv 1/N$, corresponding to standard random sampling with replacement. Mutations occur with a uniform probability μ per genomic site per individual and simply "flip" the allele $a_{ij} \to -a_{ij}$ in a randomly chosen individual i and site j. The selection effect of a mutation, and whether it is beneficial or deleterious depends on the fitness landscape and the state of the population, as will be discussed later. In the following, we introduce the different fitness landscapes used in this work.

Static additive fitness landscape

Consider as simplest evolutionary case a genomic sequence that encodes for some biologically important feature, such as a protein. In the simplest possible case, the function of the encoded feature depends only on the sequence and furthermore depends only additively on the individual genomic sites. To model this case, I consider this additive fitness function:

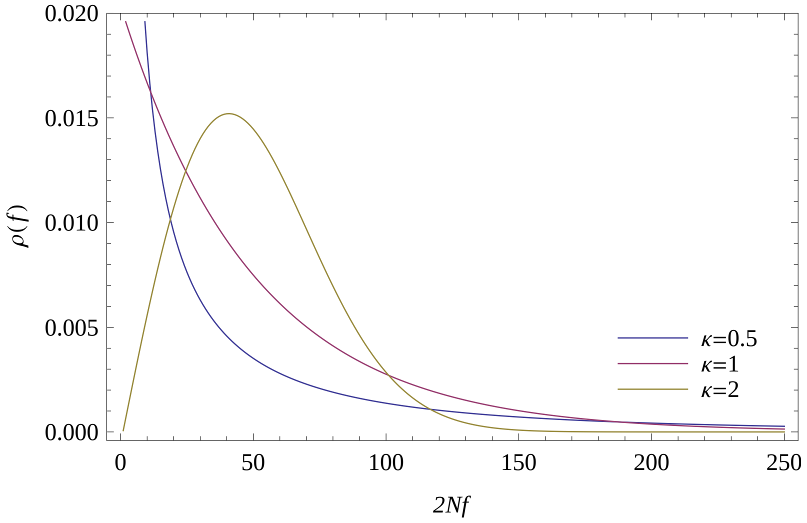
$$F(\boldsymbol{a}) = \frac{1}{2} \sum_{j=1}^{L} a_j \, f_j \tag{2}$$

where the selection coefficients $f_j \geq 0$ are fixed random numbers, drawn from a normalized distribution $\rho(f)$. According to this fitness landscape, each genomic site has a “preferred” allele $a_j = 1$ and an “unpreferred” allele $a_j = -1$. The fitness difference between these two alleles is determined by the selection coefficients f_j and, due to the additivity in equation 2, independent between locations j.

Not every genomic site has equal importance for the biological function. Some sites have zero or weak fitness effects, while other sites are strongly selected. The distribution of selection coefficients, $\rho(f)$ depends on the protein or feature, as well as on environmental and ecological conditions and on other features that are encoded in the genomic background. For most parts of our theory, this distribution will not explicitly enter the derivations. In the simulations (chapter 3), we use an exponential distribution for $\rho(f)$, and - where explicitly stated - a Weibull-distribution [58], parametrized by two parameters ζ and κ:

$$\rho(f) = \frac{\kappa}{\zeta} \left(\frac{f}{\zeta} \right)^{\kappa - 1} e^{-(f/\zeta)^{\kappa}}, \tag{3}$$

which is shown in Figure 1. For $\kappa = 1$, this distribution is exponential with mean $\overline{f} = \zeta$. For $\kappa < 1$, the tail falls off slower than exponential, whereas $\kappa > 1$ yields a steeper tail. For general values of the shape parameter κ, the mean fitness of this distribution is directly connected to ζ and κ via $\overline{f} = \zeta \, \Gamma(1 + 1/\kappa)$. To uniquely refer to a given distribution, I will use the parameter pair $\overline{f}$ and κ, rather than ζ and κ. The shape parameter κ controls how *similar* the selection coefficients of mutations are: the higher κ, the more similar the fitness effects. This has consequences for the dynam ics, as will be discussed in section 3.5.



▲ **Figure 1. Distributions of fitness effects.** This plot shows the distribution of selection coefficients, $\rho(f)$, given by equation 3, for three different shape parameters and a mean selection coefficient $2N\bar{f} = 50$.

The Weibull-distribution has the advantage that it is numerically easy to create random deviates from it: The cumulative probability density

$$\Omega(f) = \int_0^f \rho(f)\, df = 1 - e^{-(f/\zeta)^\kappa} \tag{4}$$

can be trivially inverted:

$$\Omega^{-1}(x) = \zeta \log\left(\frac{1}{1-x}\right)^{1/\kappa} \tag{5}$$

Random deviates are obtained by drawing uniform continuous deviates from the interval $x \in [0; 1]$ and transforming via equation 5.

Time-dependent additive fitness landscape

I now extend the simple fitness landscape (equation 2) to include time-dependent selection. This time-dependence can be caused by changing external conditions, by changing ecological pressure or even by other genes or genomic features that change inside the genomic background (epistatic fitness interactions). A minimal time-dependence of the fitness landscape was suggested by Mustonen and Lässig [48] where a single locus with two-alleles evolves under “selection flips”. A selection flip changes the preferred allele to the unpreferred allele and vice versa, but keeps the amplitude of selection, i.e. the fitness difference between the two alleles unchanged. Here we use this concept to extend the multi-locus fitness landscape defined in equation 1:

$$F(\boldsymbol{a}, t) = \frac{1}{2}\sum_{j=1}^{L} a_j\, f_j\, \eta_j(t) \tag{6}$$

where $\eta_j\,(t) = \{-1, 1\}$ are j independent stochastic processes, each with the same “flip rate” γ:

$$\overline{\eta_j(t)} = 0 \quad \text{and} \quad \overline{\eta_j\,(t)\,\eta_k\,(t')} = \delta_{j\,k}\, e^{-\gamma|t-t'|}. \tag{7}$$

Note that this particular time-dependent fitness landscape presents a *minimal* model in the sense that it is the simplest time-dependent fitness landscape that is still analytically tractable. It generates a surplus of beneficial over deleterious fixations, which is an essential non-equilibrium property of adaptation.

Epistatic fitness landscape

The additive fitness landscapes above have the important property that fitness effects of mutations at one site j are *independent* of mutations at another site k. In real populations, this independence cannot strictly hold. In a protein, for example, the encoded three-dimensional structure introduces dependencies between distant sites. These dependencies are called *epistasis* and can be very complex. Here, we consider a simple extension of our additive fitness landscapes that introduces epistatic interactions between loci, which has been suggested as “pairwise epistatic model” by Neher and Shraiman [53]. In addition to the additive part of the fitness landscape, we consider a pairwise term that includes all possible pairwise interactions in the sequence:

$$F(\boldsymbol{a}, t) = \frac{1}{2}\sum_{j=1}^{L} a_j\, f_j\, \eta_j(t) + e \sum_{j<k} a_j\, a_k\, f_{j\,k}. \tag{8}$$

It has two fitness components: the familiar additive component presented before, and an additional non-additive (epistatic) component with individual terms $f_{j\,k}$ for all pairwise allele combinations. The tunable parameter e controls the relative weight of the epistatic component. Here, we fill the matrix f_{jk} with normally distributed random numbers with mean 0 and variance 1. This fitness landscape is analyzed in section 3.5, where we show that epistasis does *not* have a sizable influence on our results.

Rates and dynamical regimes

Our model crucially depends on two rates: the mutation rate μ determines the rate of allelic mutations $a_{ij} \to -a_{ij}$, and the flip rate γ determines the rate of selection flips $\eta_j \to -\eta_j$. Two dynamical regimes are distinguished by the mutation rate: For $\mu N \ll 1$, each locus is fixed most of the time and polymorphisms need to be considered only in order to compute the transition probabilities between these fixed states. In contrast, if $\mu N > 1$, each locus is polymorphic at each locus most of the time, which results in entirely different dynamics than a sequence of substitutions between fixed states. In this thesis, we only consider the regime, in which $\mu N \ll 1$, because it is relevant for many natural populations [19]. Note however, that the genomewide mutation rate, $\mu N L$, can still be much larger than 1, which results in simultaneous substitution events than interfere with each other and influence fixation probabilities.

Another important parameter combination is given by $\gamma \big/ \left(\mu N \overline{f}\right)$, which governs the speed with which the fitness seascape changes its shape. We can again distinguish two regimes, defined by comparing this parameter combination with 1. For $\gamma \ll \mu N \overline{f}$, selection flips occur with a rate that is much lower than the rate with which beneficial mutations arise and fix ($\sim \mu N \overline{f}$) in the population. In other words, with a flip rate that is low in the sense defined here, flip dynamics do not interfere with substitution dynamics and we expect selected sites to substitute with rate γ (see equation 24 and below). The other dynamical regime, in which $\gamma \gtrsim \mu N \overline{f}$ is governed by fast fluctuating selection [49], which is not studied here.

As discussed, we restrict the mutation rate to the regime $\mu N \ll 1$. As a result, the state of the population is determined by the probabilities of fixed states

$$\lambda_+(f) + \lambda_-(f) \approx 1, \tag{9}$$

where $\lambda_+(f)$ is the probability that a site with selection coefficient f is fixed at the beneficial allele, and $\lambda_-(f)$ is the probability that it is fixed at the deleterious allele.

We will refer to the fixed state probabilities as *genome state*.

We expect our model to be applicable to microbial laboratory populations, which often fall into the range of evolutionary parameters covered by this study. For example, the population- and genome-wide mutation rate in an *E.coli* population of size $N = 10^5$ is $\mu N L = 250$ [19]. Our simulations cover system sizes up to $\mu N L = 2000$.

Mutations and substitution rates

The genome state determines the fraction of mutations that are beneficial or deleterious. As an example, in a perfectly adapted population with $\lambda_+(f) = 1$ for all f, all new mutations are deleterious. In general, the distribution of mutations can be expressed as

$$U(\sigma) = \begin{cases} \rho(\sigma)\,\mu\,L\,\lambda_-(\sigma) & \sigma > 0 \\ \rho(|\sigma|)\,\mu\,L\,\lambda_+(|\sigma|) & \sigma < 0. \end{cases} \tag{10}$$

Substitutions take place with a rate $V(\sigma)$, which is given by the product of the mutation rate $U(\sigma)$ and the fixation probability $G(\sigma)$:

$$V(\sigma) = N\,U(\sigma)\,G(\sigma). \tag{11}$$

To compute this fixation probability is the key challenge of the analysis of this model: Without interference, the fixation probability is given by Kimura's well known formula [41]

$$G_0(\sigma) = \frac{1 - e^{-2\sigma}}{1 - e^{-2N\sigma}}, \tag{12}$$

which is determined by selection and genetic drift. This famous equation has three important limits: Strongly deleterious mutations have an exponentially decreasing fixation probability: $G_0(\sigma) \to 2\,|\sigma| / \exp(-2\,N\,\sigma)$. Neutral mutations have a fixation probability $G_0(0) = 1/N$. Beneficial mutations fix with a linear probability $G_0(\sigma) \to 2\,\sigma$, which can also be derived from branching process arguments [30, 5].

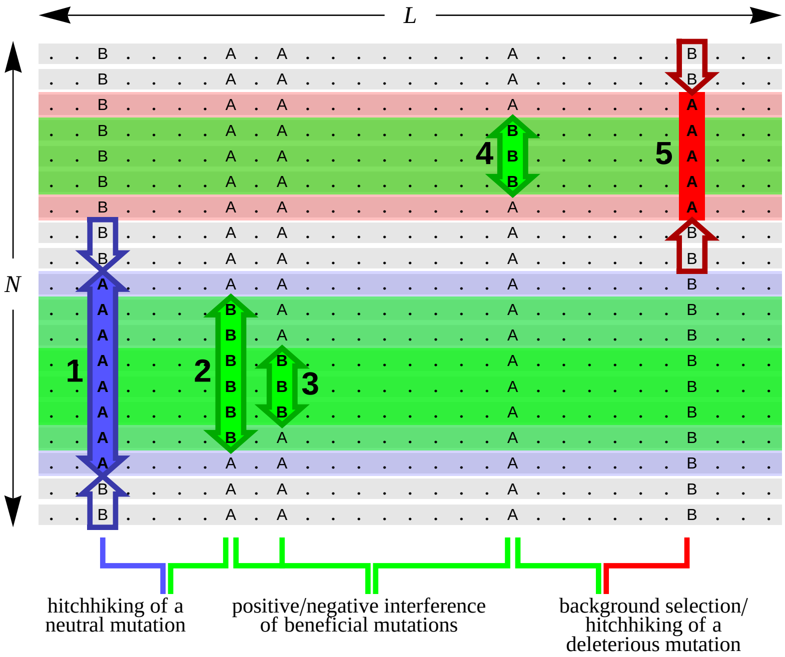
Linkage interferences

Under linkage, new mutations are not only influenced by selection and genetic drift, but also by mutations at linked loci. These mutations introduce interference interactions that have severe consequences on the dynamics of these sites, even under the simple additive fitness landscapes considered here. The complexity of these interference interactions is reflected by the long history of the subject in population genetics literature, which dates back to Fisher and Muller in the 1930's [25, 46]. The key observation of the Fisher-Muller theory is that in the absence of recombination, two mutations can both reach fixation only if the second mutation occurs in an individual that already carries the first. In other words, mutations occurring in different individuals interfere with each other. Interference inevitably causes a fraction of all mutations to be lost, even if they are beneficial and have already reached substantial frequencies in the population (i.e., have overcome genetic drift). Following a further seminal study, the interference between linked mutations is commonly referred to as Hill-Robertson effect [35]. This term is also used more broadly to describe the interplay between linkage and selection: interference interactions reduce the fixation probability of beneficial mutations and enhance that of deleterious ones. Hence, they reduce the effect of selection on substitution rates [24, 5]. More recently, a number of theoretical and experimental studies have addressed evolution under linkage, focusing on particular interaction types that are shown schematically in Figure 2 and summarized in the following:

- *Clonal interference:* Two beneficial mutations that appear on different genetic backgrounds can not both fix in the population. Instead, one of the two mutations will eventually be outcompeted by the other. This effect on average results in a loss of beneficial mutations in comparison to unlinked loci [28, 55, 77, 57].
- *Multiple mutations:* The opposite effect of clonal interference occurs if the second beneficial mutation appears on the *same* background like the first beneficial mutation. This effect has been mainly described via traveling wave theory [64, 32, 18].
- *Hitchhiking of neutral mutations:* A new beneficial mutation appears on a genetic background with other neutral mutations already present. If the beneficial mutation fixes, neutral mutations in its background will “hitchhike” to fixation. This effect has been mainly used to describe the removal of neutral diversity by linked positive selection (“genetic draft”) [72, 6, 29, 40, 34, 2, 76].
- *Hitchhiking of deleterious mutations:* Even strongly deleterious mutations can fix in the population by hitchhiking with linked beneficial mutations [8, 33].

- *Background selection:* If a beneficial mutation appears on a genetic background that carries a strongly deleterious mutant, the beneficial mutation will have a reduced probability of fixation [11, 39, 37, 12, 13, 3]

Although this summary is not an exhaustive description as we only considered pairwise interference effects, it is already clear that the fixation probability of both beneficial, neutral or deleterious mutants can be strongly affected by linkage, which will be quantified in chapter 2.


L
N
1
2
3
4
5
hitchhiking of a neutral mutation
positive/negative interference of beneficial mutations
background selection/ hitchhiking of a deleterious mutation

▲ **Figure 2.** **Interference interactions of mutations in linked genomes.** Here we schematically show a population of *N* individuals with two-nucleotide genomes of length *L*. In a non-recombining genome, this process is governed by positive and negative interference interactions between beneficial (green), neutral (blue), and deleterious mutations (red). The figure shows five mutations simultaneously present in the population; their expected frequency changes in the absence of genetic linkage are indicated by arrows. The fitness contribution of each mutation additively effects the fitness of all individuals carrying that mutation, which is indicated by the background color of the sequences. Linkage introduces the following interactions: allele 1 may be driven to fixation by allele 2 (hitchhiking of a neutral mutation), alleles 2 and 3 enhance each other's probability of fixation (positive interference between beneficial mutations), alleles 3 and 4 compete for fixation (negative interference between beneficial mutations), allele 4 may be driven to loss by allele 5 (background selection), or allele 5 may be driven to fixation by allele 4 (hitchhiking of a deleterious mutation).

1.2 Adaptive dynamics

Degree of adaptation and fitness flux

We introduce two observables that characterize the efficiency of the adaptive process. First, the degree of adaptation is defined as

$$\alpha(f) = \lambda_+(f) - \lambda_-(f) \tag{13}$$

and as such is a number between 0 and 1. Note that randomly fixed sites have $\lambda_+(f) = \lambda_-(f) = 1/2$ and hence $\alpha(f) = 0$. In contrast, sites with perfect adaptation have $\lambda_+ = 1$ and $\lambda_- = 0$ and hence $\alpha(f) = 1$. The total degree of adaptation is defined as a weighted average across all sites:

$$\alpha = \frac{1}{\overline{f}} \int_0^\infty \alpha(f)\, \rho(f)\, f\, d f, \tag{14}$$

A more intuitive definition of the degree of adaptation can be obtained as follows. By equation 2, we can write the mean fitness of the population as

$$\begin{aligned} \langle F \rangle &= \frac{1}{2} \sum_{j=1}^{L} \langle a_j \rangle\, f_j \\ &= \frac{L}{2} \int_0^\infty (\lambda_+(f) - \lambda_-(f))\, \rho(f)\, f\, d f = \frac{L}{2} \int_0^\infty \alpha(f)\, \rho(f)\, f\, d f = L\, \overline{f}\, \alpha / 2. \end{aligned} \tag{15}$$

In a similar way, we also define the fitness of a perfectly adapted population $\langle F_{\text{max}} \rangle = L\, \overline{f} / 2$ and the fitness of a population of random sequences, which in our fitness landscape is zero, $\langle F_0 \rangle = 0$. We can then define the degree of adaptation in a way that is applicable to general fitness landscapes:

$$\alpha = \frac{\langle F \rangle - \langle F_0 \rangle}{\langle F_{\text{max}} \rangle - \langle F_0 \rangle} \tag{16}$$

which is closely related to various concepts of genomic *loads*, introduced by Haldane and others [30, 31, 47]. Equation 16 yields the intuitive interpretation of $1 - \alpha$ as the normalized “distance” of a population to the global fitness optimum. Due to mutations and drift, this distance will always be non-zero (i.e. $\alpha < 1$), since deleterious mutations lower a population’s fitness and hence lower α. Linkage interferences increase this effect as we will see in chapter 3.

As a second observable, we define the fitness flux, which has been introduced by Mustonen and Lässig [48] as a measure of the speed of adaptation. In the above described regime of $\mu N \ll 1$, it can be defined via the substitution rate $V(\sigma)$, defined by equation 11:

$$\Phi(f) = f(V(f) - V(-f)). \tag{17}$$

The total fitness flux is defined as an integral over all sites:

$$\Phi = \int_0^\infty \Phi(f)\, d f = V\, \overline{\sigma}_V, \tag{18}$$

with the total substitution rate V and the average selection coefficient of *fixed* mutations $\overline{\sigma}_V$.

In general, under time-dependent selection, the state probabilities $\lambda_+(f)$ and $\lambda_-(f)$ change according to a Markov-process in time due to selection flips and substitutions. The master equation of this process reads [68]:

$$\frac{d}{dt}\lambda_+(f) \;=\; \frac{1}{L\,\rho\,(f)}[V(f) - V(-f)] + \gamma[\lambda_+(f) - \lambda_-(f)] \;=\; -\frac{d}{dt}\lambda_-(f). \tag{19}$$

Here, $V(\sigma)$ depends on the mutation rate via equation (11) and hence - via the genome state - on time. The full dynamics are thus quite complex. Before we discuss special cases, it will prove useful to rewrite equation (19) in terms of the above introduced degree of adaptation and the fitness flux:

$$\Phi(f) = L\,\rho(f)\, f\left(\frac{1}{2}\,\frac{d\alpha(f)}{dt} + \gamma\,\alpha(f)\right) \tag{20}$$

and

$$\Phi = L\,\overline{f}\left(\frac{1}{2}\,\frac{d\alpha}{dt} + \gamma\,\alpha\right). \tag{21}$$

These two equations show that fitness flux and degree of adaptation are intuitively connected: The degree of adaptation describes the adaptive *state* of the population, while the fitness flux describes the adaptive *process* that changes this state. This connection becomes clear when we now focus on two particular special cases of equations 20 and 21.

Stationary adaptation in a fitness-seascape

Stationary adaptation is achieved, if the substitution rates become stationary and exactly balance the time-dependence of the fitness function. From equation (19) and (11) it follows from $d\lambda_\pm(f)/dt = 0$:

$$\lambda_+(f) = \frac{N\,G(f) + \gamma/\mu}{N\,G(f) + N\,G(-f) + 2\,\gamma/\mu} = 1 - \lambda_-(f). \tag{22}$$

Equation 22 links the fixation probabilities $G(\pm f)$ to the genomic state $\lambda_\pm(f)$. The more efficient the fixation process is in fixing beneficial mutations and removing deleterious mutations, the higher is the probability of observing a given genomic site at its fitter state, $\lambda_+(f)$ and vice versa for $\lambda_-(f)$.

Note that under strong selection $2\,N\,f \gg 1$ and a low rate of selection flips $\gamma \ll \mu\,N\,f$ and no interference effects, we can set $G(f) \sim 2\,f$ and $G(-f) = 0$ and get

$$\lambda_+(f) \approx 1, \qquad \lambda_-(f) = \frac{\gamma/\mu}{2\,N\,f} \tag{23}$$

and hence

$$\Phi(f) = \mu\,N\,f(\lambda_-(f)\,G(f) + \lambda_+(f)\,G(-f)) \approx \gamma\,f \tag{24}$$

with the total substitution rate $\sim\gamma$ [48]. This result intuitively reflects that under efficient fixation processes, substitutions occur with a rate equal to the rate of selection flips.

A more general characterization of the stationary dynamics follows from equation 21:

$$\Phi = L\,\overline{f}\,\gamma\,\alpha = \alpha\,\Phi_{\max}, \tag{25}$$

where $\Phi_{\max} = L\,\overline{f}\,\gamma$ is the maximally possible fitness flux, in which all L sites substitute after each flip (with rate γ) and contribute an average increase $\overline{f}$ to the population's mean fitness. Hence, the degree of adaptation, although defined via the genomic state, also serves as the realized fraction of the fitness flux in comparison to the optimal flux. This also means that any evolutionary mechanism that *slows down* evolution due to interference effects (clonal interference, multiple mutations, background selection) also degrades the genomic *state*, as will become clearer in section 3.3.

Approach to stationary equilibrium

In this scenario, we consider a static fitness landscape, that is time-independent. In this case, the stationary solution necessarily has fitness flux zero, following directly from equation 21 with $\gamma = 0$. We are in this case therefore more interested in non-stationary adaptation, such as the approach to a mutation-selection-drift equilibrium state. In this case, it follows from equation 21 that the fitness flux is simply the change in α

$$\Phi = \frac{\overline{f}\,L}{2}\,\frac{d\alpha}{dt} = \frac{d\,\langle F\rangle}{dt}, \tag{26}$$

where the right hand side of this equation follows from equation 15 and intuitively interprets the fitness flux as the speed of adaptation as it is traditionally defined. It is simply the rate of increase of the mean fitness $\langle F\rangle$. An approach to equilibrium is a relevant scenario for laboratory evolution experiments, in which populations of bacteria are evolved to adapt to some environment that is constant in time [4].

2 Fixation probability under linkage

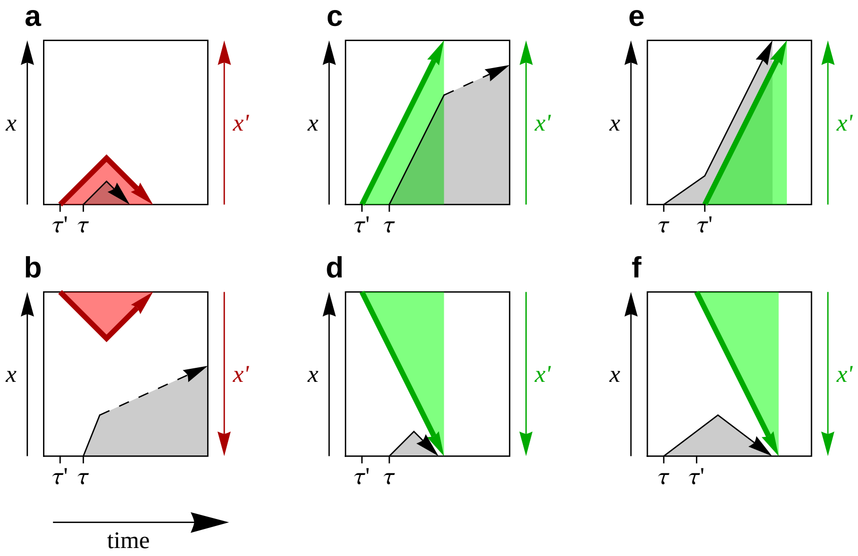
In this chapter, we derive the fixation probability of mutations under linkage. Our result applies to beneficial, neutral and deleterious mutations under general distributions of fitness effects. The full solution is a self-consistent summation of linkage-interferences and can be approximated by analytical expressions.

2.1 Pairwise interaction diagrams

Consider the situation from figure 2, where many linked mutations simultaneously segregate in the population and interfere with each other. We first classify the different pairwise interference interactions, shown in figure 2, systematically. We distinguish a *target mutation* from an *interfering mutation*: the target mutation is the one that we want to compute the fixation probability for, while the interfering mutation is a passive part of our calculation. This asymmetry is necessary to close the formalism, as will become clear later. We identify five different binary criteria that the two interacting mutations can be classified by:

- **Time ordering:** the target mutation appears *before* of *after* the interfering mutation.
- **Selection sign of the interfering mutation:** the interfering mutation can be *beneficial* or *deleterious*.
- **Selection sign of the target mutation:** also the target mutation can be either *beneficial* or *deleterious*.
- **Selection strength:** the interfering mutation can have a *larger* or *smaller* absolute selection coefficient than the target mutation.
- **Allele association:** interfering mutation and target mutation can be linked on the *same* or on a *different* genotypic background.

These five binary criteria yield $2^5 = 32$ different pairwise interaction scenarios. To be able to illustrate these interaction scenarios with reasonable overview, we drop two of the above criteria: First, we will only consider those interference interactions, in which the interfering mutation has a larger absolute selection coefficient than the target mutation, which is justified because weaker interfering mutations have a negligi ble effect on the target mutation. We refer to this assumption as the *hierarchy assumption.* Second, we will keep the sign of the selection coefficient of the target mutation undetermined in the diagrams. In systematically illustrating all pairwise interactions, this leaves $2^3 = 8$ different interference scenarios. Of these 8 types, we show 6 in Figure 3, leaving out the case of future deleterious interfering mutations, since these can never affect the fixation probability of the target mutation: Because future deleterious mutations will only affect a subpopulation of individuals that already carry the target mutation, there will always remain a fitter larger subpopulation *not* carrying the interfering deleterious mutation.



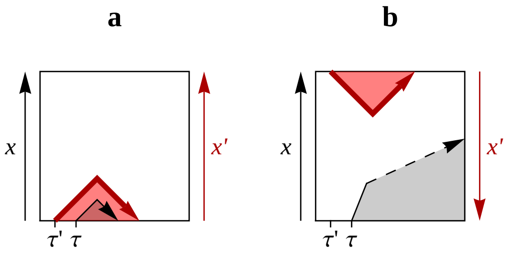
▲ **Figure 3. Linkage Interaction diagrams.** A target mutation with origination time τ and frequency $x(t)$ (black arrow) is subject to a stronger interfering mutation with origination time τ' and frequency $x'(t)$ (colored arrow). The interactions between this pair of mutations can be classified as follows: (a,b) *Interference by a deleterious background mutation* (red arrow):(a) The target mutation originates on the deleterious allele of the interfering mutation and is driven to loss, (b) the target mutation originates on the ancestral (beneficial) allele of the interfering mutation and is enhanced in frequency. (c,d) *Interference by a beneficial background mutation* (green arrow): (c) The target mutation originates on the beneficial allele of the interfering mutation and is enhanced in frequency, (d) the target mutation originates on the ancestral(deleterious) allele of the interfering mutation and is driven to loss. (e,f) *Interference by a beneficial future mutation* (green arrow): (e) The interfering mutation originates on the new allele of the target mutation and drives it to fixation, (f) The interfering mutation originates on the ancestral background of the target locus and drives the target mutation to loss.

In the following, we compute the conditional fixation probability $G(\sigma, \tau \,|\, \sigma', \tau')$ of a target mutation with selection coefficient σ and origination time τ, which is subject to an interfering mutation with selection coefficient σ' and origination time τ', for the different cases of Figure 3:

- **a, b:** Interference by deleterious background mutations ($\tau' < \tau$ and $\sigma' < 0 < |\sigma|$),
- **c, d:** Interference by beneficial background mutations ($\tau' < \tau$ and $\sigma' > |\sigma| > 0$)
- **e, f:** Interference by beneficial future mutations ($\tau' > \tau$ and $\sigma' > |\sigma| > 0$)

We neglect the effects of interference mutations weaker than the target mutations ($-|\sigma| < \sigma' < |\sigma|$), which is consistent with the hierarchy approximation. Note that we treat the fate of the target mutation probabilistically, while we treat interfering mutations as destined for fixation or extinction, depending on the sign of their selection coefficient.

Interference by deleterious background mutations



The diagrams of Figure 3 (a,b) describe background selection caused by strongly deleterious alleles originating before the target mutation. We assume for now that at the time where the target mutation appears, τ, the frequency of the background allele is at frequency x' with some probability $Q(x'; \tau', \sigma')$. For a given frequency x', we can then simply give the probabilities for positive (a) or negative (b) interference: Case a) occurs with probability x', while b) occurs with probability $1 - x'$. Following our deterministic assumption of interfering mutations, the first case results in the loss of the target mutation. Case b), however results in a small boost of the frequency of the target mutation, because the target mutation will gain a small advantage by the loss of a deleterious genotype not linked to it, as indicated in diagram b). Given the quick loss of the interfering mutation, we can model this frequency boost as an increase of the initial frequency of the target mutation by a factor $1/(1 - x')$. Together with the probability for case b) to occur in the first place, the resulting fixation probability of the target mutation is then

$$G(\sigma, \tau \mid \sigma', \tau') = \int_0^1 Q(x'; \sigma)\,(1 - x')\,G_0\left(\frac{1}{1 - x'}, \sigma\right) dx' \tag{27}$$

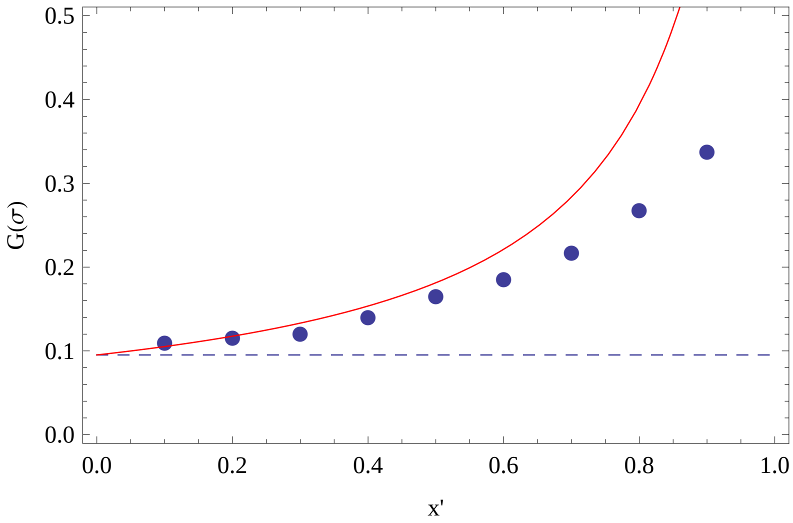
with Kimura’s unlinked fixation probability:

$$G_0(x_0, \sigma) = \frac{1 - e^{-2 N \sigma x_0}}{1 - e^{-2 N \sigma}}, \tag{28}$$

from which equation 12 is derived as the special case $x_0 = 1/N$. We can test the model of the frequency boost numerically by a simple Wright-Fisher model with the three haplotypes of diagram 3b:

- the wildtype (WT) with fitness 0.
- WT + target mutation with fitness σ.
- WT + background mutation with fitness σ'.

The fixation probability increase by diagram (b) can now be tested against simulations, as shown in Figure 4. For small frequencies of the background mutation, the prediction is very accurate, but deviates for large background frequencies and hence strong boosts. This is expected, because we assumed an immediate extinction of the interfering mutation, which overestimates the effect for larger initial frequencies of the background allele.



▲ **Figure 4. Frequency increase of a target mutation by deleterious background mutations.** We plot the fixation probability of a beneficial target mutation $2N\sigma = 10$ with a strongly deleterious background mutation with $2N\sigma' = -50$, that is not linked to the target mutation and initially present with frequency x'. The target mutation has initial frequency $x_0 = 0.01$. The dots are simulation results, the blue dashed line is the expected fixation probability without background selection, the red line is the theory prediction by the frequency increase $G_0(x_0/(1-x'), \sigma)$. The population size is $N = 1000$.

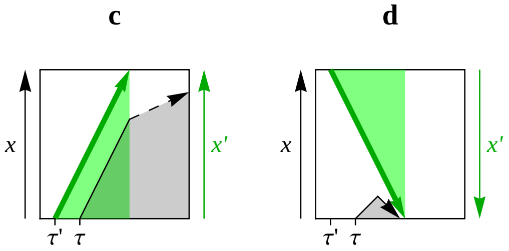
Because the interfering mutation is deleterious, its frequency distribution $Q(x'; \tau', \sigma)$ is dominated by very small frequencies $x' \ll 1$, for which our prediction is shown to be very accurate.

For $\sigma \geq 0$ and $x' \ll 1$, the fixation probability $G_0(x_0, \sigma) \approx 2\sigma$ is in good approximation linear in its first argument. In that case, the factor $(1 - x')$ cancels out and we recover the unlinked fixation probability $G(\sigma, \tau \,|\, \sigma', \tau') = G_0(1/N, \sigma)$. This argument does not apply to deleterious target mutations, but this case can be neglected because $G_0(x_0, \sigma)$ is exponentially small for $\sigma < 0$, even including a boost in the initial frequency.

The effects of background selection have been subject to a large number of articles. Often, these studies find that background selection in fact retains a substantial net effect on the fixation probability of a target mutation [59, 5, 55]. These studies typically assume a mutation-selection balance of many deleterious mutations, with a constant deterministic influx of deleterious mutations. For this case, an argument from Fisher [25] shows that the fixation probability of a beneficial mutation is reduced (see [5] and [55]) by a factor $\exp(-U_d / \sigma_d)$, where U_d is the rate and σ_d is the selection coefficient of deleterious mutations.

In our model, Fisher's argument does not hold, for two reasons: i) because of the presence of adaptive substitutions (selective sweeps), variance in the population is constantly removed, hence a stationary mutation-selection balance is never maintained; ii) because we consider an exponential distribution of selection coefficients, the number of deleterious mutations that are *stronger* in effect than the target mutation, are typically rare enough to be treated stochastically, as done in our pairwise interaction scheme using the diagrams of Figure 3.

Interference by beneficial background mutations



The diagrams of Figure 3(c,d) describe positive and negative interference by a selective sweep starting at time $\tau' < \tau$. As in diagrams (a,b), the interfering mutation is present at the time where the target mutation appears. Given that the frequency of the interfering mutation is x' at time τ, we can again distinguish the two cases: Case d) occurs with probability $1 - x'$, while c) occurs with probability x'. Both diagrams have opposite effect on the fate of the target mutation: d) describes partial hitchhiking, where the target mutation gets an advantage in the beginning, until the driver is fixed. Diagram d) results in extinction of the target mutation because we again treat the interfering mutation deterministically and as destined to fix. Similar to diagram b), we will again model the advantage of partial hitchhiking by a boost of the initial frequency x_0 of the target mutation by a factor $1/x'$. In contrast to diagram b) however, this frequency-boost can now be large, when the frequency of the past interfering mutation is still small at time τ. Hence, the net effect is non-zero, since the fixation probability (given by Kimura's equation 28) is non-linear for large initial frequencies. Also, since the interfering mutation is beneficial, we now have to explicitly account for the time-dependence on $\tau - \tau'$. Treating the interfering mutation as deterministic, its frequency x' at time τ is given by

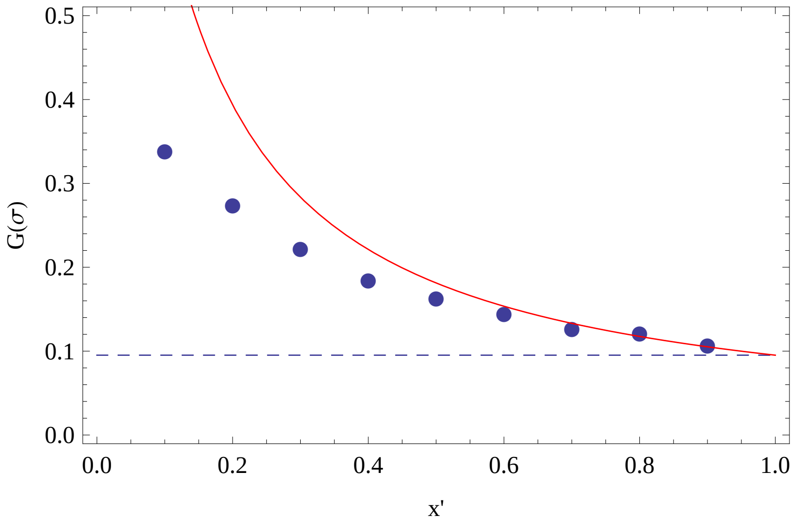
$$x' = x_{\text{det}}(\tau - \tau'; \sigma') = \frac{1}{1 + (N-1)\, e^{-\sigma'(\tau - \tau')}}, \tag{29}$$

which simply solves the deterministic evolution equation $dx/dt = \sigma\, x(1-x)$ (see [21]) under initial condition $x_{\text{det}}(0) = 1/N$. The full fixation probability under partial hitchhiking of the target mutation is then:

$$G(\sigma,\ \tau \,|\, \sigma',\ \tau') = \int_0^1 \delta(x' - x_{\text{det}}(\tau - \tau';\ \sigma'))\, x'\, G_0\!\left(\frac{1}{N\, x'},\ \sigma\right) dx', \tag{30}$$

where $\delta(x)$ is Dirac's Delta-distribution and $G_0(x_0,\ \sigma)$ is again Kimura's equation given by equation 28.

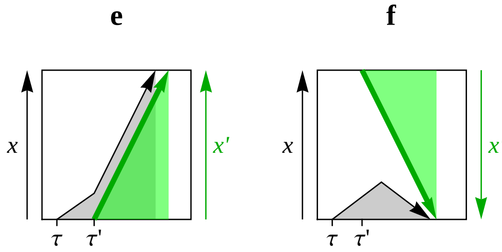
We show a comparison with the simulation scheme described above in Figure 5. The effect depends in an opposite way on x' than in Figure 4, because now the advantage is given by a factor $1/x'$ instead of $1/(1-x')$. The deviation between simulation and theory increases for small frequencies x', which corresponds to the case of very short times $\tau - \tau'$.



▲ **Figure 5.** **Frequency increase of a target mutation by partial hitchhiking.** We plot the fixation probability of a beneficial target mutation $2N\sigma = 10$ on the background of a strongly selected mutation with $2N\sigma' = 50$, initially present with frequency x'. The target mutation has initial frequency $x_0 = 0.01$. The dots are simulation results, the blue dashed line is the expected fixation probability without background selection, the red line is the theory prediction by the frequency increase $G_0(x_0/(1-x'),\ \sigma)$. The population size is $N = 1000$.

In a previously described model by Otto and Whitlock [56] the fixation probability in an expanding subpopulation is computed explicitly, which yields a result that is comparable to our approach (see their equation 11). Their result is more accurate than the expression we provide above (expanding the initial frequency in Kimura's formula). Here, modeling the effect of the expansion as an increase in the initial frequency is accurate enough. This is true in particular since the dominant interference effect is provided by future interfering mutations and not by background mutations.

Interference by future beneficial mutations



Finally, the diagrams of Figure 3(e,f) describe positive and negative interference by a selective sweep starting at time $\tau' > \tau$. These two diagrams are different from the diagrams (a-d), because we now consider the case where the interfering mutation appears *after* the target mutation, $\tau' > \tau$. In contrast to the previous cases, we now have to take into account the time-evolution of the target mutation explicitly. For that purpose, we will make use of the propagator $G_0(x, \tau' - \tau; x_0, \sigma)$, by which we denote the probability, that a polymorphism that appeared at time τ with frequency x_0 with selection coefficient σ will have reached frequency x at time τ'. Given this propagator, we can again simply distinguish between the two cases e) and f): Case (e) occurs with probability $x\, G_0(x, \tau' - \tau; x_0, \sigma)$, while case f) occurs with probability $(1 - x)\, G_0(x, \tau' - \tau; x_0, \sigma)$. Because of the deterministic nature of the interfering mutation, the first case results in fixation of the target mutation by complete hitchhiking, while the second case results in extinction of the target mutation. We therefore obtain the hitchhiking probability as an integral over all frequencies x:

$$G(\sigma, \tau \,|\, \sigma', \tau') = \int_0^1 x\, G_0(x, \tau' - \tau; x_0, \sigma)\, dx. \tag{31}$$

As will be derived in section 2.2, this integral can be solved in the diffusion approximation (equations 41 and 42). We get

$$G(\sigma,\, \tau \,|\, \sigma',\, \tau') = \begin{cases} G_0(x_0,\, \sigma) \big/ \left(1 + e^{-\hat{\sigma}(\tau'-\tau)}\left(G_0(x_0,\, \sigma)\, {x_0}^{-1} - 1\right)\right) & \text{for } \sigma > 0 \\ x_0\, e^{\hat{\sigma}(\tau'-\tau)} + \left(1 - e^{\hat{\sigma}(\tau'-\tau)}\right) G_0(x_0,\, \sigma) & \text{for } \sigma < 0 \end{cases}. \tag{32}$$

The regularized selection coefficient $\hat{\sigma}$ is a shorthand for the crossover from strong to weak selection: $\hat{\sigma} \simeq \sigma$ for $N\sigma \gtrsim 1$ and $\hat{\sigma} \simeq 1/2N$ for $N\sigma \lesssim 1$. The exact form of this crossover is not important. Here we choose

$$\hat{\sigma} = \begin{cases} 1/2N & \text{for } N\sigma \leq 1 \\ \sigma & \text{for } N\sigma > 1. \end{cases} \tag{33}$$

2.2 Single site propagator

The missing piece in the evaluation of the pairwise interaction diagrams is the integral in equation 31, which describes the time evolution of the mean allele frequency. Here we derive a solution within the diffusion approximation. The Fokker-Planck equation for a single site under drift and selection reads:

$$\partial_t G = \left[\frac{1}{2N}\, \partial_x^2 (x(1-x)) - \sigma\, \partial_x (x(1-x))\right] G. \tag{34}$$

To derive an expression for the integral $M(t,\, x_0,\, \sigma) = \int_0^1 x\, G_0(x,\, \tau' - \tau;\, x_0,\, \sigma)\, dx$, we multiply equation 34 with x and integrate by parts, neglecting boundary terms:

$$\partial_t \int_0^1 x\, G\, dx = \sigma \int_0^1 x(1-x)\, G\, dx \tag{35}$$

or

$$\partial_t M = \sigma\left(M - \int_0^1 x^2\, G\, dx\right) \tag{36}$$

We introduce the centered second moment $M_2 = \sigma\left(\int_0^1 x^2\, G\, dx - M^2\right)$ to write

$$\partial_t M = \sigma\, M(1-M) - M_2. \tag{37}$$

An exact solution of this equation can not be given, since M_2 depends on the third moment of G, which itself depends on the fourth moment, and this infinite chain of dependencies is not closed. We therefore make an heuristic ansatz for the solution of equation 37, which is motivated by the known limits of the solution: First, for $t = 0$, the variance term M_2 must vanish, because G is a Green's function and has the initial condition $G(x, 0; x_0, \sigma) = \delta(x - x_0)$. We then see from equation 37 that M evolves logistically with initial value $M(0) = x_0$. Secondly, for times larger than the characteristic polymorphism lifetime $\sim 1/\sigma$, any polymorphism initially present will have gone extinct or fixed. Therefore, G must become stationary after time $\sim 1/\sigma$ and will consist of two delta peaks at 0 and 1 with weights reflecting the fixation probability G_0.

$$G(x, t, x_0, \sigma) \xrightarrow{t \gtrsim 1/\sigma} (1 - G_0)\, \delta(x) + G_0\, \delta(1 - x), \tag{38}$$

and hence

$$M_2 \xrightarrow{t \gtrsim 1/\sigma} G_0(1 - G_0). \tag{39}$$

For the stationary first moment we get

$$\partial_t M_{\text{stat}} = 0 \quad \Longrightarrow \quad M_{\text{stat}} = G_0 \tag{40}$$

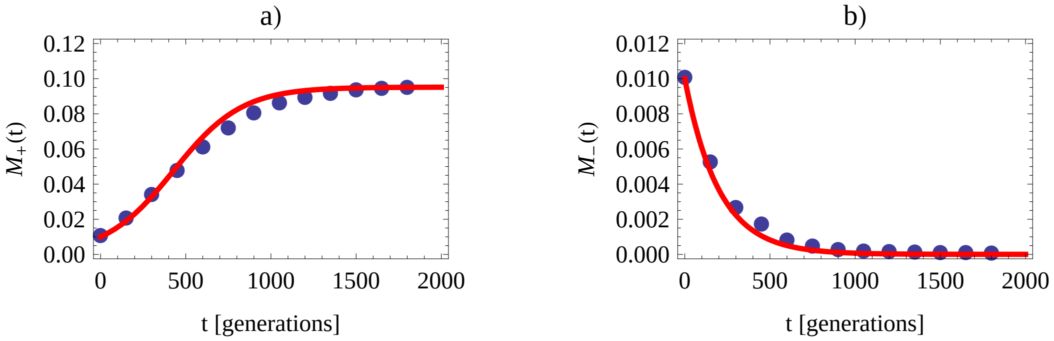
Knowing these two boundary cases and given the logistic form of the differential equation 37, we choose a logistic equation for beneficial and an exponential for deleterious mutations:

$$M_+(t, x_0, \sigma) = \frac{G_0(x_0, \sigma)}{1 + e^{-\hat{\sigma} t}\left(G_0(x_0, \sigma)\, x_0^{-1} - 1\right)} \tag{41}$$

and

$$M_-(t, x_0, \sigma) = x_0\, e^{-\hat{\sigma} t} + \left(1 - e^{-\hat{\sigma} t}\right) G_0(x_0, -\sigma) \tag{42}$$

with the standard Kimura fixation probability G_0, given by equation 28 and the regularized selection coefficient introduced in equation 33. Note that both of these equations yield the correct limit $M_\pm(t, x_0, \sigma) \to G_0(x_0, \sigma)$. This limit reflects that the target mutation is only interfered, if the second mutation appears within the lifetime of the target mutation. The heuristic predictions from equations 41 and 42 agree very well with simulations, as shown in Figure 6.



▲ **Figure 6.** **Single site propagator.** This plot shows the mean frequency of a beneficial (a) and deleterious (b) mutation as a function of time, measured in generations. Black circles are data obtained from simulations, red solid lines are the theory predictions of equations 41 and 42. Simulation data has been obtained from many trajectories started with different random seeds. Error bars indicate the standard error of the mean. Parameters: $N = 1000$, $2N\sigma = 10$, $x_0 = 10/N$.

2.3 From pairwise to many locus interactions

Rate of drivers

We now derive an approximate expression for the total fixation probability of a target mutation based on pair interactions with multiple interfering mutations. Clearly, a straightforward “cluster expansion” makes only sense in a regime of *dilute* events at sufficiently low rates of beneficial mutations, where the interference interactions of Figure 3(c-f) are infrequent. However, we are primarily interested in adaptive processes under linkage in the *dense-interactions* regime at high rates of beneficial mutations (the crossover between these regimes is further quantified below). In this regime, dense beneficial mutations generate strongly correlated clusters of fixed mutations nested in each other's background, called *selective sweeps*. This nesting has the simple topological reason that without recombination, no two beneficial mutations can fix simultaneously if they are not nested in the same cluster. We treat the dense-interactions regime by an approximation: Each sweep is associated with a unique *driver mutation*, which is the strongest beneficial mutation in its cluster. The driver mutation itself evolves free of interference, but it influences other mutations by interference; that is, we neglect the feedback of weaker beneficial and deleterious mutations on the driver mutation. This hierarchy approximation has already been anticipated in section 2.1, where we neglected pairwise interactions of a weaker interfering mutation with a stronger target mutation. Here, we use it to derive the *rate* of driver mutations.

The coherence time of a selective sweep is set by the fixation time of its driver mutation

$$\tau_{\text{fix}}(\sigma) = \frac{2\log(2N\sigma)}{\sigma}, \tag{43}$$

which is the solution of equation 29 for starting frequency $x_0 = 1/(2N\sigma)$ and final frequency $x = 1 - 1/(2N\sigma)$. These frequencies determine the regime of deterministic growth, in contrast to the drift dominated boundaries [63]. Since driver mutations are by definition non-interfering (because they are the strongest mutation in their sweep-cluster), we can treat their occurrence approximately as a Poisson process. The sweep rate is then equal to the rate of driver mutations, $V_{\text{drive}}(\sigma)$, and is given by the condition that *no stronger* selective sweep occurs during the interval $\tau_{\text{fix}}(\sigma)$. This condition can be written as a waiting time probability, which is a negative exponential

$$p_{\text{drive}}(\sigma) = e^{-\tau_{\text{fix}}(\sigma)\,V_{>}(\sigma)} \tag{44}$$

with the total rate of *stronger* driver mutations:

$$V_{>}(\sigma) = \int_{\sigma}^{\infty} V_{\text{drive}}(\varsigma)\,d\varsigma \tag{45}$$

and the rate of drivers, which again depends circularly on the driver probability:

$$V_{\text{drive}}(\sigma) = p_{\text{drive}}(\sigma)\,G_0(\sigma)\,U(\sigma). \tag{46}$$

Because of the circular dependence between equations 44, 45 and 46, they have to be solved *self-consistently*, which can only be achieved numerically (see section 3.2). This self-consistent solution can be regarded as a partial summation of higher-order interference interactions characteristic of the dense-sweep regime [68].

We note that the arguments leading to equations 46 and 44 must be modified, if the distribution of selection coefficients $\rho(f)$ falls off much faster than exponentially (as shown numerically by Fogle et al. [27]). In that case, selective sweeps may contain several driver mutations of comparable strength. In section 3.5, we show simulation results for non-exponential distributions and show that our approximations still work to a reasonable degree. But we clearly expect our hierarchy assumption to break down if selection coefficients become very similar.

We can compare our derivation to the model of clonal interference by Gerrish and Lenski [28] (GL), which has some similarities to our approach. The GL-model determines an approximation of the sweep rate, $V_{\mathrm{GL}}(\sigma)$, by requiring that no *negative* interference by a future mutation occurs (see section 3.3). In GL-theory, the fixation probability is

$$G_{\mathrm{GL}}(\sigma) = U(\sigma)\, G_0(\sigma) \exp\left(-\frac{1}{2}\, \tau_{\mathrm{fix}}(\sigma) \int_{\sigma}^{\infty} U(\sigma')\, G_0(\sigma')\, d\sigma'\right), \qquad (47)$$

where the fixation time is given by equation 43. The factor 1/2 in the exponent follows from counting only those stronger mutations that appear on the background *not carrying* the target mutation. Only these mutations decrease the fixation probability of the target mutation.

Equation 47 is analogous to equation 44, in which we require that a driver mutation is not interfered with by any stronger driver mutation. There are, however, two important differences: i) Equation 44 has no factor 1/2, since we exclude from the driver rate those mutations that fix by positive interference (hitchhiking, diagram 3e). In contrast, in GL-theory, the only mode of fixation are driver mutations that are not suffering negative interference. ii) Equation 44 reflects a self-consistent closure, in which each interfering driver must itself be free from even stronger interfering drivers. Therefore, apart from the factor 1/2, equation 47 can be seen as a first iteration loop of our self-consistent driver rate (equation 46). Taking into account hitchhiking as a positive outcome of interference dramatically enhances the fixation probability of weakly beneficial mutations and in particular allows us to compute the influence on deleterious alleles, a case which is not covered by GL-theory.

Full fixation probability

We now evaluate the pairwise interaction diagrams from Figure 3 in the context of the driver rate derived above. We need to find a way to combine the different diagrams into one fixation probability. We again summarize their expected contributions: diagrams 3a) and b) describe deleterious background selection and we showed that positive and negative interference cancel out exactly, because of the expected low frequency of background deleterious alleles. Diagrams c) and d) describe interference by a past driver mutation and these interactions retain a net effect on the fixation probability of the target mutation. Finally, diagrams e) and f) describe interference by future drivers and they have a strong expected contribution.

Based on the pairwise cancellation of diagrams 3a) and b) and the results on the remaining diagrams, we make a combination ansatz for multiple interaction scenarios: We assume that of the many potentially interfering mutations, exactly two will strongly interfere with the target mutation: the *last* driver mutation *before* its origination (with parameters $\sigma' > \sigma$ and $\tau' < \tau$), and the *first* driver *after* its origination (with parameters $\tau'' > \tau$ and $\sigma'' > \sigma$). These two drivers affect the fixation probability $G(\sigma)$ in a combined way: the target mutation can only be fixed if it appears on the background of the last background sweep *and* if it is itself the background of the first future sweep. The resulting conditional fixation probability of the target mutation, $G(\sigma, \tau \,|\, \sigma', \tau', \sigma'', \tau'')$, is a straightforward combination of equations 30 and 31:

$$G(\sigma, \tau \,|\, \sigma', \tau', \sigma'', \tau'') = \\ \int_0^1 dx' \int_0^1 dx\, \delta(x' - x_{\text{det}}(\tau - \tau'; \sigma'))\, x'\, x\, G_0\!\left(x, \tau'' - \tau; \frac{1}{N x'}, \sigma\right). \tag{48}$$

This expression is based on the assumption that the two sweeps act sequentially and independently, that is, the target mutation can only fix if it is free of interference or positively interfered with by both sweeps. Interactions between the sweeps themselves are neglected (such as rescue of the target mutation by a future sweep, following negative interference by a past sweep). This is in tune with our self-consistent determination of the sweep rate, which absorbs the overlap exclusion between driver mutations (i.e., the condition $\tau'' - \tau' > \tau_{\text{fix}}(\sigma')$) into a reduced uniform or “mean-field” rate $V_{\text{drive}}(\sigma)$, given by equations 46 and 44. In this approximation, a target mutation of selection coefficient σ is subject to interference by stronger selective sweeps at a total rate $V_>(\sigma)$. We can now integrate equation 48 over past and future sweeps (i.e., driver mutations) with an exponential distribution of waiting times $\tau - \tau'$ and $\tau'' - \tau$,

$$G(\sigma) = \int_{-\infty}^{\tau} d\tau' \int_{\tau}^{\infty} d\tau'' \int_{\sigma}^{\infty} d\sigma' \int_{\sigma}^{\infty} d\sigma'' \times \\ V_{\text{drive}}(\sigma')\, V_{\text{drive}}(\sigma'')\, e^{-V_>(|\sigma|)\,(\tau'' - \tau')}\, G(\sigma, \tau \,|\, \sigma', \tau', \sigma'', \tau''). \tag{49}$$

The exponential factor is based on an argument similar to the derivation of the driver probability, equation 44: Driver mutations are by construction independent and form a Poisson process, so the waiting time *between* drivers is again exponential. Using equation 32, the integrations over σ'' and τ'' can be treated analytically, and we obtain

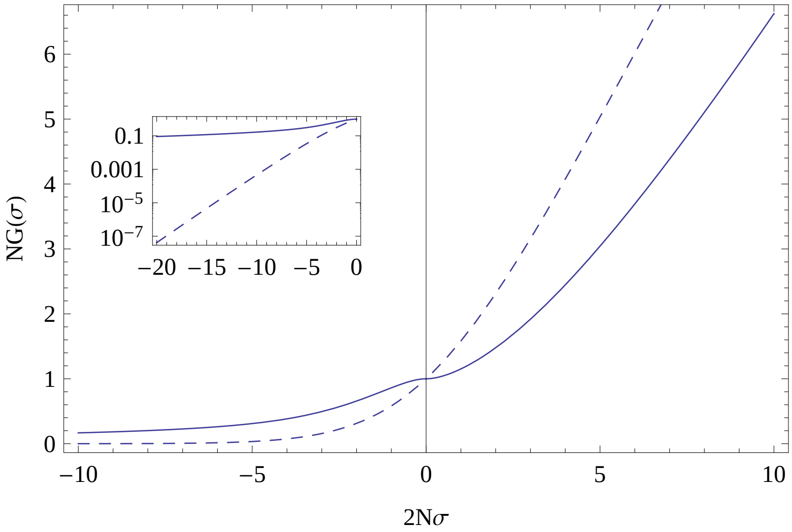
$$
\begin{aligned}
G(\sigma) = {} & \\
& \int_{-\infty}^{\tau} d\tau' \int_{|\sigma|}^{\infty} d\sigma' \, V_{\mathrm{drive}}(\sigma') \, e^{-V_>(|\sigma|)\,(\tau-\tau')} \int_0^1 dx' \, \delta(x' - x_{\mathrm{det}}(\tau-\tau'; \sigma')) \\
& \times \begin{cases} x' \, G_0\!\left(\frac{1}{N\,x'}, \sigma\right) {}_2F_1\!\left(1, \frac{V_>(\sigma)}{\hat{\sigma}}; 1 + \frac{V_>(\sigma)}{\hat{\sigma}}; 1 - N\,x' \, G_0\!\left(\frac{1}{N\,x'}, \sigma\right)\right) & (\sigma > 0), \\ \frac{1}{N\left(|\hat{\sigma}| + V_>(|\sigma|\right)} \left(N\,x' \, G_0\!\left(\frac{1}{N\,x'}, \sigma\right) |\hat{\sigma}| + V_>(|\sigma|)\right) & (\sigma < 0), \end{cases}
\end{aligned}
\tag{50}
$$

where ${}_2F_1(a, b; c; z)$ is the hypergeometric function [68]. The remaining integrals in this expression can be evaluated numerically, using standard numerical integration methods, as implemented for example by *Mathematica*'s *NIntegrate* routine.

In order to understand the dependence of the fixation probability on the driver rate better, we try to approximate equation 50 by a closed expression. A simple approximation can be achieved by neglecting the integral over past sweeps. As argued in section 2.1, the net effect of past sweeps is small in comparison to the effect by future sweeps. We therefore obtain a closed form of equation 50 by setting $x' = 1$ and omitting the integral over x':

$$
\begin{aligned}
G(\sigma) = {} & \\
& \begin{cases} G_0\!\left(\frac{1}{N}, \sigma\right) {}_2F_1\!\left(1, \frac{V_>(\sigma)}{\hat{\sigma}}; 1 + \frac{V_>(\sigma)}{\hat{\sigma}}; 1 - N\,x' \, G_0\!\left(\frac{1}{N}, \sigma\right)\right) & (\text{for}\ \sigma > 0), \\ \frac{1}{N\left(|\hat{\sigma}| + V_>(|\sigma|\right)} \left(N \, G_0\!\left(\frac{1}{N}, \sigma\right) |\hat{\sigma}| + V_>(|\sigma|)\right) & (\text{for}\ \sigma < 0). \end{cases}
\end{aligned}
\tag{51}
$$

An evaluation of this equation for a *constant* driver rate $N\,V_>(\sigma) \equiv 1$ is shown in Figure 7. We observe two main features of the fixation probability in comparison to the classical Kimura results (equation 12). First, both beneficial and deleterious mutations have a fixation probability that is more neutral than expected without linkage. Secondly, the fixation probability of deleterious mutations is dramatically increased: while the classical result predicts an exponential decrease $\sim \exp(-2\,N\,|\sigma|) / |\sigma|$ our result predicts an algebraic decrease $\sim V_>(|\sigma|) / |\sigma|$ (see inset of Figure 7). Note however, that $V_>$ will itself become exponentially small for large $|\sigma|$, because of the tail of the (typically exponential) distribution of selection coefficients, $\rho(f)$. Both of these effects will be discussed in more detail in sections 2.4 and 3.3.



▲ **Figure 7. Fixation probability under linked drivers.** This figure shows the fixation probability under linkage with a uniform driver rate $N\,V_> = 1$. Here we use the approximate closed form equation 51. In comparison to the standard Kimura equation (shown as dashed curve), the beneficial fixation probability is reduced, while the deleterious fixation probability is dramatically increased, as emphasized in the inset.

Knowing the full fixation probability under driver mutations, we can now distinguish between *Driver* and *Passenger* contributions. Since we know that the driver contribution is given by equation 44 as $p_{\text{drive}}(\sigma)\,G_0(\sigma)$, we can define the passenger part as

$$G_{\text{pass}}(\sigma) = \frac{G(\sigma) - p_{\text{drive}}(\sigma)\,G_0(\sigma)}{1 - p_{\text{drive}}(\sigma)}, \tag{52}$$

which is discussed in more detail in section 3.3.

2.4 Selection regimes and emergent neutrality

With an increasing rate of driver mutations, the fixation probability of beneficial mutations decreases, and the fixation probability of deleterious mutations increases (see Figure 7). This suggests a reduction factor of the efficacy of selection. Indeed, a Taylor expansion of equation 51 gives:

$$G(\sigma) = \frac{1}{N} + \frac{\sigma}{1 + 2\,N\,V_>(\sigma)} + O\!\left(\sigma^2\right) \tag{53}$$

We can compare this expansion with the corresponding expansion of Kimura's formula (equation 12):

$$G_0(\sigma) = \frac{1}{N} + \sigma\left(1 - \frac{1}{N}\right) + O\left(\sigma^2\right) \approx \frac{1}{N} + \sigma + O\left(\sigma^2\right) \tag{54}$$

We see that linkage leads to a linear reduction of the strength of selection by a factor $1 + 2\,N\,V_>(\sigma) \approx 1 + 2\,N\,V_>(0)$ compared to unlinked sites. We define the threshold of emergent neutrality as $2\,N\,\tilde{\sigma} = 1 + 2\,N\,V_>(0)$. For mutations of sufficiently weak effect, the fixation probability then takes the particularly simple form

$$G(\sigma) \simeq G_0\left(\frac{\sigma}{2\,N\,\tilde{\sigma}}\right) \quad (\text{for } -\tilde{\sigma} < \sigma < \tilde{\sigma}), \tag{55}$$

where the neutrality threshold $\tilde{\sigma}$ is given by the total sweep rate $V_>(0) = \int_0^\infty V_{\text{drive}}(\varsigma)\, d\varsigma$

$$\tilde{\sigma} = \frac{1}{2\,N} + V_>(0). \tag{56}$$

These equations show how neutrality emerges for strong adaptive evolution under linkage [68]. Specifically, the relation for $\tilde{\sigma}$ delineates two dynamical modes: the *dilute sweep mode* ($V_>(0) \lesssim 1/2\,N$), where the neutrality threshold is set by genetic drift to the Kimura value $\tilde{\sigma} \simeq 1/2\,N$ (see [41]), and the *dense sweep mode* ($V_>(0) \gtrsim 1/2\,N$), where interference effects generate a broader neutrality regime with $\tilde{\sigma} \simeq V_>(0)$. The transition between these modes marks the onset of clonal interference as defined in previous work [77, 57]. For stationary adaptation in a time-dependent fitness seascape, the upper bound $V_>(0) \simeq \gamma\,L$ produces the estimate $2\,N\,\gamma\,L > 1$ for the crossover from dilute to dense sweeps. However, equations 55 and 56 remain valid for non-stationary adaptation, where the neutrality threshold $\tilde{\sigma}$ becomes time-dependent (see below).

We also anticipate results seen in section 3.3: Strongly beneficial mutations do not show emergent neutral behavior. In contrast, for strongly deleterious mutations, we see a reduction of the efficacy of selection similar to the emergent neutrality regime.

In summary, interference interactions in the dense-sweep mode produce the following selection classes of mutations and genomic sites:

1. **Emergent neutrality regime:** Mutations with selection coefficients $-\tilde{\sigma} < \sigma < \tilde{\sigma}$ fix predominantly as passenger mutations. Their near-neutral fixation probability (equation 55) is the joint effect of positive and negative interference. Compared to unlinked mutations, $G(\sigma)$ is reduced for beneficial mutations and enhanced for deleterious mutations. Accordingly, sites with selection coefficients $f < \tilde{\sigma}$ have near-random probabilities of their alleles.
2. **Adaptive regime:** Mutations with effects $\sigma > \tilde{\sigma}$ have a fixation rate significantly above the neutral rate and, hence, account for most of the fitness flux. Moderately beneficial mutations ($\sigma \gtrsim \tilde{\sigma}$) still fix predominantly as passengers, whereas strongly beneficial mutations ($\sigma \gg \tilde{\sigma}$) are predominantly drivers. Hence, the fixation rate increases to values of order $G_0(\sigma) \simeq 2\,\sigma$, which are characteristic of unlinked mutations. Accordingly, sites with $f > \tilde{\sigma}$ evolve towards a high degree of adaptation.
3. **Deleterious passenger regime:** Mutations with $\sigma < -\tilde{\sigma}$ can fix by positive interference, i.e., by hitchhiking in selective sweeps. This effect drastically enhances the fixation rate in comparison to the unlinked case (see Inset of Figure 7). It follows the heuristic approximation $G(\sigma) \approx G_{\text{pass}}(\sigma) \sim \exp(-|\sigma|/\tilde{\sigma})$, which extends the linear reduction in the effective strength of selection, equation 55, obtained in the emergent neutrality regime.

The cross-over between adaptive and emergent-neutrality regime implies a non-monotonic dependence of the substitution rate $V(\sigma)$ on the population size: In sufficiently small populations sizes (where $\tilde{\sigma} < \sigma$), beneficial mutations of strength σ are likely to be driver mutations. Hence, $V(\sigma)$ is an increasing function of N with the asymptotic behavior $V(\sigma) \simeq \mu\,N\,\sigma$ familiar for unlinked sites. In larger populations (where $\tilde{\sigma} > \sigma$), the same mutations are likely to be passenger mutations and $V(\sigma)$ decreases with N towards the neutral rate $V(\sigma) \simeq \mu$. The maximal substitution rate is expected to be observed in populations where $\tilde{\sigma}$ is similar to σ. By the same argument, deleterious mutations have a minimum in their substitution rate in populations where $\tilde{\sigma}$ is similar to $|\sigma|$.

3 Analysis of adaptive scenarios

Here we analyze our results for two specific adaptive scenarios: stationary adaption in a fitness seascape and approach to equilibrium in a static fitness landscape. Detailed comparisons of our analytical results with numerical simulations show that our approach is valid in both cases. We show that genomic linkage can drastically affect process and state in large asexual populations: Although adaptation generates beneficial *driver* mutations, a substantial fraction of sequence changes are *passenger* mutations, whose chance of fixation depends only weakly on their fitness effect. In particular, we find a regime of emergent neutrality with a threshold $\tilde{\sigma}$, which is time-dependent for non-stationary processes. Due to this *effectively neutral* dynamics, a fraction of genomic sites has nearly random fixed alleles, which do not reflect the direction of selection at these sites. Thus, linkage interactions not only reduce the speed of adaptation, but also degrade the genome state and the population's fitness in its current environment.

3.1 Computer simulation scheme

We represent the population as a list of haplotypes $(\boldsymbol{a}_k, n_k, f_k)$ with $k = 1 \dots n \leq N$. Each haplotype is a structured data type that consists of a sequence $\boldsymbol{a}_k$ (which is a bit-string in our case of two alleles per locus), the "occupation number" of individuals carrying that haplotype, n_k, and the fitness of the haplotype f_k. In each generation, we first apply mutations, then we apply selection flips, and finally we sample the next generation from the current generation.

Mutations are realized by creating new haplotypes. More specifically, every generation we draw the number of mutations from a Poisson distribution with mean $\mu N L$. For every mutation, we then randomly draw a haplotype and a site. The haplotype is drawn with the weight corresponding to the number of individuals carrying that haplotype. We create a new haplotype $(\boldsymbol{b}_k, 1, g_k)$ as a copy of the mutating haplotype with one individual, a mutated sequence vector $\boldsymbol{b}_k$ and a modified fitness g_k. The new fitness is computed by evaluating the fitness landscape, equation 2 or 6, on the new sequence.

We draw the number of selection flips each generation from a Poisson distribution with mean γL. For each flip we draw a random location and flip the selection coefficient of the fitness landscape. We then recompute the fitness of each haplotype in the population accordingly.

Finally, the next generation is determined as new occupation numbers $m_1, m_2 \ldots m_n$, drawn as multinomial random deviates with the probability distribution

$$P(m_1, m_2 \ldots m_n) = \frac{N!}{\prod_{k=1}^{n} m_k!} \prod_{k=1}^{n} \left(n_k\, e^{f_k - \bar{f}} \right)^{m_k} \tag{57}$$

with the mean fitness

$$\bar{f} = \log\left(\sum_{k=1}^{n} n_k\, e^{f_k} \right). \tag{58}$$

If the drawn number of individuals of some haplotype is zero ($m_k = 0$), we delete the corresponding haplotype. The multinomial random deviate is realized by the conditional binomial method [17]. This method uses binomial random deviates, for which we use the *Boost-C++ Libraries* [9].

In a running simulation, substitutions are observed if sites that were monomorphic at some allele become polymorphic and then fixed at the other allele. Each substitution can be categorized as *beneficial*, if the new allele is the currently fitter, or *deleterious* if it is the currently less fit of the two alleles. We also keep track of the fixation state: Over a sufficiently large number of generations, we can compute $\lambda_{\pm,j}$ of site j as the fraction of time at which the frequency of the currently fitter allele was below (above) 0.5. Since we generally consider low local mutation rates $\mu N \ll 1$, this fraction yields approximately the fraction of time the population was *fixed* at the locally fitter (less fit) allele. At every site, the origination rate of new beneficial (deleterious) mutations can be computed as $\mu \lambda_{\pm,j}$. Fixation probabilities can be computed by dividing the rate of beneficial/deleterious substitutions by the rate of beneficial (deleterious) mutations.

For the stationary adaptation simulations (section 3.3), we let the population reach stationarity for $\sim 1/\gamma$ generations before initiating any measurement to ensure that the population is in the stationary state. In case of the approach to equilibrium (section 3.4), we first run the above protocol for stationary adaptation (for some parameter γ as given in the particular presentation of the results) for sufficiently long time to ensure stationarity. Then we stop flipping, by setting $\gamma = 0$ and start obtaining measurements as described above. All results are now time-dependent. To obtain averages, we repeat this program many times and average over the full ensemble of simulations for each time point.

The theoretical derivations for our model suggest some simple scaling laws. As can be seen, the population size N and the genome length L are only relevant in the parameter combinations $(N\,L\,\mu)$, $(N\,L\,\gamma)$ and $\left(N\,\overline{f}\right)$. As long as we keep these parameter combinations fixed, we can use smaller values for N and L than in real populations to speed up the simulations. This scaling holds up to the following conditions: i) $\mu\,L \ll 1$ to avoid two mutations in the same individual in the same generation, ii) $\mu\,N \ll 1$ so that sites follow substitution dynamics with short polymorphic times, and iii) $N\,\gamma \ll 1$ so that the time between selection flips is larger than the time needed for a fixation. These conditions can always be fulfilled for given parameter products $(N\,\mu\,L)$, $(N\,\gamma\,L)$ and $\left(N\,\overline{f}\right)$.

3.2 Numerical solution

Here we show how the coupled set of equations derived above can be solved. For the following, we define a set of discrete values for the selection coefficients $f_i \geq 0$ with $i = 1 \ldots n$. We typically choose at least $n = 40$ equidistant points in the interval $0 \leq f_i \leq 10\,\overline{f}$.

General solution

Given the genome state at some time t_0, $\lambda_\pm(t_0,\ f_i)$, we want to propagate equation 19 a small time-interval Δt. This is achieved via the following steps:

1. Use $\lambda_\mp(t_0,\ f_i)$ to compute the mutation rate $U(\pm\, f_i)$ from equation 10.
2. Use $U(\pm\, f_i)$ to compute the driver rate $V_{\text{drive}}(f_i)$ from equations 46 and 44.
3. Use $V_{\text{drive}}(f_j)$ to compute the cumulative rate of drivers,

$$V_>(f_i) = \sum_{j=i+1}^{n} (f_j - f_{j-1})\,(V_{\text{drive}}(f_j) - V_{\text{drive}}(f_{j-1})), \tag{59}$$

4. Use $V_>(f_i)$ to compute the fixation probabilities $G(\pm f_i)$ from equation 50.
5. From $U(\pm f_i)$ and $G(\pm f_i)$ compute the substitution rate $V(\pm f_i)$ via equation 11.
6. Compute the genome state at time $t + \Delta t$ from equation 19:

$$\begin{aligned}\lambda_+(t_0+\Delta\mathrm{t},\ f_i) = \lambda_+(t_0,\ f_i) + \frac{1}{L\rho(f_i)}\,(V(f_i) - V(-f_i)) + \\ \gamma(\lambda_+(f_i) - \lambda_-(f_i))\,\Delta\mathrm{t} = 1 - \lambda_-(t_0+\Delta\mathrm{t},\ f_i)\end{aligned} \tag{60}$$

Iterative stationary solution

The selfconsistent solution of the dynamics is found by the following numerical procedure: We initialize the iteration by setting $V_{\mathrm{drive}}(f_i) \equiv 0$. We then iterate the following steps:

1. As above, use $V_{\mathrm{drive}}(f_j)$ to compute the cumulative rate of drivers (equation 59).
2. Use $V_>(f_i)$ to compute the fixation probabilities $G(\pm f_i)$ from equation 50.
3. Use $G(\pm f_i)$ to compute the stationary state probabilities $\lambda_+(f_i) = 1 - \lambda_-(f_i)$ from equation 22.
4. Use $\lambda_\mp(f_i)$ to compute the mutation rate $U(\pm f_i)$ from equation 10.
5. Use $U(\pm f_i)$ to compute the driver rate $V_{\mathrm{drive}}(f_i)$ from equations 46 and 44. Go back to step 1.

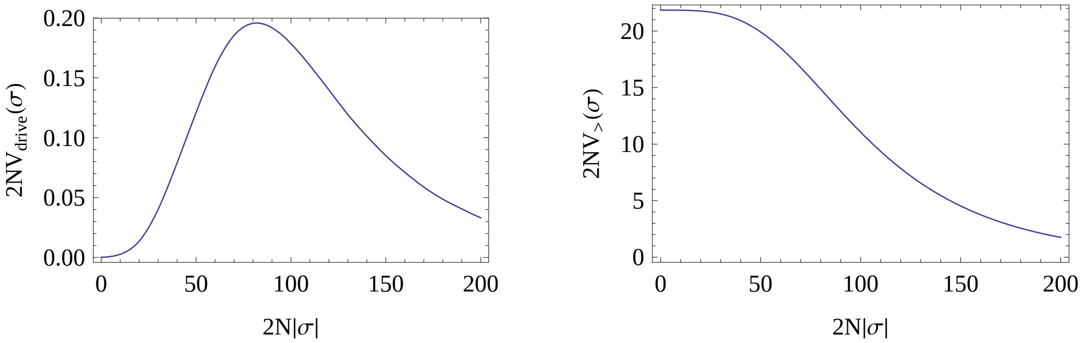
For the parameters used in this work, this algorithm converges already after 6 iterations.

3.3 Stationary adaptation

Here we show results of simulations and theory for stationary adaptation in our minimal fitness seascape (equation 6). This stationary solution is characterized by ongoing selection flips, which occur with rate γ per site and generate an excess of beneficial over deleterious substitutions, with rates $V(\sigma) > V(-\sigma)$ (see equation 19).

Driver rate

A key concept of our solution is the rate of driver mutations, $V_{\mathrm{drive}}(\sigma)$ and the cumulative rate of drivers, $V_>(\sigma)$, defined by equations 46 and 45. These two quantities are themselves an outcome of the dynamics, since they depend on the mutation rate and hence on the genome state. An example is shown in Figure 8, where we show the driver rate and the cumulative driver rate in stationarity. Note that the non-monotonicity of the driver rate comes from the underlying exponential distribution of selection coefficients. The driver rate is small for very large selection coefficients, because not many mutations with that high selection coefficients occur. It is again low for very small selection coefficients, because the driver probability $p_{\text{drive}}(\sigma)$ is low for these mutations (see equation 44). Therefore, the cumulative rate of drivers, $V_>(\sigma)$ saturates for low selection coefficients.

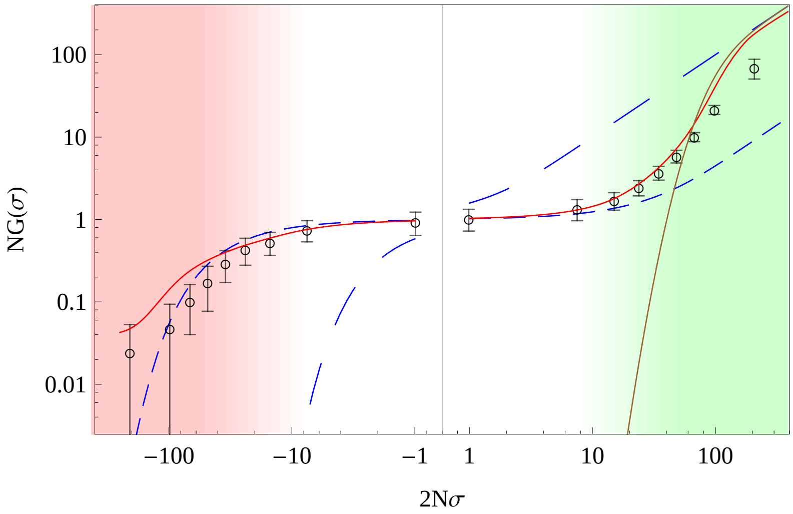


▲ **Figure 8.** **Driver rate under stationarity.** This figure shows the driver rate $V(\sigma)$ and the cumulative driver rate $V_>(\sigma)$. Parameters are $N = 2000$, $L = 1000$, $2N\gamma = 0, 1$, $2N\mu = 0.025$ and $2N\bar{f} = 50$.

Fixation probability

The effect of the driver rate on the fixation probability is shown in Figure 9, which shows the selection-dependent fixation probability $G(\sigma)$ in a linked genome undergoing stationary adaptive evolution. The emergent neutrality regime ($|\sigma| < \tilde{\sigma}$), the adaptive regime ($\sigma > \tilde{\sigma}$), and the deleterious passenger regime ($\sigma < -\tilde{\sigma}$) are marked by color shading (using equation 56). The self-consistent solution of our model (red line) is in good quantitative agreement with simulation results for a population of linked sequences (open circles). Data and model show large deviations from single-site theory (long-dashed blue line), which demonstrate strong interference effects in the dense-sweep regime.

Figure 9 also shows an effective single-site probability with a globally reduced efficacy of selection, $G_0(\sigma / 2N\tilde{\sigma})$ (short-dashed blue line). While this correctly predicts mutations in the emergent neutrality regime and in the deleterious passenger regime, it fails to capture the adaptive regime, where mutations have a fixation probability approaching the single-site value 2σ.



▲ **Figure 9. Fixation probability in stationary.** Selection-dependent fixation probability of mutations $G(\sigma)$, scaled by the population size N. Analytic model solution (red line) and simulation results (circles) show three regimes of selection: (i) Effective neutrality regime (white background): $G(\sigma)$ takes values similar to the fixation probability of independent sites with reduced selection, $G_0(\sigma/2N\tilde{\sigma})$ (short-dashed blue line). (ii) Adaptive regime (green): $G(\sigma)$ crosses over to the fixation probability for unlinked sites with full selection, $G_0(\sigma)$ (long-dashed blue line). The strong-selection part of this crossover is captured by the Gerrish-Lenski model, $G_{\mathrm{GL}}(\sigma)$ (brown line). (iii) Strongly deleterious passenger regime (red): $G(\sigma)$ is exponentially suppressed, but drastically larger than for unlinked sites (long-dashed blue line) due to hitchhiking in selective sweeps. Simulation results are averaged over intervals of σ, with error bars obtained by the standard deviation within an interval.

It is instructive to compare our results for stationary adaptation with one particular mutation-based model, namely the classical clonal interference model by Gerrish and Lenski (GL) [28]. As discussed in section 2.3, GL-theory predicts the fixation probability of a beneficial mutation in a manner similar to how we compute the rate of driver mutations. However, to directly compare our model with the GL-model, we need to make an important assumption on the mutation rate. Our genomic model yields beneficial mutation rates and their distribution as an *outcome*, while GL-theory takes distribution and rate of beneficial mutations as an *input*. In GL-theory all stronger driver mutations in the exponent of equation 47 are assumed to be free of interference. A straight-forward comparison between our theory and GL-theory is possible, if we use the following mutation rate as an input for GL-theory:

$$U_{\mathrm{GL}}(\sigma) = \lambda_{-}^{\mathrm{GL}}(\sigma)\,\rho(\sigma) \tag{61}$$

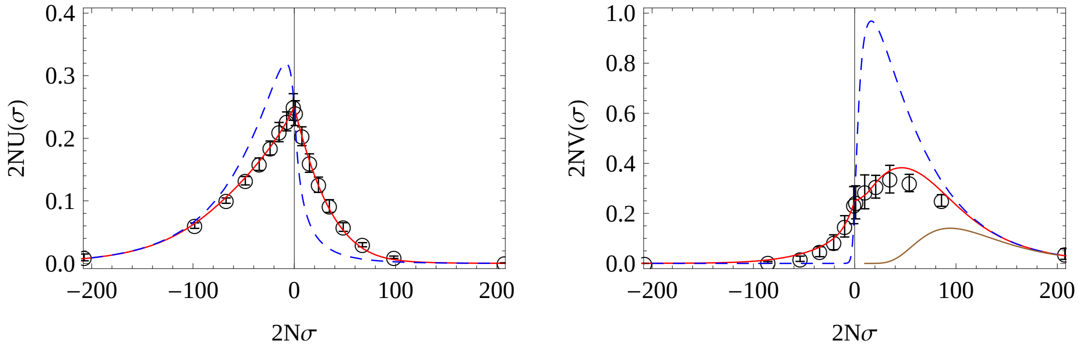
with

$$\lambda_{-}^{\mathrm{GL}}(\sigma) = \frac{N\,G_0(-\sigma) + \gamma/\mu}{N\,G_0(\sigma) + N\,G_0(-\sigma) + 2\,\gamma/\mu} \tag{62}$$

with the unlinked fixation probability, equation 12. A comparison with the fixation probability of GL-theory is shown in Figure 9 and some of the following plots as brown curve. GL-theory captures two salient features of the stationary adaptive process: the behavior of strongly beneficial driver mutations and the drastic reduction of the total substitution rate caused by interference. However, the full spectrum of $G(\sigma)$ requires taking into account positive and negative interference. Furthermore, mutation-based models with rate and effect of beneficial mutations as input parameters cannot predict the degree of adaptation, as discussed in section 3.3.

Mutation and substitution rates

In Figure 10 we plot the rate of mutations, $U(\sigma)$ and their substitutions, $V(\sigma)$ compared to the expectation without interference from linkage. The rate of beneficial mutations is enhanced, while the rate of deleterious mutations is decreased in comparison to single site theory. This increase reflects the fact that the population under linkage is less adapted to its environment and many sites in the genome are not fixed at the locally fitter allele. Hence mutations at these maladapted sites emit more beneficial mutations. The substitution rate also shows large deviations from the classical theory. In particular, GL-theory underestimates the rate of weakly beneficial mutations by neglecting hitchhiking, as emphasized in section 2.3.

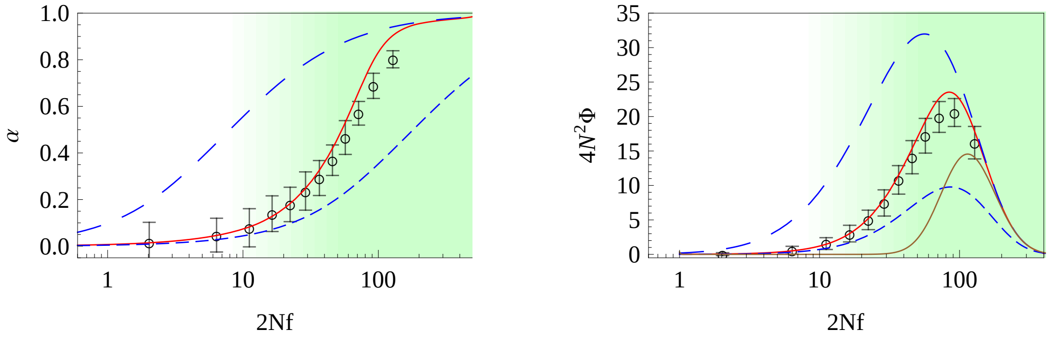


▲ **Figure 10. Mutation and substitution rates.** a) rate of mutations, b) rate of substitutions, c) theory prediction of substitutions with partitioning into passengers (gray) and drivers (green). Black circles: Simulation data, blue line: single site theory, red line: self-consistent theory, brown line: GL-theory. Parameters are the same as in Figure 9.

Degree of adaptation and fitness flux

In Figures 11, we plot the selection-dependent degree of adaptation $\alpha(f)$ and the fitness flux $\Phi(f)$ at stationarity, which are proportional to each other according to equation 20. Recall that $\alpha(f) = 0$ for randomly fixed sites and $\alpha(f) = 1$ for perfect adaptation. Simulation results are again in good agreement with our self-consistent theory, but they are not captured by single-site theory, single-site theory with globally reduced selection efficacy, or Gerrish-Lenski theory. The functions $\alpha(f)$ and $\Phi(f)$ display the emergent neutrality regime ($f < \tilde{\sigma}$) and the adaptive selection regime ($f > \tilde{\sigma}$) for genomic sites, which are again marked by color shading. Using equations 22 and 55, we can obtain approximate expressions for both regimes. Consistent with near-neutral substitution rates, sites in the emergent neutrality regime have a low degree of adaptation and fitness flux:

$$\alpha(f) = \frac{\Phi(f)}{f\,\gamma\,L\,\rho(f)} \simeq \frac{1}{1+\gamma/\mu}\,\frac{f}{2\,\tilde{\sigma}} \tag{63}$$



▲ **Figure 11.** **Degree of adaptation.** Selection dependent degree of adaptation $\alpha(f)$ and fitness flux $\Phi(f)$. Analytical model solution (red line) and simulation results (circles) show two regimes of selection: (i) Effective neutrality regime (white background): $\alpha(f)$ and $\Phi(f)$ take values similar to those of unlinked sites with reduced selection (short-dashed blue line). (ii) Adaptive regime (green): $\alpha(f)$ and $\Phi(f)$ cross over to values of unlinked sites with full selection (short-dashed blue lines). For $\Phi(f)$, the strong-selection part of the crossover is captured by the Gerrish-Lenski model, $\Phi_{\mathrm{GL}}(f)$ (brown line). Parameters are the same as in Figure 9.

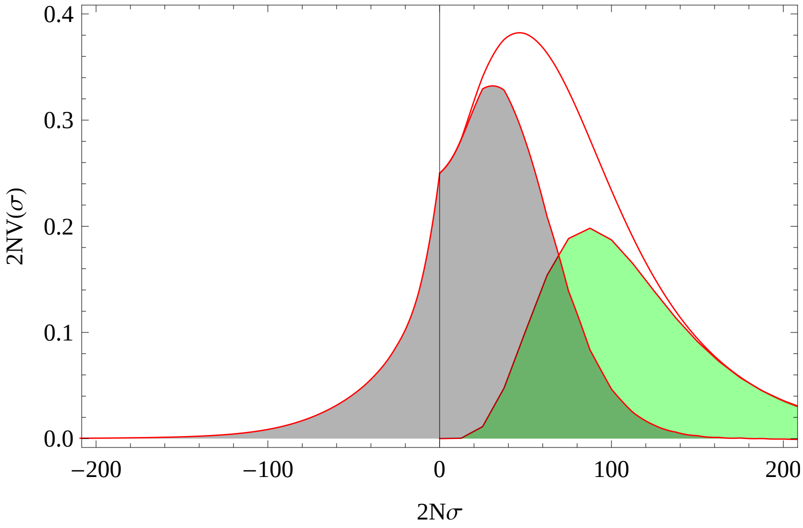
Hence, fixed sites in this regime have nearly random alleles: they cannot carry genetic information. Two processes contribute to this degradation: negative interference slows down the adaptive response to changes in selection, and hitchhiking in selective sweeps increases the rate of deleterious substitutions. By contrast, sites in the adaptive regime ($f > \tilde{\sigma}$) have a high degree of adaptation and generate most of the fitness flux. Sites under moderate selection ($f \gtrsim \tilde{\sigma}$) are still partially degraded by interference, and the negative component of fitness flux (i.e., the contribution from deleterious substitutions) is peaked in this regime (Figure 13). Strongly selected sites ($f \gg \tilde{\sigma}$) are approximately independent of interference. Hence, their degree of adaptation and fitness flux increase to values characteristic of unlinked sites,

$$\alpha(f) = \frac{\Phi(f)}{f\,\gamma\,L\,\rho(f)} \simeq \frac{f}{f + \gamma/\mu\,N} \tag{64}$$

Drivers vs. passengers

An important feature of the adaptive dynamics under linkage is the relative weight of driver and passenger mutations in selective sweeps. The separation into drivers and passenger is given by equations 44 and 52. The fixation probability is highest for strongly beneficial mutations ($\sigma \gg \sigma$), which are predominantly driver mutations.

Nevertheless, the majority of observed substitutions can be moderately adaptive or deleterious passenger mutations. Figure 12 shows this separation as different shadings. For the process displayed here, for example, about 60% of all substitutions are passengers, 20% of which are deleterious.

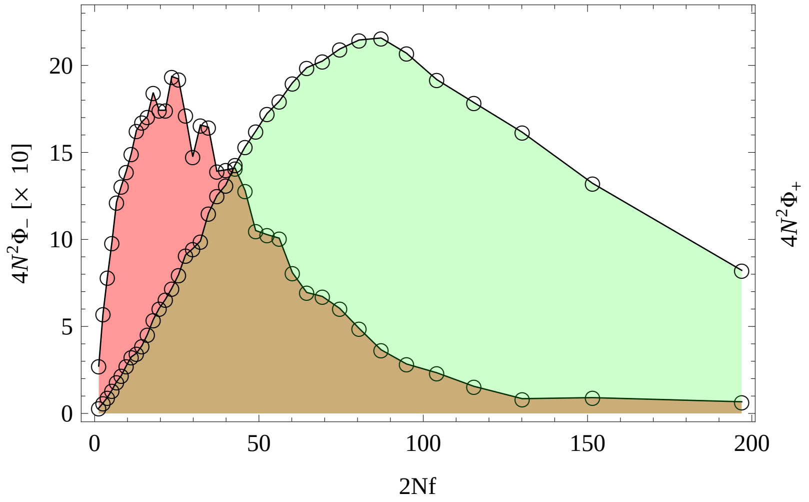


▲ **Figure 12.** **Driver and passenger substitutions.** This plot shows the separation of the total substitution rate into drivers (green) and passengers (gray). Parameters are the same as in Figure 9.

In our simulations, we can not trivially disentangle passengers from drivers, so Figure 12 shows only theory predictions. However, a related decomposition can be done for the fitness flux, which we can disentangle into a positive fitness flux, constituted by the beneficial mutations, and a negative part, caused by the fixation of deleterious mutations:

$$\Phi(f) = \Phi_{+}(f) - \Phi_{-}(f) = f(V(f) - V(-f)). \tag{65}$$

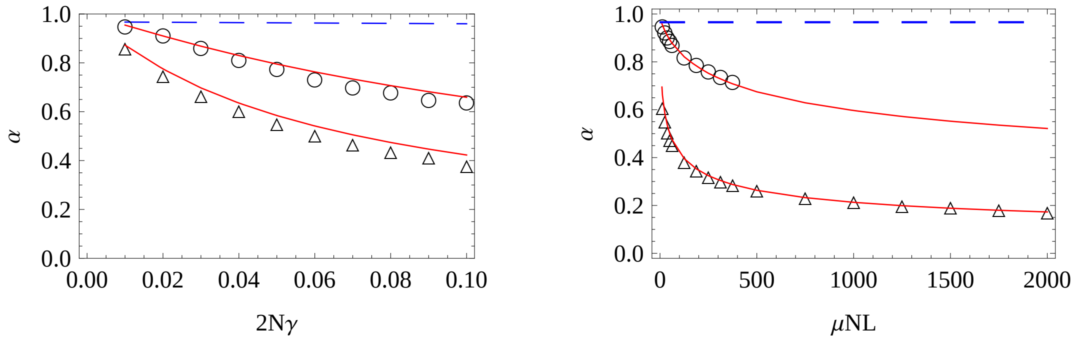
In figure 13 we show this decomposition with a ten-fold amplification of Φ_{-}. As can be seen, the two terms of the fitness flux have their main contributions coming from different parts of the spectrum of selection coefficients: While the positive flux is mainly carried by strongly beneficial mutations, the negative flux consists of weaker deleterious fixations. In total, the positive flux is always much larger than the negative one, but the decomposition reveals an interesting pattern: Remarkably, even very strongly selected genomic sites provide a significant contribution to the negative flux, reflecting ubiquitous (strongly) deleterious passenger mutations.



▲ **Figure 13. Positive and negative fitness flux.** In green: Positive fitness flux, in red: negative fitness flux. For clarity, the negative fitness flux is shown with a ten-fold amplification to give a more direct comparison. Circles: Simulation data. Parameters are the same as in Figure 9.

Genome-wide characteristics

In addition to the selection-dependent quantities discussed so far, our theory also predicts how genome-wide characteristics of the adaptive process depend on its input parameters. The adaptively evolving genome is parametrized by the mutation rate μ, the effective population size N, and by three parameters specific to our genomic model: average strength $\overline{f}$ and flip rate γ of selection coefficients, and genome length L. As an example, Figure 14 shows the dependence of the average degree of adaptation α on γ and on L, with all other parameters kept fixed (recall that according to Equation 25, this also determines the behavior of the total fitness flux, $\Phi = \alpha\,\overline{f}\,\gamma L$). The genome-wide rate of selection flips, γL, describes the rate at which new opportunities for adaptive substitutions arise at genomic sites. With increasing supply of opportunities for adaption, interference interactions become stronger. This leads to an increase in the neutrality threshold $\tilde{\sigma}$, a decrease in the degree of adaptation α, and a sub-linear increase of the fitness flux Φ. All of these effects are quantitatively reproduced by the self-consistent solution of our model. As shown in Figure 14, low values of the degree of adaptation α are observed over large regions of the evolutionary parameters γ and L. This indicates that a substantial part of the genome can be degraded to a nearly random state, implying that interference effects can compromise biological functions.

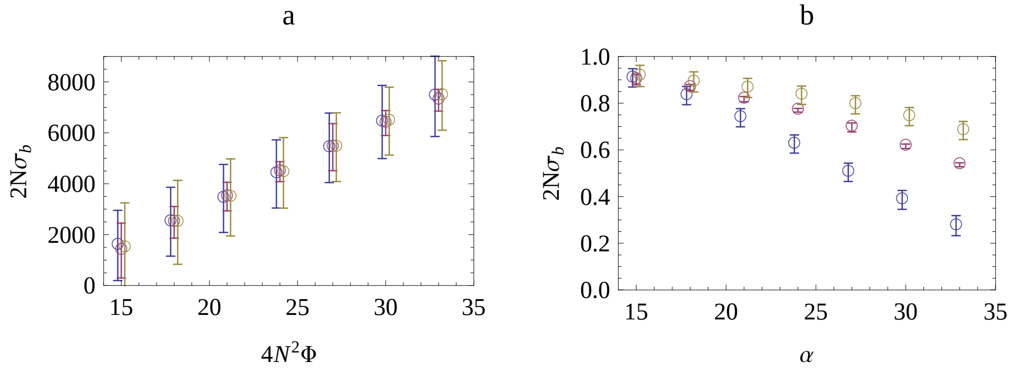


▲ **Figure 14. Degree of adaptation at stationarity.** Results from our model (red lines), simulation results (circles), and values for independent sites (dashed blue lines) of the degree of adaptation are plotted (a) against the total selection flip rate γL for two different values of the genome length, $L = 200$ (circles) and $L = 2000$ (diamonds), (b) against the total genomic mutation rate $\mu N L$ for two different values of the selection flip rate, $2N\gamma = 0.01$ (circles) and $2N\gamma = 0.1$ (diamonds). Note that for the smaller value of γ, the time to reach stationarity is very long, which limits the numerical results to smaller values of the system size. Other system parameters: $N = 4000$, $2N\mu = 0.025$, $2N\bar{f} = 50$, simulation time: $8 \cdot 10^5$ generations.

This shows that interference can strongly reduce the degree of adaptation of an evolving population, and hence its viability. This result is likely to be valid beyond the specifics of our model: in any ongoing adaptive process driven by time-dependent selection, a large reduction in the speed of adaptation due to interference is inextricably linked to a large fitness cost compared to unlinked sites.

Comparison to mutation-based models

The self-consistency of genomic state and mutations is one feature that distinguishes our model from most previous studies of adaptation under linkage [28, 77, 18, 57]. These mutation-based models constrain the distribution of selection coefficients for beneficial mutations, $u(\sigma) = (1/U_b)\,U(\sigma)$, to a fixed shape and use the total rate of beneficial mutations, U_b, and their mean effect, $\sigma_b = \int_0^\infty \sigma\, u(\sigma)\, d\sigma$, as independent input parameters. This is a suitable setup to evaluate the speed of adaptation at stationarity, because Φ depends in good approximation only on the distribution of beneficial mutations (see Figure 15a). However, mutation-based approaches of this type cannot predict genomic quantities such as the average degree of adaptation, α, which arguably is the most appropriate measure of the efficiency of the adaptive process over long periods of time. Even at stationarity, the average degree of adaptation is not uniquely determined by U_b and σ_b, but depends on all three genomic parameters $\overline{f}$, γ, and L in a nontrivial way (see Figure 15b). In our model, the rates $U(\sigma)$ of beneficial mutations (and, hence, U_b and σ_b) are dependent quantities, which must be derived from the self-consistent solution of the genome dynamics described in the main text. Changing any of the genomic parameters, say L, will change U_b and σ_b, so that these parameters are not suitable as input if we want to evaluate the dependence of the model on L. Similarly, U_b and σ_b change with time in a non-stationary adaptive process.



▲ **Figure 15.** **Comparison with mutation based models.** This plot shows the fitness flux Φ and the degree of adaptation α as a function of σ_b at *fixed* $2NU_b = 30$. The data points have been obtained by interpolating simulation results with varying γ, L and μ with three different fixed values of $2N\overline{f} = 40, 50, 60$ in blue, red and yellow. Error bars are predicted by the interpolation method (see Numerical Recipes 3rd edition [61] "Kriging"). In a) circles have been shifted horizontally by $\epsilon = \pm 0.2$ to make distinction between the data points possible.

3.4 Approach to equilibrium in a static fitness landscape

A particular non-stationary adaptive process is the approach to evolutionary equilibrium in a static fitness landscape, starting from a poorly adapted initial genome state.

In contrast to the above described stationary adaptation scenario, we can analyze this process under less generality. The reason is, that the process now depends on the initial state, for which we can choose *any* configuration of fixed or polymorphic alleles. For example, we could start with a population that is perfectly adapted, with all sites being fixed at their better alleles. This state corresponds to an initial degree of adaptation $\alpha = 1$. Since the equilibrium degree of adaptation is, due to drift, smaller than 1, we expect α to decrease with this initial state. On the other side, we could start with a random population, with alleles fixed at random, and hence $\alpha = 0$, or even more extreme, with $\alpha = -1$, which corresponds to all sites being fixed at the worse of the two alleles.

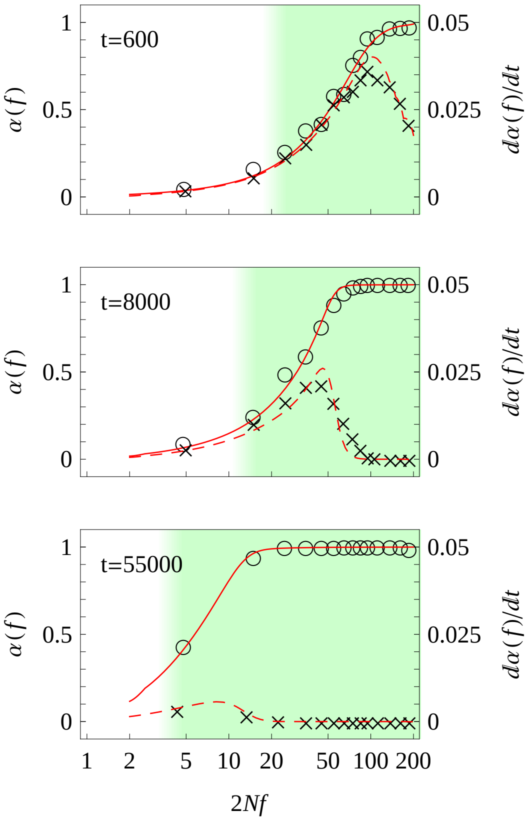
We do not expect our theoretical analysis to predict all of the approach-to-equilibrium processes equally well. In particular, if the initial state has a very poor degree of adaptation, we expect our description of effective pairwise interactions to break down, as many sites simultaneously try to adapt to their better allele and hence form combined haplotypes of multiple beneficial mutations (see [18] and [27]). We therefore restrict this part of the work to a particular initial state and its subsequent approach to equilibrium. This initial state belongs to the above analyzed family of stationary states in a fitness seascape with $\gamma > 0$. This particular choice of an initial state is relevant for the study of laboratory evolution experiments, in which a population that has undergone long-term adaption in environments that slowly changed over time ($\gamma > 0$), is put into controlled laboratory conditions that are kept fixed over time ($\gamma = 0$).

Time-dependent degree of adaptation

Figure 16 shows the selection-dependent degree of adaptation $\alpha(f)$ of this process at three consecutive times. The numerical solution has been obtained with the protocol described above (section 3.2). The self-consistent solution of our model is again in good agreement with simulation data. There is still a clear grading of genomic sites into an emergent neutrality regime and an adaptive regime, which is again marked by color shading. The neutrality threshold $\tilde{\sigma}(t)$ can be obtained numerically from equation 56 and is now a decreasing function of time. Figure 16 also shows the time-derivative of the degree of adaptation, which equals half the adaptive substitution rate per site by equations 21 and 17:

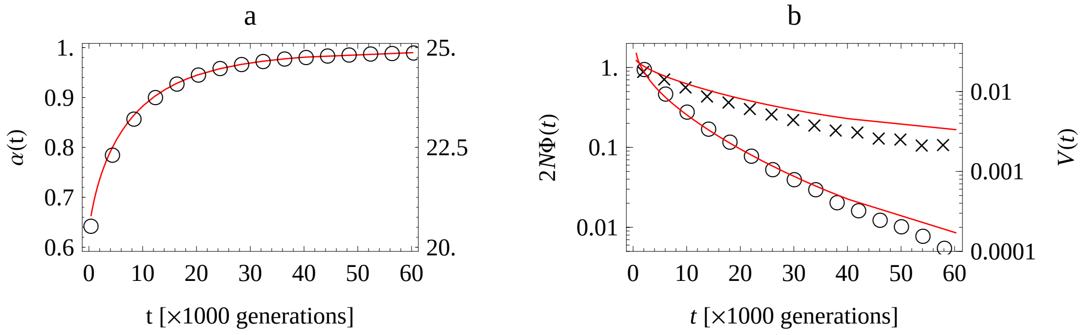
$$\frac{d\alpha(f)}{dt} = \frac{1}{2\,\rho(f)}\,(V(f) - V(-f)) \tag{66}$$

Data and model solution show that the adaptive process is non-uniform: at a given time t, adaptation is peaked at sites of effect $f \sim \tilde{\sigma}(t)$, while sites with stronger selection have already adapted at earlier times and sites with weaker selection are delayed by interference. Thus, our model predicts a non-monotonic behavior of the adaptive rate $d\alpha(f)/dt$ on time: For sites with a given selection coefficient f, this rate has a maximum at some intermediate time when $\tilde{\sigma}(t) = f$, after interference effects have weakened and before these sites have reached equilibrium. This result mirrors the maximum of the substitution rate $V(\sigma)$ at some intermediate population size for stationary adaptation. As before, a substantial fraction of substitutions are passengers in selective sweeps.



▲ **Figure 16.** **Selection regimes for approach to equilibrium.** The population evolves from a poorly adapted initial state to a high-fitness equilibrium state. The degree of adaptation (circles and solid line) and its time-derivative (which is related to the adaptive rate per site, crosses and dashed line) are shown for three consecutive times, $t = 600$, 8000 and 55.000 generations. Theory lines are obtained by numerically solving equation 21. The emergent neutrality regime ($\sigma < \tilde{\sigma}$) and the adaptive regime ($\sigma > \tilde{\sigma}$) are shown by color shading; the neutrality threshold $\tilde{\sigma}$ decreases with time. Parameters are $N = 1000$, $L = 1000$, $2N\bar{f} = 50$, $2N\mu = 0.025$. The initial state is a stationary state with $2N\gamma = 0.1$.

Figure 17a shows the evolution of the genome-averaged degree of adaptation, α, and of the mean population fitness, F, which are linearly related by equation 15. The fitness flux $\Phi = dF/dt$ and the total substitution rate V are plotted in Figure 17b. According to equation 18, these quantities are linked in a time-dependent way, $\Phi(t) = V(t)\,\overline{\sigma}_V(t) \approx V(t)\,\tilde{\sigma}(t)$. We observe that fitness increases monotonically with time. Its rate of increase Φ rapidly slows down as the system comes closer to evolutionary equilibrium, whereas the total substitution rate V shows a slower approach to equilibrium. A qualitatively similar time-dependence of fitness and substitution rate has been reported in a long-term bacterial evolution experiment by [4].

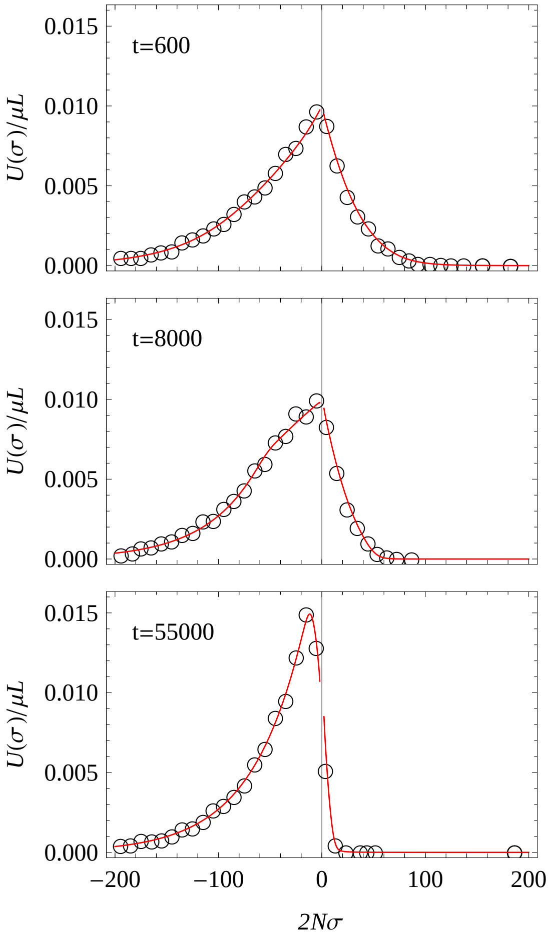


▲ **Figure 17. Degree and speed of adaptation for approach to equilibrium.** (a) The degree of adaptation evolves from a poorly adapted initial state towards equilibrium at high fitness. Simulation results are shown as circles, theory predictions as red lines. (b) The fitness flux (circles) and the substitution rate (crosses) as a function of time, together with theory (red). The substitution rate decays slower than the fitness flux, due to a decrease in $\tilde{\sigma}$. Theory predictions are obtained by numerically solving equation 21. Parameters are the same as in Figure 16.

Time-dependent mutation rates

As discussed in detail, our model involves a coupling between the genomic state and the rate of beneficial and deleterious mutations. This has consequences for an approach to equilibrium. Specifically, we expect the distribution of beneficial and deleterious mutations, $U(\sigma)$, to shift towards a higher deleterious mutation rate as the population adapts towards the equilibrium.

Figure 18 shows the distribution of fitness effects of new mutations at three different time points. Initially, the population is poorly adapted, so there are many beneficial mutations available. As the population approaches equilibrium, more mutations become deleterious, as seen in the plot. Note that this time-dependence is not to be confused with a particular type of epistasis, which generates diminishing returns in the effect of beneficial mutations [70]. Under diminishing-returns-epistasis, the fitness effect of an *individual* mutation depends on the genome state. As a consequence, this type of epistasis generates a dependence of the distribution of fitness effects on the genome state. In our model, the fitness effect of an individual mutation is independent of the genome state, and yet the distribution depends on it, simply as a consequence of the finite sequence.

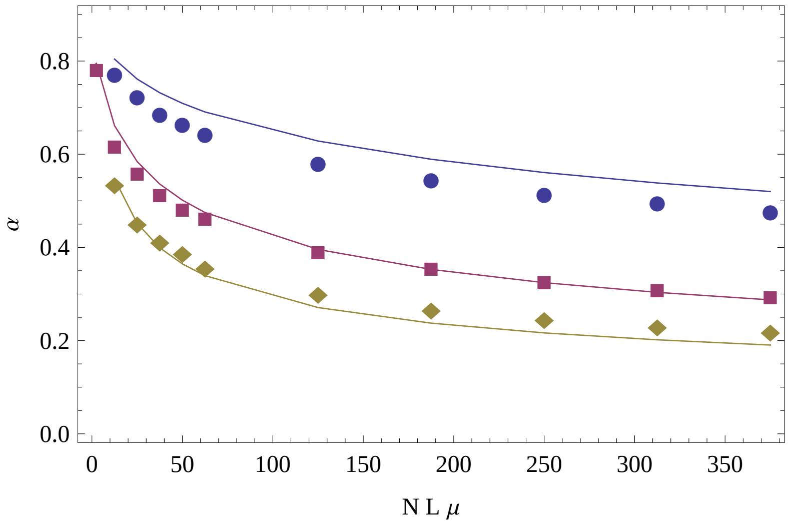


▲ **Figure 18.** This Figure shows simulation and theory results for an approach to equilibrium, starting from an initially poorly adapted state. We plot the distribution of new mutations for three different time points. Simulation results are shown in circles, theory predictions are shown in red. Parameters are the same as in Figure 16.

3.5 Non-exponential and epistatic fitness landscapes

Non-exponential distribution of selection coefficients

In addition to the exponential distribution results shown in section 3.3, we present two additional cases of distributions. Recall that the shape parameter κ (in equation 3) controls the slope of the tail of the distribution: For $\kappa = 1/2$ the tail of the distribution is a stretched exponential with broadly distributed selection coefficients. For $\kappa = 2$ we get a Gaussian tail and hence more sharply distributed selection coefficients than for the exponential case, $\kappa = 1$. In Figure 19 we show simulation results with theory predictions for the three distribution shapes $\kappa = \{0.5,\ 1,\ 2\}$. In all three cases, theory and simulation results agree very well, with the exponential case, $\kappa = 1$, exhibit ing the best agreement. Note that we expect our approach to break down for very large values of κ, i.e. very steep tails the distribution of selection coefficients. The reason is that our approximation of associating a selective sweep with a *single* driver mutation will become invalid in that case. If the distribution of selection coefficients is very narrow, beneficial mutations will have very similar selection coefficients, which means that selective sweeps are associated with *multiple* driver mutations. This regime is covered by previous studies, e.g. Desai and Fisher, 2007 [18], Rouzine et al., 2008 [65], Hallatschek [32].



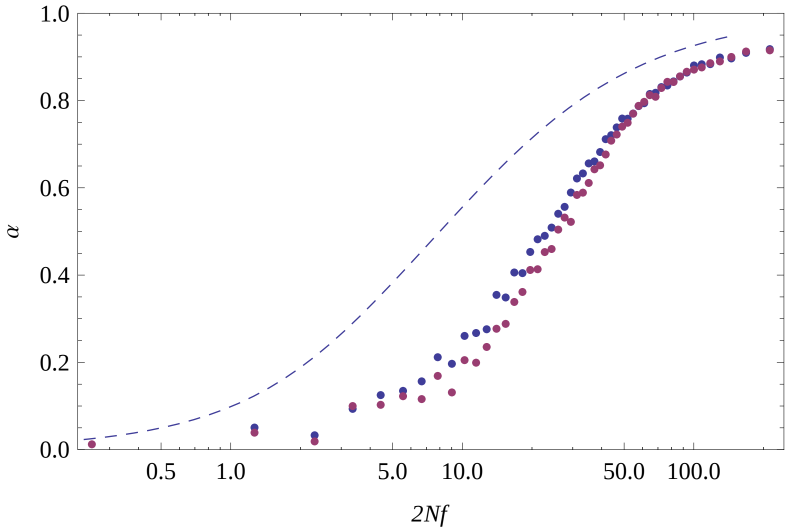
▲ **Figure 19.** **Degree of adaptation for different fitness distributions.** This plot shows simulation results for the degree of adaptation as a function of the total mutation rate. We show results for three values of $\kappa = 0.5$ (Circles), $\kappa = 1$ (Rectangles) and $\kappa = 2$ (Diamonds). Other parameters are $N = 4000$, $2N\bar{f} = 50$, $2N\gamma = 0.1$, $2N\mu = 0.025$.

Epistasis

In section 1.1, we introduced not only the additive fitness landscape studied so far, but with equation 8 also an epistatic fitness landscape. This landscape exhibits explicit interactions between genomic sites at the level of fitness, as opposed to the additive landscape, in which mutations interact only through linkage.

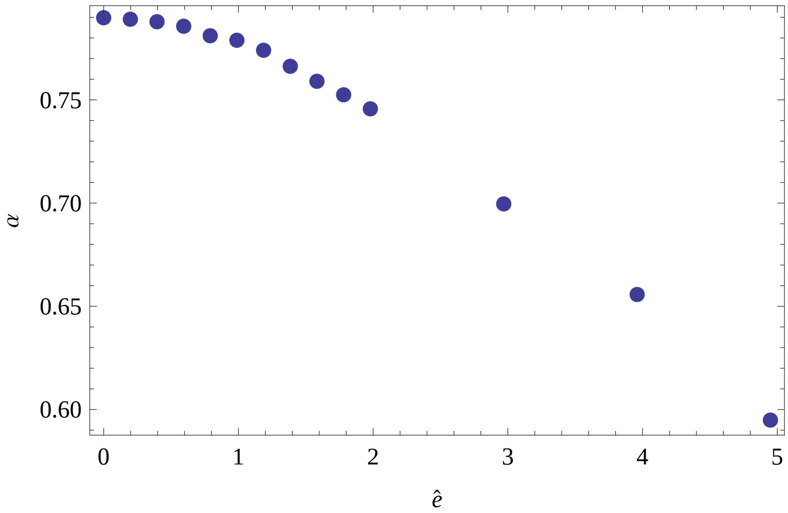
The impact of the epistatic interactions can be estimated by comparing the magnitude of the additive fitness effects with the magnitude of the epistatic effects. As can be seen from equation 8, the epistatic term scales as L^2 because of the double sum, whereas the additive term scales as L. We therefore have $e \sim \bar{f} / L$ as the typical crossover at which epistatic interactions dominate the additive fitness.

We simulate a population in the epistatic fitness landscape with the same protocol as for the additive landscape (see section 3.1). Figure 20 shows the degree of adaptation as a function of the site selection coefficient f, for both the strictly additive model (blue curve) and the epistatic model with $e = \bar{f} / L$, which is exactly the crossover value. Clearly, although the epistatic contribution to the fitness function is of the same magnitude as the additive part, the degree of adaptation does not change substantially.



▲ **Figure 20. Degree of adaptation under epistasis.** This plot shows the degree of adaptation as a function of the local selection coefficient f_j. The blue curve is the result from a simulation with no epistasis, $e = 0$, while the red curve was simulated with the epistatic fitness component being equal to the additive component: $e = \overline{f}/(L-1)$. Other parameters are $N = 1000$, $L = 1000$, $2N\gamma = 0.1$, $2N\mu = 0.025$, $2N\overline{f} = 50$.

Figure 21 shows the total degree of adaptation for varying values of the scaled epistasis parameter $\hat{e} = e(L-1)/\overline{f}$. The predicted crossover is given by $\hat{e} = 1$. In summary, epistasis does not quantitatively change our model and results, as long as the epistatic interactions are of the same magnitude as the additive component of the fitness.



▲ **Figure 21.** **Total degree of adaptation under epistasis.** Here we show the total degree of adaptation with the same parameters as in Figure 20, with varying values of scaled epistasis $\hat{e}(L-1)/\bar{f}$. For $\hat{e} > 1$ the epistatic component of the fitness is larger than the additive component.

3.6 Fluctuations and intermittency of the adaptive process

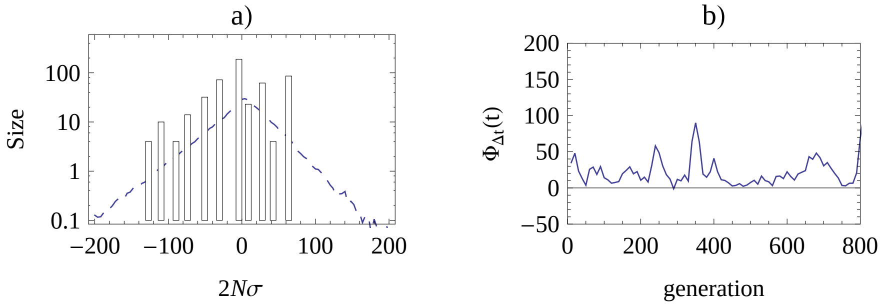
So far, we have been focusing on the mean values of fitness flux, degree of adaptation and other observables, but given genetic drift and interference as stochastic driving forces, how stochastic is the resulting adaptive substitution dynamics? This question has been addressed in several recent studies, which treat the adaptive process as a traveling fitness wave [64, 18, 65, 32]. If all mutations are assumed to have the same effect, these models are solvable. One finds a traveling wave with a deterministic bulk of stationary shape (given by a mutation-selection flux state) and a stochastic tip. The variance of this wave determines its speed (i.e., the fitness flux) by Fisher's Fundamental Theorem. Given a stationary bulk of the wave, the fitness flux has only small fluctuations around its mean value, whereas the tip of the wave is explicitly stochastic.

The adaptive process studied here shows a drastically different behavior. In our model, fitness effects at genomic sites follow a distribution $\rho(f)$ with shape parameter κ. For the case of exponential $\rho(f)$ (given by $\kappa = 1$), a snapshot of the population's fitness distribution at a given point in time is shown in Figure 22a). This distribution has large shape fluctuations throughout its bulk, not just at the tip. It shows that the adaptive process is dominated by *few* co-occurring beneficial mutations of large effect, whereas a stationary wave is maintained by many mutations of smaller effect. To measure the stochasticity of the fitness flux, we define the short time cumulative fitness flux [51]:

$$\Phi_{\Delta t}(t) = \sum_{j=1}^{L} \Delta x_j(t)\, \Delta F_j(t), \tag{67}$$

where $\Delta x_j(t)$ is the frequency change at site j between time-points t and Δt and $\Delta F_j(t)$ is the change in the population's mean fitness due to that frequency change.

As shown in Figure 22b) the fitness flux becomes *intermittent*: on small time scales (here we choose $\Delta t = 20$ generations), it has large fluctuations around its mean value. Importantly, this strong stochasticity accelerates evolution: at given rate U_b and mean effect $\sigma_b = 1 / U_b \int_0^\infty \sigma\, U(\sigma)\, d\sigma$ of beneficial mutations, our model produces a much higher mean fitness flux than the traveling-wave solution. The reason is simple: a distribution of selection coefficients generates a dynamics dominated by strong driver mutations, whose effect is substantially larger than the mean [58].

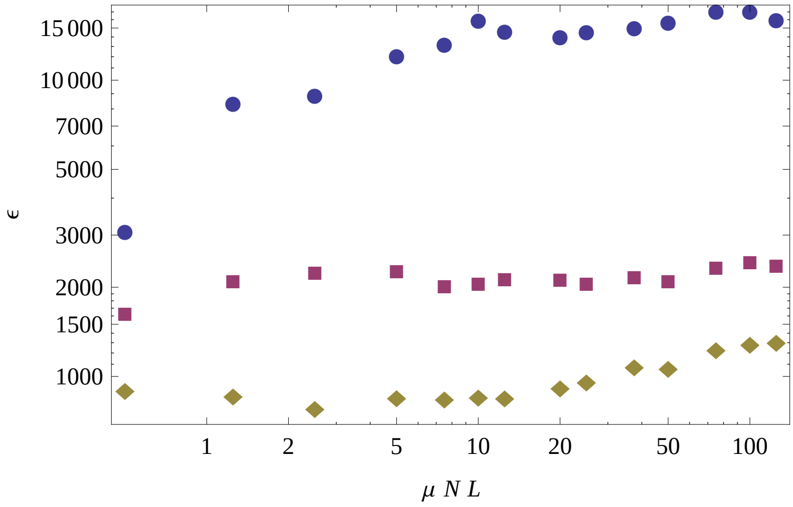


▲ **Figure 22. Stochasticity of the adaptive process.** The speed of adaptation and the shape of the distribution of fitnesses in the population are governed by large fluctuations. (a) Snapshot of the fitness distribution in the population, centered around the mean fitness. The shape of this distribution is very different from the average shape, shown as dashed line. The dynamics is governed by few fitness classes with a large number of individuals (note the logarithmic axis). (b) Time-series of the fitness flux (we use $\Delta t = 20$ generations). This flux is intermittent, i.e., the traveling fitness wave has short-term boosts in its speed. Other simulation parameters: $N = 500$\$ (c), $L = 500$, $2N\gamma = 0.1$, $2N\mu = 0.025$, $\kappa = 1$, $2N\bar{f} = 50$.

To further test the range of applicability of our model, we evaluate the stochasticity of the fitness flux for different evolutionary parameters. In Figure 23, we plot the ratio of the variance and the mean fitness flux as a function of the genome length L for fitness effect distributions $\rho(f)$ of different shapes, which have a stretched exponential ($\kappa = 1/2$), exponential ($\kappa = 1$), or Gaussian ($\kappa = 2$) tail for large values of f. As expected, we observe a decrease in stochasticity with increasing shape parameter κ. This is consistent with a crossover to fitness waves with deterministic bulk shape in the limit of a sharp distribution ($\kappa \to \infty$). However, we do not see any evidence of a cross-over to deterministic fitness waves with increasing genome length L: the stochasticity ratio

$$\epsilon = \frac{\left\langle (\Phi_{\Delta t})^2 \right\rangle}{\langle (\Phi_{\Delta t}) \rangle} \tag{68}$$

stays roughly constant and the fitness wave retains strong shape fluctuations even for the largest values of L. At the same time, our model predictions of the fixation probability $G(\sigma)$ overestimate the simulation results at large L, in particular for strongly beneficial driver mutations. This may indicate a crossover to a new mode of adaptive evolution: selective sweeps are driven cooperatively by multiple beneficial mutations, but the adaptive dynamics remains intermittent. This regime is not yet covered in the existing literature, but we expect that our interaction calculus can be extended to more complex multi-driver sweeps.



▲ **Figure 23.** **Stochasticity of the fitness flux.** The ratio of variance and mean fitness flux $\Phi_{\Delta t}(t)$ over a population's history is plotted as a function of the total mutation rate $\mu N L$ for selection shape parameters $\kappa = 1/2$ (circles), $\kappa = 1$ (squares), and $\kappa = 2$ (triangles). For a given total mutation rate, the stochasticity is highest for $\kappa = 1/2$ and decreases with increasing κ. However, the stochasticity remains approximately constant for increasing system size. Other simulation parameters: $N = 1000$, $L = 500$, $2 N \gamma = 0.1$, $2 N \mu = 0.025$, $\kappa = 1$, $2 N \bar{f} = 50$.

4 Inference in genomic data

In this chapter, we apply our developed theory to genomic data. More specifically, we want to infer the amount of *adaptive* evolution from polymorphism and divergence data of real genomes, explicitly taking into account linkage. Standard methods for genomic inference of adaptation are based on comparing the rate of species divergence with the frequencies of polymorphisms (these include the MK-test[44], Tajima's D [73] and the Fay and Wu test [22]). Most of these tests either ignore linkage completely or take into account only partial aspects, such as linkage effects on neutral polymorphisms. This is in particular worrying since many studies have shown that linkage can in fact play a crucial role, even in recombining species, such as yeast [36], fruit fly [76] and humans [14]. There is an urgent need for new methods that quantify adaptive evolution while considering linkage as an explicit feature of genome evolution, rather than a caveat (see also the recent review from Fay [23]).

Here we develop a novel inference method for genomic data that aims to quantify adaptive evolution using intra- and interpopulation sequencing data. We show that linkage generates spurious signals of positive selection, which lead to an overestimation in the rate of adaption when interpreted with many existing inference methods. Given this finding and several recent reports on pervasive adaptive evolution in the *Drosophila* genome [43, 1, 69, 23], the amount of positive selection in *Drosophila* needs to be re-examined. We apply our method to polymorphism and divergence data from *Drosophila melanogaster* and *Drosophila simulans* and find that a substantial fraction of substitutions between these species are in fact a result of linkage, rather than adaptation.

Most ingredients of the method developed in this chapter base upon concepts introduced in the previous chapters, extended with important features like recombination and allele-frequency predictions. We derive a likelihood-framework that we test with simulated data and show that it correctly infers the fraction of driver mutations in a given sequence under finite recombination. In the first four sections of this chapter, we will derive dynamics for so called *passenger* sites, that evolve under *static* selection $\gamma = 0$. These passenger sites will be linked or partially linked to *driver* sites, whose only property is that they emit *driver substitutions* with rate γ per site. This setup is a simplification of the full distribution of selection coefficients $\rho(f)$, which is necessary to make the inference scheme, developed in the last section of this chapter, numerically feasible.

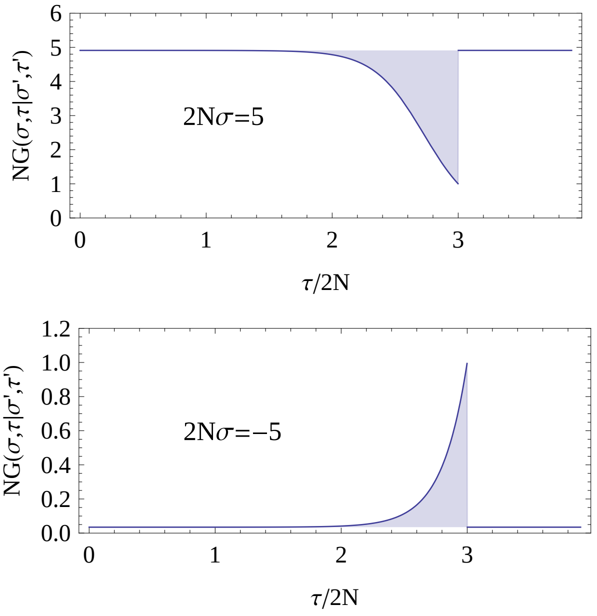
4.1 Substitution rate of passenger mutations

Single driver

We consider a single passenger site with site selection coefficient σ at time τ and a single driver mutation with selection σ' at time τ'. We assume for now that the driver is close enough to remain linked throughout its fixation. For the purpose of this chapter, we neglect effects from past sweeps ($\tau' < \tau$), but the theory can easily be extended to include the same calculation including past sweeps as in chapter 2. The fixation probability of the passenger facing a driver in the future, $G(\sigma, \tau \,|\, \sigma', \tau')$ is an integral over all possible frequencies that the passenger has reached when the driver appears

$$G(\sigma, \tau \,|\, \sigma', \tau') = \int_0^1 x\, G_0(x, \tau' - \tau; x_0, \sigma)\, dx, \tag{69}$$

with the neutral propagator of allele frequencies, $G_0(x, \tau' - \tau; x_0, \sigma)$, as derived in section 2.2. The integral in equation 69 can be solved in the diffusion approximation (equation 32): For $\tau' - \tau > \tau_{\text{fix}}(|\sigma|)$, the fixation probability $G(\sigma, \tau \,|\, \sigma', \tau')$ is equal to Kimura’s unlinked probability, $G_0(\sigma)$, while it is more neutral (see *Emergent neutrality*, section 2.4) for times $\tau' - \tau < \tau_{\text{fix}}(|\sigma|)$. Figure 24 shows $G(\sigma, \tau \,|\, \sigma', \tau')$ as a function of the passenger origination time τ, if the driver occurs at time $\tau' / 2N = 3$. The effect of the driver consists in a decrease of the beneficial fixation rate, and an increase of the deleterious fixation rate, as discussed in chapter 2. Note that the selection coefficient of the driver mutation, σ', does not enter at this point.



▲ **Figure 24. Time-dependent passenger fixation probability.** We consider the fixation probability of a passenger with selection $2N\sigma = \pm 5$, conditioned on a driver mutation appearing at time $\tau'/2N = 3$. For a beneficial passenger (upper plot), the fixation probability is reduced, while for a deleterious passenger (lower plot), it is increased by the driver mutation (lower plot). Note that the fixation probability of neutral mutations is $N\,G_0(\sigma) = 1$.

We define the fixation rates $u_\pm(\tau' - \tau) = \mu\, N\, G(\sigma,\, \tau \,|\, \sigma',\, \tau')$. For a static fitness landscape, the genome state generally evolves according to:

$$\frac{d}{d\tau}\lambda_a = -\lambda_a\, u_{-a}(\tau' - \tau) + \lambda_{-a}\, u_a(\tau' - \tau) \tag{70}$$

(see also equation 19) or in matrix notation:

$$\frac{d}{d\tau}\begin{pmatrix}\lambda_+\\ \lambda_-\end{pmatrix} = \begin{pmatrix}-u_-(\tau' - \tau) & u_+(\tau' - \tau)\\ u_-(\tau' - \tau) & -u_+(\tau' - \tau)\end{pmatrix}\begin{pmatrix}\lambda_+\\ \lambda_-\end{pmatrix} \tag{71}$$

where u_+ and u_- are now explicitly time-dependent quantities, that depend on the driver time τ'. This ordinary differential equation (ODE) is formally solved as

$$\begin{pmatrix}\lambda_+(\tau_0 + \tau_{\text{div}})\\ \lambda_-(\tau_0 + \tau_{\text{div}})\end{pmatrix} = \boldsymbol{g}(\tau_{\text{div}}) \cdot \begin{pmatrix}\lambda_+(\tau_0)\\ \lambda_-(\tau_0)\end{pmatrix} \tag{72}$$

with the divergence time τ_{div} and the propagator matrix

$$\boldsymbol{g}(\tau_{\text{div}}) = \exp\left(\int_{\tau_0}^{\tau_0+\tau_{\text{div}}} \begin{pmatrix} -u_-(\tau'-\tau) & u_+(\tau'-\tau) \\ u_-(\tau'-\tau) & -u_+(\tau'-\tau) \end{pmatrix} d\tau\right). \tag{73}$$

The matrix $\boldsymbol{g}(\tau_{\text{div}})$ defines the transition probabilities between the ancestral genome state $\lambda_\pm(\tau_0)$ and the final genome state $\lambda_\pm(\tau_0+\tau_{\text{div}})$. It absorbs the influence of genetic drift, selection and interference by the driver mutation into one propagator.

To solve the integral in equation 73, we first define the *waiting time* $\tau_w = \tau' - \tau$ and note that the driver affects the functions $u_\pm(\tau_w)$ only for waiting times $\tau_w < \tau_{\text{fix}}(|\sigma|) \ll \tau_{\text{div}}$. For all other times, in particular also for $\tau_w < 0$ (i.e. a driver that appears before the passenger), the rates $u_\pm(\tau'-\tau) \equiv \mu N G_0(\pm\sigma)$ are constant in time. We can therefore shift the boundaries of the integral:

$$\boldsymbol{g}(\tau_{\text{div}}) = \exp\left(\int_0^{\tau_{\text{div}}} \begin{pmatrix} -u_-(\tau_w) & u_+(\tau_w) \\ u_-(\tau_w) & -u_+(\tau_w) \end{pmatrix} d\tau_w\right). \tag{74}$$

where we neglected cases in which the driver appears very early $\tau' < \tau_0 + \tau_{\text{fix}}(|\sigma)|$. To solve the integral in equation 74, we now introduce the normalized waiting time distribution $P_w(\tau_w; 1/\tau_{\text{div}}) = \exp(-\tau_w/\tau_{\text{div}})/\tau_{\text{div}}$. This enables us to integrate from 0 to infinity, which does hardly change the integral, because $P_w(\tau_w) \approx 1$ for $\tau_w < \tau_{\text{fix}} \ll \tau_{\text{div}}$, and $u_\pm = u_0$ for $\tau_w > \tau_{\text{fix}}$:

$$\boldsymbol{g}(\tau_{\text{div}}) \approx \exp\left(\tau_{\text{div}} \int_0^\infty P_w(\tau_w; 1/\tau_{\text{div}}) \begin{pmatrix} -u_-(\tau_w) & u_+(\tau_w) \\ u_-(\tau_w) & -u_+(\tau_w) \end{pmatrix} d\tau_w\right). \tag{75}$$

The integrals in equation 75 have been computed before (see equation 51). Here we use the notation $u(\sigma, V) = \mu N G(\sigma, V)$ to denote the full fixation rate under linkage with driver rate $V = 1/\tau_{\text{div}}$, which was derived for stationary adaptation in chapter 2 and can be given in terms of hypergeometric functions (see equation 51). The propagator matrix then gives

$$\begin{aligned} &\boldsymbol{g}(\tau_{\text{div}}, 1/\tau_{\text{div}}) = \\ &\quad g(a, a'; \tau_{\text{div}}, V) = \exp\left(\tau_{\text{div}} \begin{pmatrix} -u(-\sigma, 1/\tau_{\text{div}}) & u(\sigma, 1/\tau_{\text{div}}) \\ u(-\sigma, 1/\tau_{\text{div}}) & -u(\sigma, 1/\tau_{\text{div}}) \end{pmatrix}\right). \end{aligned} \tag{76}$$

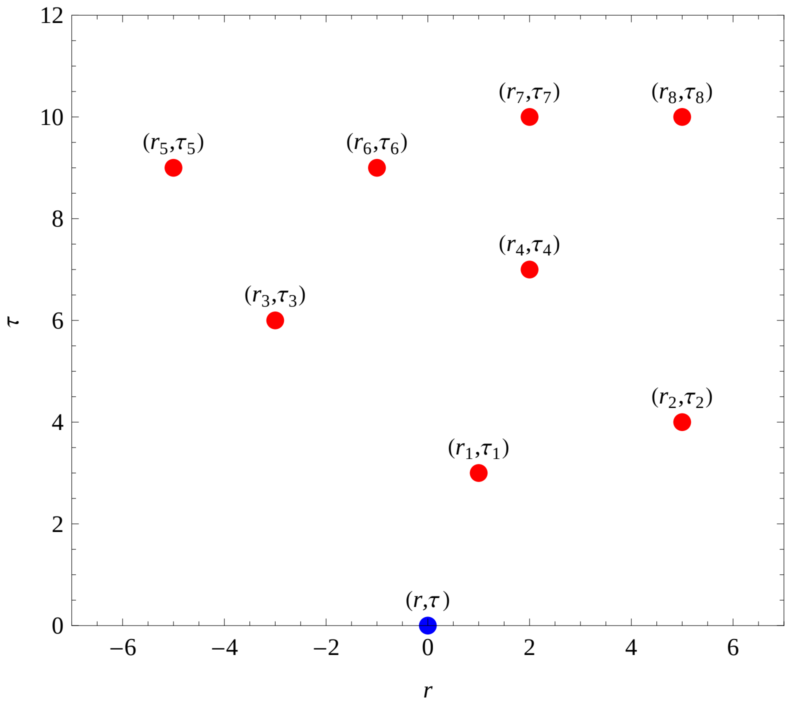
Equation 76 shows that the substitution matrix of passengers under a single driver can be computed as if every passenger faces drivers with a *rate* that equals $1/\tau_{\text{div}}$

instead of taking into account this single driver explicitly. Note that the time of the driver mutation, τ', disappeared from equation 76, because the relevant time was the time difference between the passenger and the driver, which is integrated out. This has an important consequence: since passenger mutations occur randomly and independently at every site in the genome, every passenger locus evolves *independently* according to the same ODE as above (equation 71), even if we initially consider the *same* driver at time τ'. We will use this independence for our likelihood method, derived in section 4.5.

Many drivers and recombination

In addition to the previous case, we now include partial linkage and recombination. Recombination is a process that only occurs in sexually reproducing organisms, where every sampled individual of the current generation is sampled from *two* individuals of the previous generation. The resulting sequence is a hybrid combination of the two parent sequences, with break-points (*recombination event,* or *crossing over),* separating chunks of one parent from chunks of the other parent chromosome. These break-points occur with rate ρ per site and per generation and have an important effect on the interference dynamics: they can decouple driver and passenger, and hence counteract interference effects. The full dynamics of recombination on interference effects such as hitchhiking is complicated and has been subject to detailed studies (see for example [5]).

Consider a single passenger locus at site r and time τ that is potentially interfered with by many driver mutations occurring at positions r_i and times τ_i with selection coefficients σ_i as sketched in Figure 25.



▲ **Figure 25.** **Interference by many drivers.** Schematic plot of a passenger locus at time $\tau = 0$ (blue) facing several driver mutations (red).

Not all of these driver mutations are fully linked to the passenger site, due to recombination. Here we make the approximation, that every driver can only be either fully linked, or completely unlinked. This binary approximation will prove accurate enough for our purposes and in addition provides a means to combine many drivers easily under recombination. We associate with every driver-passenger pair a binary variable, l, denoting full linkage ($l_i = 1$) or not ($l_i = 0$). We assume that l_i is a random variable that reflects the probability that recombination breaks linkage between the particular driver i and the passenger:

$$p_l(l_i = 1; |r_i - r|) = \exp(-|r_i - r|\,\rho\,\tau_{\text{fix}}(\sigma_i)). \tag{77}$$

It is derived by a simple assumption: Recombination events occur as a Poisson process with rate $\rho\,|r_i - r|$, where ρ is the recombination rate per site per generation and we assume that the recombination rate between to loci depends linearly on their distance. Equation 77 then simply describes the probability that *no* recombination event occurs between the driver and the passenger throughout its fixation time $\tau_{\text{fix}}(\sigma_i)$, which is equivalent to the probability that the two mutations remain fully linked.

How do these many linked and unlinked driver mutations interact with the single passenger? In case of the fully linked genome, we solved this problem by a topological argument: Only the *first* driver mutation after the passenger will have an effect, since it completely determines the fate of the passenger (see section 2.3). Here this argument needs to be modified, since the first driver mutation after the passenger does *not* necessarily determine the fate of the passenger, as recombination could break the linkage. Our binary approximation of linkage, using the variable l_i, provides an easy solution: we simply take the *first linked* driver mutation after the passenger to be the one that matters. That is, we define the waiting time

$$\tau_{\min} = \min_{\{i,\, l_i = 1, \tau_i > \tau\}} \tau_i. \tag{78}$$

as the minimum of all future drivers, conditioned on $l_i = 1$. We then express the fixation probability of the passenger as an integral over the unlinked propagator G_0, evaluated at time $\tau_{\min} - \tau$, analogous to equation 69:

$$G(\tau_{\min} - \tau,\, \sigma) = \int_0^1 G_0(x;\, \tau_{\min} - \tau,\, \sigma)\, x\, dx. \tag{79}$$

Following the arguments from above, we see $(\tau_{\min} - \tau)$ as a random variable of waiting times, similar to equation 75. But now, the mean value of this waiting time is not anymore given by $1/\tau_{\text{div}}$ as before, but by a smaller mean waiting time until the next linked driver occurs. This mean waiting time can be expressed as $1/V$, where V is the *rate* of linked drivers. We then yield a propagator matrix similar to equation 76:

$$\boldsymbol{g}(\tau_{\text{div}},\, V) \approx \exp\left(\tau_{\text{div}} \begin{pmatrix} -u(-\sigma,\, V) & u(\sigma,\, V) \\ u(-\sigma,\, V) & -u(\sigma,\, V) \end{pmatrix}\right), \tag{80}$$

where V is now the rate of *linked* drivers that determines the mean waiting time $\langle \tau_w \rangle = 1/V$. To compute the rate of linked drivers, we assume that drivers are emitted in the genome with rate $\gamma(r)$, which generally depends on the location. The total probability that between time τ and $\tau + d\tau$ there occurred a driver that remains linked (in Figure 25: number of red dots in a time-interval $d\tau$) is:

$$V(r)\, d\tau = \int_0^L \gamma(r')\, p_l(|r' - r|)\, dr'\, d\tau, \tag{81}$$

where we have integrated over all distances in the full sequence (from 0 to L) with the linkage probability p_l defined in equation 77. Equation 81 defines the *driver field*, which counts all nearby drivers weighted with their genetic distance. It will be discussed in more detail in section 4.4.

Total substitution rate of passengers

The substitution matrix $\boldsymbol{g}(\tau_{\mathrm{div}}, V)$ describes how the genome state propagates through time under selection, drift and partial linkage to drivers. It enables us, to compute the first ingredient for our cross-species analysis, namely the expected rate of substitutions between related species. We assume that passenger sites of the ancestral population have evolved for a long time under stationary evolution. Their equilibrium state probabilities are then given by

$$\lambda_a(\sigma, V) = \frac{u(a\,\sigma, V)}{u(\sigma, V) + u(-\sigma, V)}, \tag{82}$$

which are stationary solutions of equation 71. The *total* rate of substitutions is then given by the same equation as (see also [48]):

$$u_{\mathrm{tot}}(\sigma, V) = \\ \lambda_-(\sigma, V)\, u(\sigma, V) + \lambda_+(\sigma, V)\, u(-\sigma, V) = \frac{2\, u(\sigma, V)\, u(-\sigma, V)}{u(\sigma, V) + u(-\sigma, V)}. \tag{83}$$

We now derive an approximate expression for $u_{\mathrm{tot}}(\sigma, V)$. We assume that passenger sites are most likely fixed at their beneficial allele and set $\lambda_+(\sigma, V) = 1$, $\lambda_-(\sigma, V) = 0$. Because of deleterious mutations, these sites will be in mutation-selection balance, which is characterized by an equilibrium frequency of deleterious mutants

$$x_{\mathrm{del}} = \frac{\mu}{\sigma}. \tag{84}$$

Each linked driver that occurs will create a hitchhiking substitution at the passenger locus with probability x_{del}. After hitchhiking, the passenger will soon try to substitute back to the beneficial allele. If the time between subsequent driver events is long enough to let this happen, the rate of substitutions at the passenger locus is

$$\hat{u}_{\mathrm{tot}}(\sigma, V) = 2\, V\, \frac{\mu}{\sigma}, \tag{85}$$

where V is the rate of linked drivers. We will discuss the approximate linearity in the driver rate further in section 4.3.

4.2 Allele frequency of passenger mutations

The calculation in the last section was concerned with the substitutions dynamics, i.e. the changes of *fixed* allelic states in the population. We have introduced the driver field $V(r)$, which equals the rate of linked driver mutations at site r. Here we derive the allele frequency spectrum of a passenger site that is linked to driver mutations. We will first recapitulate results without linkage to drivers, following Mustonen and Lässig [48].

Allele frequency spectrum under selection and drift

We consider a single passenger locus with two alleles $a = \{+1, -1\}$. We denote the frequency of allele $+1$ with x. The probability distribution of the frequency x, evolving under selection, drift and mutations follows from the Fokker-Planck equation:

$$\partial_t p(x, t) = \left[\frac{1}{2N}\partial_x^2 x(1-x) - \sigma\,\partial_x x(1-x) - \mu\,\partial_x(1-2x)\right] p(x, t). \tag{86}$$

Asymptotically, for small values of μN, two linearly independent *stationary* solutions exist [48]:

$$p_a(x; \mu, \sigma) = \frac{1}{Z_a(\mu, \sigma)}[x(1-x)]^{-1+\hat{\mu}}\left(1 - e^{\hat{\sigma}(x-(1-a)/2)}\right), \text{ for } a = \pm 1 \tag{87}$$

where $\hat{\mu} = 2N\mu$ and $\hat{\sigma} = 2N\sigma$. The normalization factors are

$$Z_a(\mu, \sigma) = \frac{\Gamma(\hat{\mu})^2}{\Gamma(2\hat{\mu})}\left(1 - e^{-(1-a)\hat{\sigma}/2}\,{}_1F_1(\hat{\mu}; 2\hat{\mu}; \hat{\sigma})\right) \tag{88}$$

where ${}_1F_1(a; b; z)$ is the confluent hypergeometric function. The two solutions for $a = \pm 1$ have a simple interpretation. They are *conditioned* on the fact that the polymorphic allele always begins from one of the two fixed states (see Figure 26). For $a = 1$, the population is fixed at allele $+1$ when the polymorphism starts, while for $a = -1$ it is fixed at allele -1. Note that in case of no selection ($\sigma = 0$), the solutions

p_a take the different form

$$p_a(x;\,\mu,\,0) = \frac{1}{Z_a(\mu,\,0)}\,\frac{a-1+2\,x}{a}\,(x(1-x))^{-1+\hat{\mu}} \tag{89}$$

with

$$Z_a(\mu,\,0) = \frac{\Gamma[\hat{\mu}]^2}{\Gamma[2\,\hat{\mu}]}. \tag{90}$$

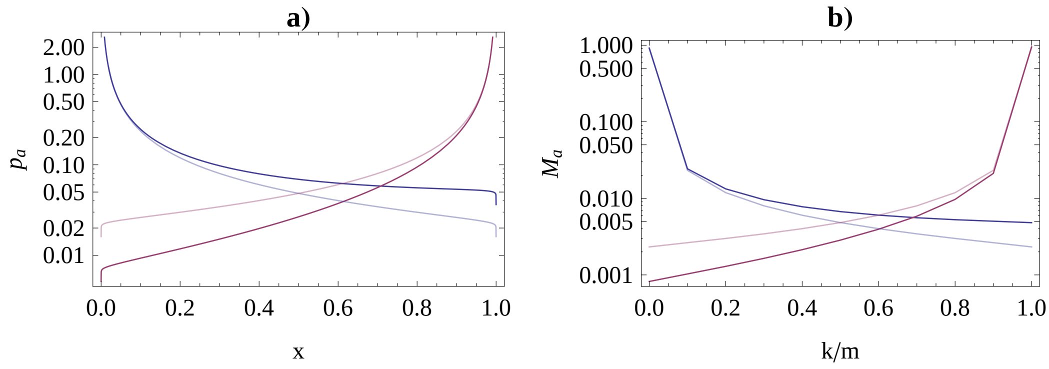
In real data, the allele frequencies are impossible to directly measure, but we can estimate x from small samples of the population. Consider a small sample of individuals, randomly drawn from the population. The probability to observe $k \leq m$ individuals with allele $+1$, can then be written as an integral over binomial moments [48]:

$$\begin{gathered} M_a(k;\,m,\,\mu,\,\sigma) = \binom{m}{k}\int_0^1 p_a(x;\,\mu,\,\sigma)\,x^k(1-x)^{m-k}\,dx = \\ \frac{1}{Z_a(\mu,\,\sigma)}\binom{m}{k}\frac{\Gamma[k+\hat{\mu}]\,\Gamma[m-k+\hat{\mu}]}{\Gamma[m+2\,\hat{\mu}]}\times \\ \left(1-e^{-\hat{\sigma}(1-a)/2}\,{}_1F_1(k+\hat{\mu};\,m+2\,\hat{\mu};\,\hat{\sigma})\right). \end{gathered} \tag{91}$$

Note again that in case of no selection, we get a different form:

$$\begin{gathered} M_a[k;\,m,\,\mu,\,0] = \\ \frac{1}{Z_a(\mu,\,0)}\binom{m}{k}\frac{\Gamma[k+\hat{\mu}]\,\Gamma[m-k+\hat{\mu}]}{a\,\Gamma[1+m+2\,\hat{\mu}]}\,(2\,k-m+a\,m+2\,a\,\hat{\mu}) \end{gathered} \tag{92}$$

Both $p_a(x;\,\mu,\,\sigma)$ and $M_a(k;\,m,\,\mu,\,\sigma)$ are shown with and without selection in figure 26. In contrast to the full probability density $p_a(x)$, the sampling probabilities produce finite probabilities for the boundary cases $k = 0$ and $k = m$.



▲ Figure 26. Allele frequency distributions. This figure shows the theory predictions of the allele frequency spectra $p_a(x)$ (a) and its moments $M_a(k;m)$ (b) Parameters are $2N\mu = 0.025$, $m = 0$, $a = -1$ (blue), $a = 1$ (red), $\nu = 2$. For comparison we also show the neutral spectrum as semi-opaque curve.

Linked drivers

We extend the above equations to include linked drivers. Their rate V enters the calculation with an approximation: We require that a given *neutral* polymorphism of frequency x needs a number of generations $x\,2\,N$ (the neutral time to reach fixation is $\sim 2\,N$, see [21]) to reach its frequency if it started at fixed state $a = -1$, and $(1-x)\,2\,N$ generations if it started at fixed state $a = 1$. The probability that no linked driver occurred during that time is a negative exponential, reflecting the waiting time in a Poisson process. Strictly, this argument holds only for neutral polymorphisms, which travel to fixation in $\sim 2\,N$ generations. In contrast, non-neutral polymorphisms travel *faster* than neutral mutations. Nevertheless, we approximate the joint effect of linked drivers and selection by applying the same argument to non-neutral polymorphisms, but expect it to break down for strongly selected mutations. We can now simply extend the stationary solution (equation 87) to include the exponential factor:

$$p_a(x;\,\mu,\,\sigma,\,V) = \frac{1}{Z_a(\mu,\,\sigma,\,V)}\left[x(1-x)\right]^{-1+\hat{\mu}}\left(1-e^{\hat{\sigma}(x-(1-a)/2)}\right)e^{-\hat{V}(-a\,x+(1+a)/2)} \tag{93}$$

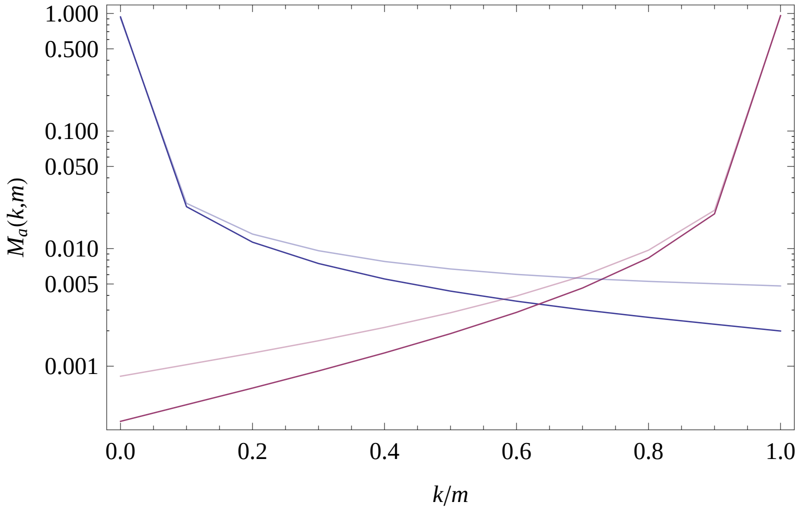
with $\hat{V} = 2\,N\,V$ and a modified normalization factor:

$$Z_a(\mu,\, \sigma,\, V) = \frac{\Gamma(\hat{\mu})^2}{\Gamma(2\,\hat{\mu})}\, e^{-\frac{1}{2}(1+a)\,\hat{V}}\Big({}_1F_1\big(\hat{\mu};\, 2\,\hat{\mu};\, a\,\hat{V}\big) - e^{\frac{1}{2}(-1+a)\,\hat{\sigma}}\; {}_1F_1\big(\hat{\mu};\, 2\,\hat{\mu};\, \hat{\sigma} + a\,\hat{V}\big)\Big) \tag{94}$$

We again use binomial sampling (see equation 91) to derive the moments:

$$\begin{aligned} M_a(k;\, m,\, \mu,\, \sigma,\, V) = {}& \frac{1}{Z_a(\mu,\, s,\, V)} \binom{m}{k} \frac{\Gamma(k+\hat{\mu})\,\Gamma(-k+m+\hat{\mu})}{\Gamma(m+2\,\hat{\mu})}\, e^{-\frac{1}{2}(1+a)\,\hat{V}} \\ & \times\Big({}_1F_1\big(k+\hat{\mu};\, m+2\,\hat{\mu};\, a\,\hat{V}\big) - e^{\frac{1}{2}(-1+a)\,s}\; {}_1F_1\big(k+\hat{\mu};\, m+2\,\hat{\mu};\, \hat{\sigma}+a\,\hat{V}\big)\Big). \end{aligned} \tag{95}$$

In Figure 27, we plot these moments with and without driving. Note that by putting $V = 0$ in equations 93, 94 and 95, we recover the known case without linkage.



▲ **Figure 27.** **Conditional polymorphism probabilities.** This figure shows the moments $M_a(k;m)$ under linkage to drivers. Parameters are $2\,N\,\mu = 0.025$, $2\,N\,V = 1$, $2\,N\,\sigma = 2$, $m = 10$. For comparison, we show the case without driving as semi-opaque curve.

Full frequency spectrum under stationarity

So far we have derived the two conditional solutions p_a and moments M_a for evolution under linkage to drivers. The full solution is a linear combination of these two conditional solutions, where the coefficients are simply the genome state probabilities λ_a:

$$p(x;\,\mu,\,\sigma,\,V) = \lambda_+(\mu,\,\sigma,\,V)\,p_+(x;\,\mu,\,\sigma,\,V) + \lambda_-(\mu,\,\sigma,\,V)\,p_-(x;\,\mu,\,\sigma,\,V), \qquad (96)$$

where not only the allele frequency spectra p_a and their moments reflect the effect of drivers, but also the stationary state probabilities λ_a, as derived in the previous section. Similarly to the full probability distributions, the discrete moments are a linear combination:

$$\begin{aligned} &M(k;\,m,\,\mu,\,\sigma,\,V) = \\ &\quad \lambda_+(\mu,\,\sigma,\,V)\,M_+(k;\,m,\,\mu,\,\sigma,\,V) + \lambda_-(\mu,\,\sigma,\,V)\,M_-(k;\,m,\,\mu,\,\sigma,\,V). \end{aligned} \qquad (97)$$

We note that one particular observable that follows from the allele frequency spectrum is the mean heterozygosity, which is here simply given by a particular moment:

$$\pi(\mu,\,\sigma,\,V) = M(1;\,2,\,\mu,\,\sigma,\,V). \qquad (98)$$

Under neutrality and without linked drivers, one can show that $\pi(\mu,\,0,\,0) \approx 2\,N\,\mu$ [21].

Outgroup-directed allele frequency spectrum

We can now use the results on the cross-species substitution rate and of the allele frequency spectrum to derive a probability distribution for cross-species polymorphism. Consider two species that diverged independently for some time τ_{div} from some common ancestor. We want to compute the probability of observing k out of m alleles with $a = 1$ from species 1 and k' out of m' alleles with $a = 1$ from species 2. We assume that both species evolve independently on this tree, but under the same evolutionary parameters, including the same driver rate V. We further assume that the common ancestor to the two diverged species is given by the *stationary* state probabilities λ_a. The probability that the two diverged species are at allelic states a' and a'' is then simply

$$p(a',\,a'') = \sum_{a=\{1,-1\}} \boldsymbol{g}_{a'a}(\tau_{\mathrm{div}},\,V)\,\boldsymbol{g}_{a''a}(\tau_{\mathrm{div}},\,V)\,\lambda_a(\mu,\,\sigma,\,V), \qquad (99)$$

where $\boldsymbol{g}_{a'a}(\tau_{\mathrm{div}},\,V)$ is the cross-species propagator under linkage to drivers (equation 80) and $\lambda_a(\mu,\,\sigma,\,V)$ is the stationary state probability (equation 82). The sum over the ancestral allele captures our lack of knowledge about the ancestral state.

Given the two allelic states a' and a'', we use the moment formulas (equation 95) to compute the likelihood of sampling k' and k'' alleles from the two species

$p(a', a'')\, M_{a'}(k'; m', \mu, \sigma, V)\, M_{a''}(k''; m'', \mu, \sigma, V)$. The full likelihood is then a sum over all unknown states a, a' and a'':

$$P(k', k''; m', m'', \tau_{\mathrm{div}}, \mu, \sigma, V) = \sum_{a,a',a''=\{1,-1\}} \boldsymbol{g}_{a'\,a}(\tau_{\mathrm{div}}, V)\, \boldsymbol{g}_{a''\,a}(\tau_{\mathrm{div}}, V) \times M_{a'}(k'; m', \mu, \sigma, V)\, M_{a''}(k''; m'', \mu, \sigma, V). \tag{100}$$

A special case is $m'' = 1$ and $m' > 1$, i.e. the situation in where one species is only known as a single reference sequence (*outgroup-species*), whereas the other species is sampled more deeply with $m' > 1$ (*ingroup-species*). The likelihood of observing k alleles in the ingroup that are *different* from the outgroup allele is then simply:

$$P(k; m, \tau_{\mathrm{div}}, \mu, \sigma, V) = \frac{1}{2}\left(P(k, 0; m, 1, \tau_{\mathrm{div}}, \mu, \sigma, V) + P(m-k, 1; m, 1, \tau_{\mathrm{div}}, \mu, \sigma, V)\right). \tag{101}$$

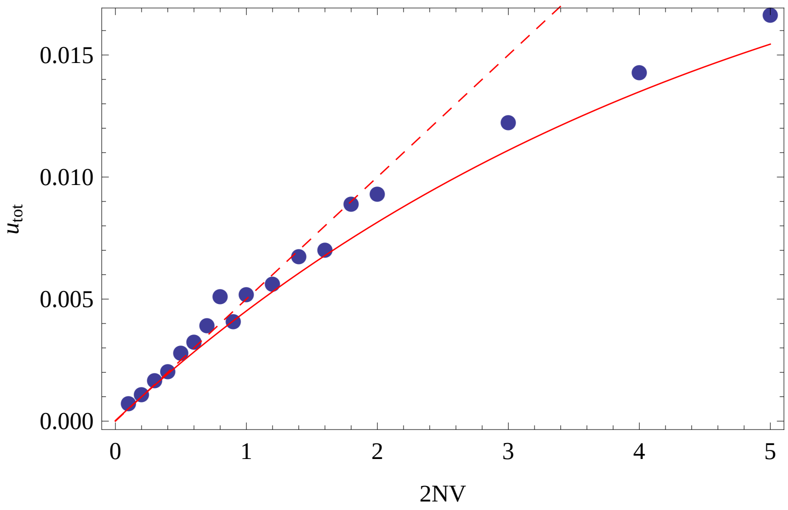
The frequency, k, of alleles that *differ* from the outgroup-species is also called the *outgroup-directed allele frequency.* Equation 101 is the key derivation of this section. It is an analytical likelihood function that can be used to *score* genomic outgroup-directed polymorphism and substitution data, as shown in section 4.5. A fit to data from synonymous sites from drosophila can be found in the Appendix (see also section 4.5).

4.3 Single-locus computer simulations

To test both allele frequency predictions and substitution dynamics under driver mutations, we introduce a simple computer simulation model. Consider a Wright-Fisher model with a single locus and two alleles under static selection and mutations. We introduce so called *Quasi-driver* events, which occur as a Poisson process with rate V. At any such event, the allele frequency at our single locus will be instantaneously “reset” to either $x = 1$ (with probability x) or $x = 0$ (with probability $1 - x$). This process mimics a linked driver locus, which emits infinitely strongly selected driver mutations with rate V. A similar model has been proposed by Gillespie [29] to study *genetic draft* of neutral diversity. In this simulation, we can then measure the substitution rate and the polymorphism spectrum. In the following, we show several observables and compare theory prediction with simulation results.

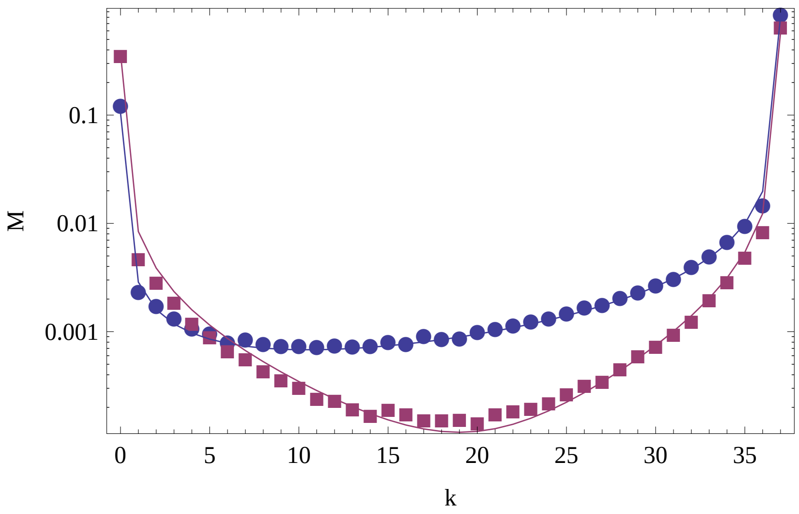
Single species

We first study the above described simulation scheme for a single population in stationarity. In this case, observables are derived by averaging over many samples taken every few generations. The total substitution rate under drivers is shown in Figure 28. The solid line shows the prediction by equation 83. The linear approximation, given by equation 85 is very accurate for low and moderate driver rates. This linearity of the total substitution rate allows a simple interpretation: Every driver event creates a passenger substitution with probability μ/σ. The higher the driver rate, the more passenger substitutions we will observe. This simple linearity breaks down if driver events become so common, that passenger sites will be fixed at their deleterious allele for substantial fraction of the time, which violates the assumptions behind the linear approximation (see derivation of equation 85).



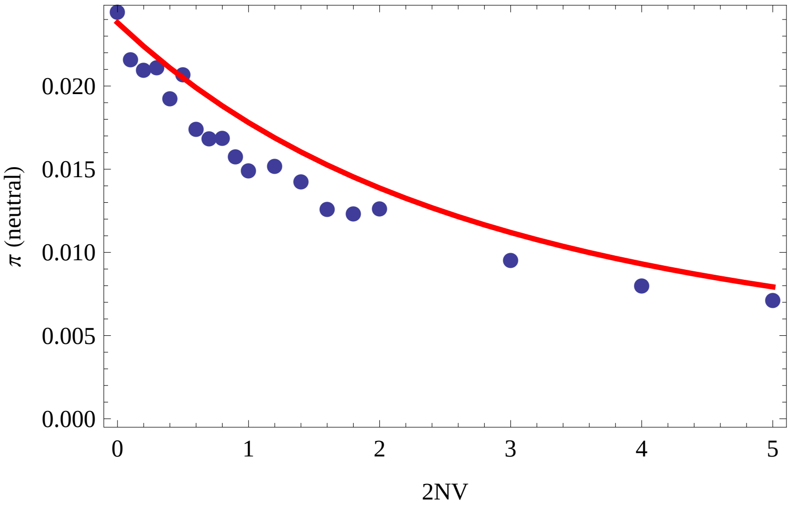
▲ **Figure 28. Passenger substitution rate with drivers.** We show simulation data in dots. The theory prediction is shown in solid red (equation 83), and the approximate formula in dashed red (equation 85). Parameters are $2N\mu = 0.025$, $2N\sigma = 10$.

Figure 29 shows the stationary allele frequency spectrum in a population with and without linkage to drivers. Linked drivers remove a substantial fraction of the polymorphisms as can be seen by comparing the red and blue curve. At the same time, they largely increase the probability that the population is fixed at the less fit state (left side of the spectrum). Note that the analytical prediction of the frequency spectrum (equation 97) agrees surprisingly well with the simulations, given the heuristic approximations of the driver-extension (equation 93).



▲ **Figure 29.** **Allele frequency spectrum.** Full allele frequency spectrum under selection, drift and linkage, simulated in a Wright-Fisher model. Parameters are $2N\mu = 0.025$, $2N\sigma = 2$, $m = 37$ and $2NV = 5$ (red), $V = 0$ (blue). Dots are simulation data, solid lines are theory.

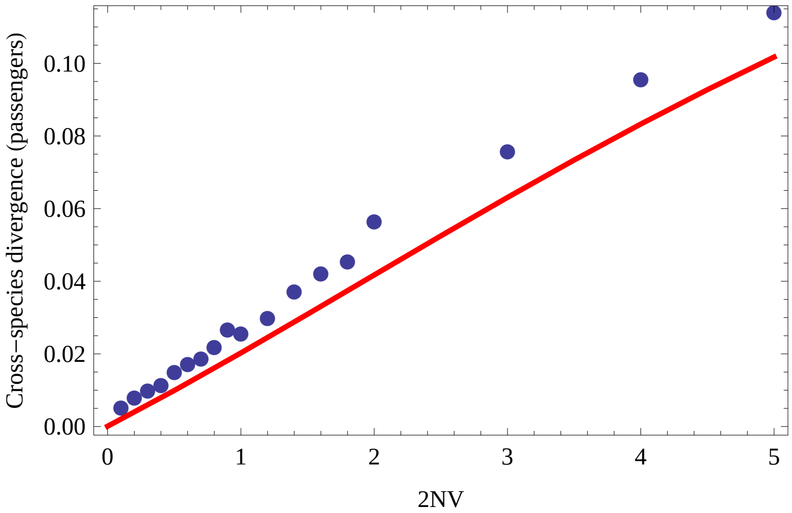
A particularly interesting summary statistics of the full polymorphism spectrum is the mean heterozygosity $\pi(\mu, \sigma, V)$, given by equation 98. Under linked driver mutations, the diversity in the population is reduced, which results in a decreasing π with V. For neutral mutations, this effect is known as genetic *draft* (see [29]). Figure 30 shows the effect of draft on the diversity of neutral passengers. Again, theory and simulations agree very well.



▲ **Figure 30. Mean heterozygosity of neutral sites.** Here we show the mean heterozygosity (equation 98) as a function of the driver rate. Parameters are $2N\mu = 0.025$ and $\sigma = 0$. Dots are simulation results, the solid red line is the prediction from theory.

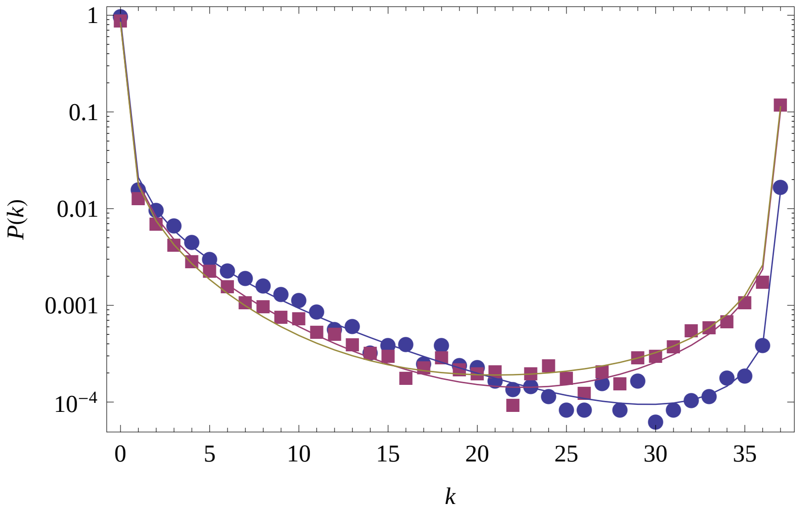
Cross-species simulations

We now apply the above described quasi-driver algorithm for a cross-species scenario. We first run the model for a single population and wait until the population has reached stationarity. This population is the ancestor-population. We then evolve both outgroup- and ingroup-species independently from that same ancestor for a time τ_{div}. To obtain statistics we run this protocol many times and average over the whole ensemble. Every simulation yields an outgroup-directed allele frequency $0 \le k \le m$, where a substitution is defined by all samples of the ingroup differing from the allele of the outgroup. For related species the substitution probability is approximately linear in the divergence time τ_{div} and in the driver rate V. Figure 31 shows the fraction of substitutions at passenger sites as a function of the rate of linked drivers, with good agreement between theory and simulations.



▲ **Figure 31.** **Cross-species divergence.** This plot shows the divergence at passenger sites between two species. Dots are simulation, solid red line is theory prediction. Parameters are $2N\mu = 0.025$, $2N\sigma_p = 10$ and $\tau_{\text{div}} = 6\times 2N$ generations.

In Figure 32 we plot the outgroup-directed allele frequency spectrum for evolution under drivers and without drivers (blue). The last value in each spectrum ($k = m$) is simply the above described cross-species divergence rate. Most interestingly, apart from the higher rate of divergence, the dominant effect of driver mutations affects the high-frequency allele frequencies of this spectrum: There are about an order of magnitude more passengers at high outgroup-directed allele frequencies under linkage to driver mutations than without drivers. Note that traditional methods to infer positive selection, such as the MK-test [44] or the method by Mustonen and Lässig [48] would infer substantial positive selection from this spectrum, as shown by the yellow curve in Figure 32, which is a maximum likelihood fit of the fluctuating selection model [48] to the simulation data.



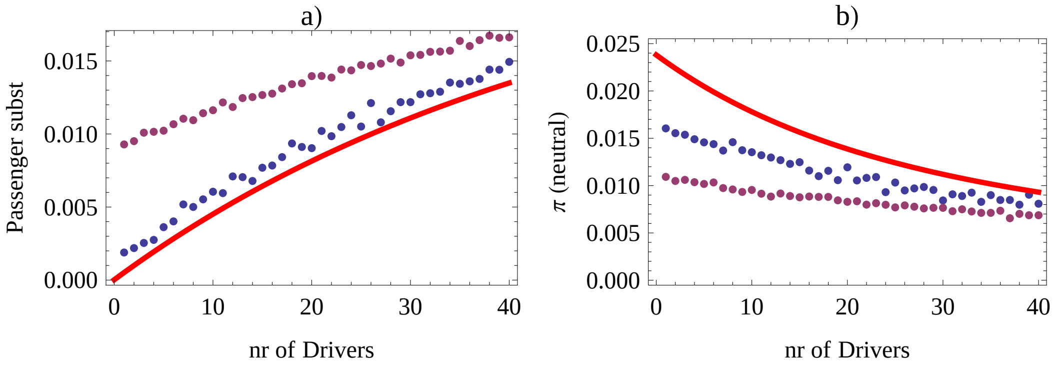
▲ **Figure 32.** **Outgroup directed polymorphism spectrum.** This plot shows simulation data from a Wright-Fisher Simulation (dots) and an evaluation of equation 101 (solid lines). Parameters are $2N\mu = 0.025$, $2NV = 2$ (red), $2N\sigma = 5$, $\tau_{\text{div}} = 6 \times 2N$, $m = 37$. For comparison we also show the case without driving, $2NV = 0$, in blue and a best fit model without drivers but with selection flips [48] in yellow. This last model contains substitutions solely as a consequence of *positive selection*, rather than hitchhiking. It is given by a fit of the simulation to the theory from Mustonen and Lässig [48], yielding a selection coefficient $2N\sigma = 9.45$ and a selection flip rate $2N\gamma = 0.0189$.

4.4 Multi-locus computer simulations

No recombination, background selection

We now test the above theory on multi-locus simulations. For that we use the following simulation setup: We define three types of sites in the sequence: i) a *passenger* site is a site with static selection coefficient σ; ii) A *driver* site has selection coefficient $\sigma_d > \sigma$ and evolves under selection flips with rate γ (see chapter 1); iii) Finally, a *neutral* site has zero selection. We run a Wright-Fisher model on this sequence and measure substitution rates and neutral diversity (based on the neutral sites only) in stationarity.

We first consider the case of zero recombination. In that case, every driver is fully linked to every passenger and neutral site and there is a global constant driver field, which simply equals the number of driver sites times their substitution rate γ. Figure 33 shows the rate of passenger substitutions and the neutral diversity as a function of the number of linked drivers with two different sequence lengths. While the theory prediction works nearly perfect for small sequence lengths, it fails to predict the large L simulation. The reason are interference effects *between* passenger sites, which are also known as *background selection* (see [11] and [12]). Background selection lowers the effective population size, which results in a higher deleterious substitution rate and in a lower neutral diversity, as seen in Figure 33 for the longer sequence length simulation data (dark red dots).



▲ **Figure 33. Multi-locus simulations without recombination.** This figure shows the rate of passenger substitutions (a) and the neutral diversity (b) as a function of linked driver sites in the sequence. Simulation results are shown in blue ($L = 200$) and red ($L = 1000$) dots. The red curve is the prediction from theory (equations 83 and 98). Other parameters are $2N\mu = 0.025$, $2N\sigma = 10$ and $2N\sigma_d = 200$.

Finite recombination

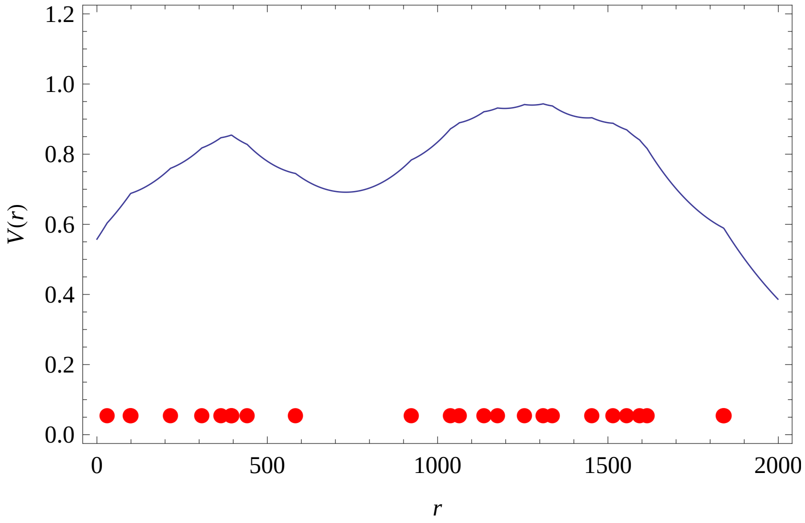
Consider now a recombining chromosome with uniform recombination rate ρ and length L. Recombination is implemented via random crossing-over events with total rate $N L \rho$. For every crossing-over we randomly pick two individuals from the population and create two new hybrid-individuals that contain all the alleles from the two individuals with a single crossing-over point that is also randomly picked. Above we already showed that recombination creates a distance dependent *driver field* (equations 77 and 81). For a site at position $1 \le r \le L$, the driver field is defined as:

$$V(r) = \sum_{r'=1}^{L} \gamma(r') \exp\left(-\frac{|r-r'|}{\xi}\right) \tag{102}$$

where $\gamma(r')$ is the rate of drivers emitted at site r'. The effect length of a driver event is given by the selection coefficient of the driver, σ_d and the recombination rate ρ:

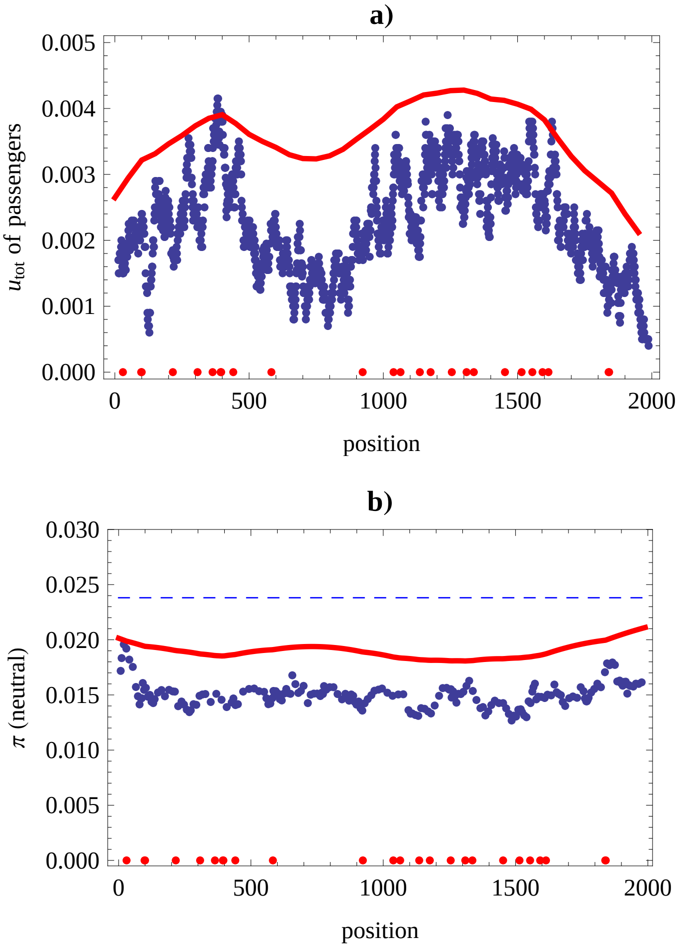
$$\xi = \frac{\tau_{\text{fix}}(\sigma_d)}{\rho} = \frac{\sigma_d}{2\,\rho} \frac{1}{\log(2\,N\,\sigma_d)}. \tag{103}$$

An example of a driver field for a random configuration of drivers is shown in Figure 34.



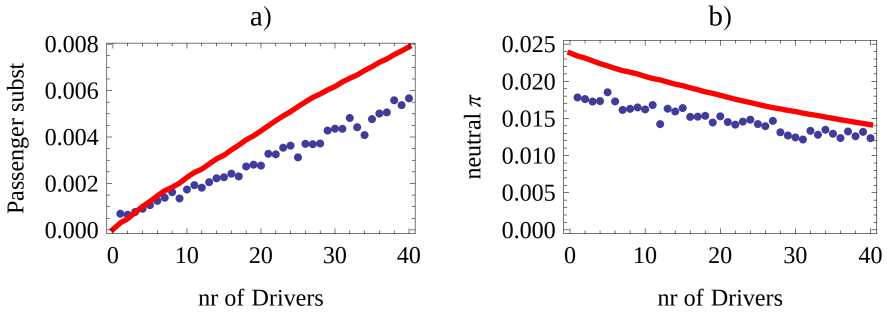
▲ **Figure 34. Example for a random driver field.** Here we show an evaluation of equation 102 with $2\,N\,\gamma = 0.1$, $2\,N\,\rho = 0.05$ and $2\,N\,\sigma = 200$. This yields a correlation length of $\xi \approx 380$ (equation 103). We distributed 25 driver sites randomly and show their positions as red dots at the bottom of the plot.

We run simulations in this driver configuration to measure the rate of passenger substitutions and neutral diversity. The result is shown in Figure 35. The effect of driver positions on the predicted and simulated passenger substitutions is clearly visible in the shown pattern. Both of the observables are qualitatively predicted correctly, however with an offset.



▲ **Figure 35. Simulations with drivers, passengers and neutral sites.** The red dots in both plots indicate the positions of the driver sites. a) Blue dots are measured substitution rates (smoothed by a moving average). The red line is the predicted substitution rate (equation 83) under the driver field created by the actual drivers in the simulation and computed by equation 102 (we took the rate of drivers that actually occurred, which is less than γ due to clonal interference). Note that the substitution rate of passengers without any drivers is quasi zero. b) Blue dots are measured neutral diversity values. Again, the red curve indicates the theory prediction by equation 98. The number of drivers in this simulation was 50, the number of neutral sites is 100, both randomly distributed along the sequence. The dashed blue curve indicates the value expected without linkage to driver sites. Other parameters are $N = 500$, $L = 2000$, $2N\mu = 0.025$, $2N\sigma_d = 100$, $2N\sigma = 10$, $2N\rho = 0.05$, $t = 1000 \times 2N$ generations.

To further explore this multi-locus model with recombination, we plot again the mean passenger substitution rate and neutral diversity as a function of the number of drivers. Comparing with the zero-recombination case (Figure 33), we first notice that the passenger substitution rate without drivers (leftmost values) shows no offset due to background selection. However, the neutral diversity still shows an offset similar to the fully linked case due to background selection.



▲ **Figure 36. Substitution and Polymorphism with recombination.** Here we show the mean passenger substitution rate (a) and the mean heterozygosity (b) as a function of the number of drivers. Parameters are $L = 1000$ and $2N\rho = 0.05$.

4.5 Likelihood-model for genomic data

We will now use the developed predictions to build a likelihood function for outgroup-directed frequency data. Consider a given dataset of length L, consisting of outgroup-directed frequencies $k(r)$ with $r = 1 \dots L$ and a set of classifications into *synonymous* and *non-synonymous* sites. We generally assume that those two classes of sites follow different dynamics. While synonymous sites are considered neutral, we will give two different “roles” to non-synonymous sites: drivers or passengers. One ingredient for this mixture model is the likelihood for stationary allele frequencies of passengers under linked drivers, as derived above (equation 101). The other ingredient is the likelihood function for independent drivers that evolve under selection flips with rate γ. This likelihood function has been derived and used for genomic inference by Mustonen and Lässig [48] and is briefly summarized in the following.

Similarly to the Markov model from section 4.1, we define a four-state Markov model for the dynamics of substitutions and selection flips. We define the four states λ_a^ϵ with two binary variables a (for the allele) and ϵ for the “preferred” allele. Similarly to the transition probability in equation 76, we can then define a 4×4 rate matrix:

$$\frac{d}{dt}\begin{pmatrix}\lambda_+^+\\ \lambda_-^+\\ \lambda_+^-\\ \lambda_-^-\end{pmatrix}=\begin{pmatrix}-u_- -\gamma & u_+ & \gamma & 0\\ u_- & -u_+-\gamma & 0 & \gamma\\ \gamma & 0 & -u_+-\gamma & u_-\\ 0 & \gamma & u_+ & -u_- -\gamma\end{pmatrix}\begin{pmatrix}\lambda_+^+\\ \lambda_-^+\\ \lambda_+^-\\ \lambda_-^-\end{pmatrix}, \tag{104}$$

where we omitted the arguments μ and σ in the substitution rates $u_\pm$. This matrix equation can be solved exactly and defines cross-species substitution rates $\boldsymbol{g}_{a'a}^{\epsilon'\epsilon}(\tau_{\text{div}}, \gamma)$ similarly to equation 99 with sums over all four states:

$$\begin{aligned}&D(k', k''; m', m'', \tau_{\text{div}}, \mu, \sigma, \gamma) =\\ &\quad\sum_{a,\epsilon,a',\epsilon',a'',\epsilon''=\{1,-1\}} \boldsymbol{g}_{a'a}^{\epsilon'\epsilon}(\tau_{\text{div}}, \gamma)\, \boldsymbol{g}_{a''a}^{\epsilon''\epsilon}(\tau_{\text{div}}, \gamma)\\ &\qquad M_{a'}(k'; m', \mu, \epsilon'\sigma)\, M_{a''}(k''; m'', \mu, \epsilon''\sigma),\end{aligned} \tag{105}$$

and for the outgroup-directed allele frequencies, similarly to equation 101:

$$\begin{aligned}&D(k; m, \tau_{\text{div}}, \mu, \sigma, \gamma) =\\ &\quad\frac{1}{2}\,(D(k, 0; m, 1, \tau_{\text{div}}, \mu, \sigma, \gamma) + D(m-k, 1; m, 1, \tau_{\text{div}}, \mu, \sigma, \gamma)).\end{aligned} \tag{106}$$

Details are found in reference [48]. We now have all ingredients for inference of passenger-driver dynamics.

We define two likelihood-models for cross-species analysis:

Mixed model I: driver-passenger

For *non-synonymous* sites in this model, we assume a mixed likelihood function: With probability η, they are driver sites with substitution rate γ and selection σ_d, while with probability $1-\eta$, they are passengers that evolve under the driver mean-field $V(r)$ created by the drivers. Since we leave the exact positioning of the driver sites undefined, we define the driver mean-field as an average over all non-synonymous sites:

$$V(r) = \sum_{r'=\{\text{nonsyn}\}}^{L} \gamma\,\eta \exp\left(-\frac{|r-r'|}{\xi}\right). \tag{107}$$

For *synonymous* sites, this model assumes neutral dynamics under linkage to the driver field $V(r)$. The full log-likelihood score of this model given a dataset of outgroup-directed allele frequencies $k(r)$ is then

$$\begin{aligned} S_1(\boldsymbol{k};\eta) = &\sum_{r=\{\text{nonsyn}\}} \log((1-\eta)\,P(k(r);\sigma_p,\ V(r)) \ +\ \eta\,D(k(r);\sigma_d,\ \gamma)) + \\ &\sum_{r=\{\text{syn}\}} \log(P(k(r);0,\ V(r))). \end{aligned} \tag{108}$$

where the passenger likelihood, $P(k;\sigma,V) = P(k;m,\tau_{\text{div}},\mu,\sigma,V)$, is given by equation 101, and the driver likelihood, $D(k;\sigma,\gamma) = D(k;m,\tau_{\text{div}},\mu,\sigma,\gamma)$, is given by a equations 106. We explicitly use the previously discussed independence between sites, which results in a factorization of the likelihood function, and the additivity of the above log-likelihood score function. Correlations between sites are implemented only via the driver field (see *Discussion*). Note that we can easily extend the driver-model (equation 105) to include effects from other drivers, i.e. their coupling to the driver field $V(r)$. But if selection on the drivers is strong as assumed here, this coupling will hardly matter.

Mixed model II: unlinked drivers

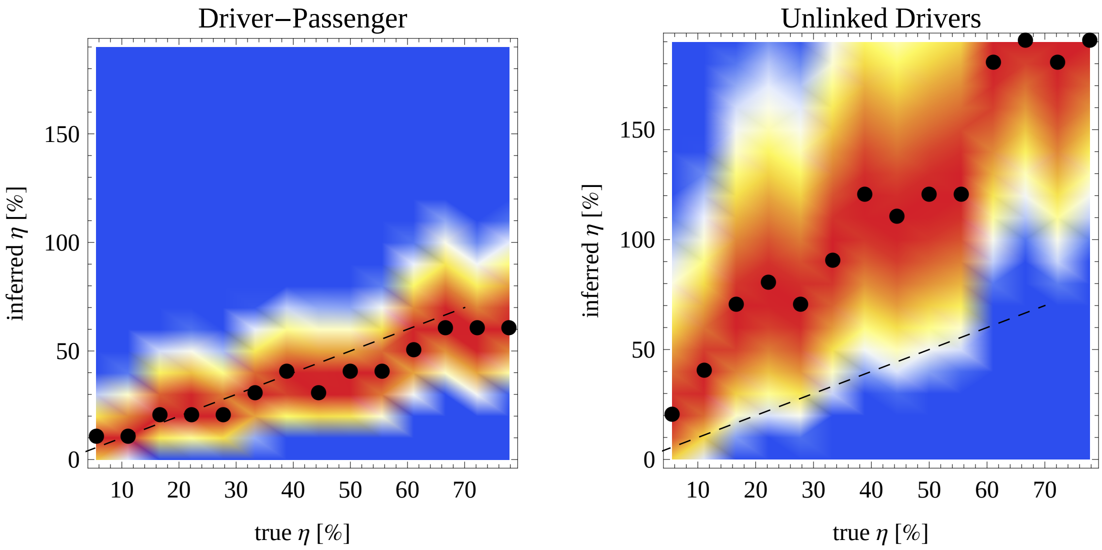
The second model is similar to the driver-passenger model, but without linkage. We therefore ignore any driver field in this model. The log-likelihood score is then

$$\begin{aligned} S_2(\boldsymbol{k};\eta) = &\sum_{r=\{\text{nonsyn}\}} \log((1-\eta)\,P(k(r);\sigma_p,\ 0) \ +\ \eta\,D(k(r);\sigma_d,\ \gamma)) + \\ &\sum_{r=\{\text{syn}\}} \log(P(k(r);0,\ 0)). \end{aligned} \tag{109}$$

In comparison to the Driver-passenger model, this model does not contain passenger substitutions, except for the very unlikely event of substitutions by drift against selection. This model therefore attributes almost all non-synonymous substitutions to driver mutations, which will overestimate the fitness flux (see below).

Application to a simulated dataset

To test the two models, we use our multi-locus simulation algorithm (see section 4.4) to create a simulated dataset with outgroup-directed polymorphism frequencies under finite recombination. We simulate the Wright-Fisher model (see section 4.4) to stationarity and define the resulting population as the common ancestor population. Subsequently we evolve both outgroup and ingroup species independently from that common ancestor for time τ_{div}. From such a simulation, we get a vector of outgroup-directed frequencies at every site in the sequence, $k(r)$ for $r = 1 \ldots \mathrm{L}$. We can then compute the log-likelihood score for both models (equations 108 and 109). In Figure 37, we show the result of such an inference, where we fixed all parameters to their true values and inferred only the parameter η, which defines the fraction of driver sites in the sequence. This one-dimensional inference has the advantage that we can plot the likelihood-surface explicitly, as shown by background-coloring. As a result, we find a slight underestimation of the driver-fraction, η, by the driver-passenger model, and a large over-estimation by the unlinked-driver model. Recall again, that we used both non-synonymous and synonymous sites for this inference method. Our results clearly indicate that finite recombination in the simulated datasets has a strong impact on the data: While the model without driver-passenger effects attributes *every* observed substitution to a driver mutation, the linkage-model attributes the correct fraction of observed substitutions to hitchhiking events.



▲ **Figure 37. Likelihood surface of the driver inference.** We simulate outgroup-directed frequency data with parameters $L = 2000$, $2N\mu = 0.025$, $\tau_{\text{div}} = 6$, $2N\rho = 0.05$, $2N\sigma = 10$ and different numbers of drivers with $\gamma = 0.1$. We then evaluate the three models (given by equations 108 and 109) for every dataset, trying different values of η. The Dots give the maximum likelihood estimates, the shading reflects the log-likelihood surface around the maximum: Red corresponds to the maximum, while blue indicates a log-score difference of 20 or higher with respect to the maximum. The dashed line is a diagonal for comparison. Note that the absolute score is comparable in model I and model II, while in model III the score is much lower, suggesting a very poor fit of the data to model III, as expected.

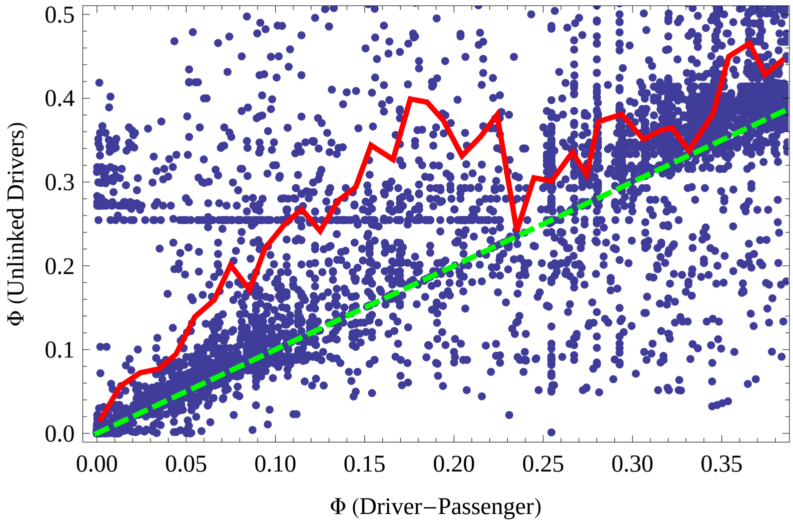
Preliminary application to *Drosophila* data

As an outlook of this work, we apply the above likelihood models to real data from the fruit-fly *Drosophila melanogaster*. For this analysis, we use re-sequencing data from the *Drosophila Population Genomics Project* [20]. This data comprises whole genome sequences for chromosomes 2L, 2R, 3L, 3R and chromosome X from 37 individuals from a wild North-American population. As an outgroup, we use the reference genome sequence of *Drosophila simulans*. We analyze coding regions from all genes in the five chromosomes, which are more than 12.000 genes. To give an overview on the dataset, we show histograms of the fraction of non-synonymous substitutions and of the heterozygosity in Figure 42 in the Appendix.

Recombination rate estimates in *Drosophila* can be obtained via a web-tool from Fiston-Lavier [26] and converted to our preferred parameter $2N\rho$ via an effective population size $N_e = 10^6$ (see [48] and the Appendix). Mutation rate and divergence time estimates are obtained from a neutral model fit to synonymous sites in highly recombining regions as $2\,N\,\mu = 0.01261$ and $\tau_{\mathrm{div}}\,/\,2\,N = 3.928$ (see Appendix). Note that the divergence time measures the branch length between the common ancestor and each of the two species.

While in our simulations, we could fix the parameters σ_p and σ_d, in real data we do not know selection coefficients of passengers and drivers a priori. Moreover, they can be different in every gene. We therefore have to find the maximum likelihood set of parameters σ_d, σ_p and η for every gene. This multi-dimensional optimization is realized via the program “amoeba” from *Numerical Recipes, 3rd edition* [61], which is a downhill-simplex algorithm by Nelder and Mead [54]. As initial values for the likelihood-maximization we choose $2\,N\,\sigma_p = 10$, $2\,N\,\sigma_d = 100$ and $\eta = 0.02$, which are motivated by inferred values from Mustonen and Lässig, 2007 [48] and by the mean divergence of non-synonymous sites (see Figure 42, Appendix).

Results are shown in Figure 38 in form of a scatter plot that shows the inferred fitness flux $\Phi = \eta\,\gamma\,\sigma_d$ for the two likelihood models, with a mean value indicated by the red line. While the optimization routine is not perfect and partly depends on the choice of the initial values, the results of this preliminary analysis are promising: The inferred fitness flux under the unlinked-driver model is much larger than for the driver-passenger model. We showed above, that the observed number of substitutions between the two species, and hence also their fitness flux, is overestimated by a model that ignores linkage. The results shown in Figure 38 are therefore consistent with our simulations results (see Figure 37) and indicate that a substantial number of substitutions in the Drosophila genome are hitchhiking as deleterious passenger substitutions rather then fixing as driver substitutions by positive selection.



▲ **Figure 38.** **Inferred driver-dynamics in Drosophila.** We show the inferred fitness flux under both the Driver-Passenger model and the unlinked-driver model. The green dashed line indicates the diagonal, the red lined the mean of the data. From this plot we excluded the highest 20% of all genes with respect to the inferred Φ under the Driver-Passenger model, since for numerical reasons we expect a very high uncertainty for high values of inferred σ_d.

Summary and discussion

In this thesis we have developed a comprehensive framework to incorporate genomic linkage into the analysis of population genetic processes and observables. A key feature of our theory was the computation of the fixation probability of mutations of arbitrary selection strengths and -sign. We find that while interference interactions in the dense-sweep regime may be complicated in their details, their net effect is simple: genomic sites with selection coefficients σ smaller than a threshold $\tilde{\sigma}$ have nearly random fixed alleles, and mutations at these sites fix with near-neutral rates. The neutrality threshold $\tilde{\sigma}$ is given by the total rate of selective sweeps, V_{drive}. Emergent neutral mutations, as well as strongly deleterious changes, fix as passengers in selective sweeps. That is, both classes of mutations are subject to interference, not genetic drift, as dominant stochastic force. Their resulting fixation probability $G(\sigma)$ depends only weakly on the effective population size N. Mutations with larger beneficial effect ($\sigma > \tilde{\sigma}$) suffer gradually weaker interference interactions. Hence, their fixation rates show a drastic increase towards the Haldane-Kimura value $G(\sigma) = 2\,\sigma$ set by genetic drift.

At a qualitative level, these results tell the story of the Hill-Robertson effect: genetic linkage reduces the efficacy of selection. Quantitatively, they demonstrate that emergent neutrality is not equivalent to a simple reduction in effective population size. The fixation rate of emergent neutral and deleterious passenger mutations can heuristically be interpreted as a linear reduction in effective population size by a factor $2\,N\,\tilde{\sigma}$, but this approximation breaks down for mutations with larger beneficial effect, as shown in Figure 9. In other words, we cannot absorb the effects of interference into a single modified strength of genetic drift. Of course, both interference and genetic drift are stochastic processes that randomize alleles of genomic sites. However, they have fundamentally different characteristics: genetic drift is a diffusion process causing *independent* changes in allele frequencies in each generation, whereas interference generates *coherent* changes over time intervals given by the inverse selection coefficient of the driver mutation.

An important concept arising from our derivation of the fixation probability is the classification of *driver*- and *passenger*- mutations. This classification arose from the necessity to close the self-consistent formalism of effective pair-interactions and it proves very useful to understand adaptive processes. We find that a substantial fraction of genomic substitutions observed in dense-sweep processes are not driver mutations, but moderately beneficial or deleterious passenger mutations fixed by hitchhiking. This fraction increases with increasing population size N or genome length L.

As shown in the last chapter of this thesis, the concept of drivers and passengers served as the key to extend the developed formalism for the fixation probability to recombining systems: Here the driver-*rate* becomes a position-dependent driver-*field*, which encapsulates genomic correlations caused by linkage: Every driver mutation affects neighboring sites up to a characteristic distance ξ, and multiple driver sites in the same region additively accumulate their effect. The dynamics of passenger mutations are then fully specified by the strength of the local driver field, with no need to specify the exact pairwise interactions between every possible pair. Moreover, we can accurately quantify how the driver field affects not only the substitution rate, but also the allele-frequency spectrum. Both of these observables are affected in a way that gave rise to a new inference scheme for genomic data, that in contrast to many previous methods explicitly takes into account genomic linkage. This method uses a mixed model, in which non-synonymous sites evolve as drivers or passengers, and synonymous sites evolve under neutral evolution with genetic draft. We have shown that this simple model is analytically tractable, such that accurate predictions for allele-frequencies and cross-species divergence could be derived. Application to real data from coding region in *Drosophila melanogaster* suggests that a substantial fraction of sites is affected from linkage. We show that traditional inference methods for positive selection (see the review from Fay [23]), that do not incorporate the possibility of hitchhiking, will *overestimate* the amount of positive selection in linked genomes.

Acknowledgments

I thank Michael Lässig for his close supervision and the inspiring years in his team. I also thank Ville Mustonen for the close collaboration and for inviting me to visit his group at the Wellcome Trust Sanger Institute, and Gergely Szöllösi for his collaboration on the publication of this work. I thank my colleagues Armita and Donate for stimulating discussions and enjoyable office-sharing, and especially, I thank my parents, my brothers and friends.

The work presented in this thesis has been supported by the *Studienstiftung des deutschen Volkes* and the *Bonn-Cologne graduate school for Physics and Astronomy*.

References

[1] Andolfatto, Peter. “Adaptive evolution of non-coding DNA in Drosophila.” Nature 437, no. 7062 (October 20, 2005): 1149–1152.

[2] Andolfatto, Peter. “Hitchhiking effects of recurrent beneficial amino acid substitutions in the Drosophila melanogaster genome.” Genome Research 17, no. 12 (December 1, 2007): 1755–1762.

[3] Bachtrog, Doris, and Isabel Gordo. “Adaptive evolution of asexual populations under Muller’s ratchet.” Evolution; international journal of organic evolution 58, no. 7 (July 1, 2004): 1403–1413.

[4] Barrick, Jeffrey E, Dong Su Yu, Sung Ho Yoon, Haeyoung Jeong, Tae Kwang Oh, Dominique Schneider, Richard E Lenski, and Jihyun F Kim. “Genome evolution and adaptation in a long-term experiment with Escherichia coli.” Nature 461, no. 7268 (October 29, 2009): 1243–1247.

[5] Barton, Nick H. “Linkage and the limits to natural selection.” Genetics 140, no. 2 (June 1, 1995): 821–841.

[6] Barton, Nick H. “Genetic hitchhiking.” Philosophical transactions of the Royal Society of London Series B, Biological sciences 355, no. 1403 (November 29, 2000): 1553–1562.

[7] Betancourt, Andrea J. “Genomewide patterns of substitution in adaptively evolving populations of the RNA bacteriophage MS2.” Genetics 181, no. 4 (April 1, 2009): 1535–1544.

[8] Birky, C W, and J B Walsh. “Effects of linkage on rates of molecular evolution.” Proceedings of the National Academy of Sciences of the United States of America 85, no. 17 (September 1, 1988): 6414–6418.

[9] www.boost.org, BOOST C++ Libraries

[10] Bush, R M, C A Bender, K Subbarao, N J Cox, and W M Fitch. “Predicting the evolution of human influenza A.” Science (New York, NY) 286, no. 5446 (December 3, 1999): 1921–1925.

[11] Charlesworth, Brian, M T Morgan, and D Charlesworth. “The effect of deleterious mutations on neutral molecular variation.” Genetics 134, no. 4 (August 1, 1993): 1289–1303.

[12] Charlesworth, Brian. “The effect of background selection against deleterious mutations on weakly selected, linked variants.” Genetical Research 63, no. 3 (June 1, 1994): 213–227.

[13] Charlesworth, Brian. “Background selection and patterns of genetic diversity in Drosophila melanogaster.” Genetical Research 68, no. 2 (October 1, 1996): 131–149.

[14] Chun, Sung, and Justin C Fay. “Evidence for Hitchhiking of Deleterious Mutations within the Human Genome.” PLoS Genetics 7, no. 8 (August 25, 2011): e1002240.

[15] Comeron, Josep M, and Martin Kreitman. “Population, evolutionary and genomic consequences of interference selection.” Genetics 161, no. 1 (May 1, 2002): 389–410.

[16] Comeron, Josep M, A Williford, and R M Kliman. “The Hill-Robertson effect: evolutionary consequences of weak selection and linkage in finite populations.” Heredity 100, no. 1 (2008): 19–31.

[17] Davis, C.S “The computer generation of multinomial random variates.” Computational statistics & data analysis 16, no. 2 (1993): 205–217.

[18] Desai, Michael M, and Daniel S Fisher. “Beneficial mutation selection balance and the effect of linkage on positive selection.” Genetics 176, no. 3 (July 1, 2007): 1759–1798.

[19] Drake, J W, Brian Charlesworth, D Charlesworth, and J F Crow. “Rates of spontaneous mutation.” Genetics 148, no. 4 (April 1, 1998): 1667–1686.

[20] www.dpgp.org, Drosophila Population Genomics project

[21] Ewens. Mathematical Population Genetics. Springer (2004)

[22] Fay, Justin C, and Chung-I Wu. “Hitchhiking under positive Darwinian selection.” Genetics 155, no. 3 (July 1, 2000): 1405–1413.

[23] Fay, Justin C. “Weighing the evidence for adaptation at the molecular level.” Trends in genetics : TIG (July 18, 2011).

[24] Felsenstein, J. “The evolutionary advantage of recombination.” Genetics 78, no. 2 (October 1, 1974): 737–756.

[25] Fisher, Ronald A. “The genetical theory of natural selection.” Clarendon Press, Oxford (1930)

[26] Fiston-Lavier, Anna-Sophie, Nadia D Singh, Mikhail Lipatov, and Dmitri A Petrov. “Drosophila melanogaster recombination rate calculator.” Gene 463, no. 1 (September 1, 2010): 18–20.

[27] Fogle, Craig A, James L Nagle, and Michael M Desai. “Clonal Interference, Multiple Mutations and Adaptation in Large Asexual Populations.” Genetics 180, no. 4 (2008): 2163–2173.

[28] Gerrish, P J, and Richard E Lenski. “The fate of competing beneficial mutations in an asexual population.” Genetica 102-103, no. 1 (1998): 127–144.

[29] Gillespie, John H. “Is the population size of a species relevant to its evolution?.” Evolution; international journal of organic evolution 55, no. 11 (November 11, 2001): 2161–2169.

[30] Haldane, JBS. “The effect of variation of fitness.” The American Naturalist 71, no. 735 (1937): 337–349.

[31] Haldane, J. “The cost of natural selection.” Journal of Genetics (1957).

[32] Hallatschek, Oskar. “From the Cover: The noisy edge of traveling waves.” Proceedings of the National Academy of Sciences of the United States of America 108, no. 5 (February 1, 2011): 1783–1787.

[33] Hartfield, M., and S P Otto. “Recombination and hitchhiking of deleterious alleles.” Evolution; international journal of organic evolution (2011).

[34] Hermisson, Joachim, and Pleuni S Pennings. “Soft sweeps: molecular population genetics of adaptation from standing genetic variation.” Genetics 169, no. 4 (April 1, 2005): 2335–2352.

[35] Hill, W, and A Robertson. “The effect of linkage on limits to artificial selection..” Genetical Research 8 (1966): 269–294.

[36] Illingworth, Christopher J R, Leopold Parts, Stephan Schiffels, Gianni Liti, and Ville Mustonen. “Quantifying selection acting on a complex trait using allele frequency time-series data..” Molecular Biology and Evolution (November 23, 2011).

[37] Kaiser, Vera B, and Brian Charlesworth. “The effects of deleterious mutations on evolution in non-recombining genomes.” Trends in genetics : TIG 25, no. 1 (2009): 9–12.

[38] Kao, K, and G Sherlock. “Molecular characterization of clonal interference during adaptive evolution in asexual populations of Saccharomyces cerevisiae.” Nature genetics 40, no. 12 (November 23, 2008): 1499–1504.

[39] Kim, Yuseob, and Wolfgang Stephan. “Joint effects of genetic hitchhiking and background selection on neutral variation.” Genetics 155, no. 3 (July 1, 2000): 1415–1427.

[40] Kim, Yuseob, and Wolfgang Stephan. “Selective sweeps in the presence of interference among partially linked loci.” Genetics 164, no. 1 (May 1, 2003): 389–398.

[41] Kimura, Motoo. “On the Probability of Fixation of Mutant Genes in a Population.” Genetics 47 (1962): 713–719.

[42] Kinnersley, Margie A, William E Holben, and Frank Rosenzweig. “E Unibus Plurum: genomic analysis of an experimentally evolved polymorphism in Escherichia coli..” PLoS Genetics 5, no. 11 (November 2009): e1000713.

[43] Macpherson, J Michael, Guy Sella, Jerel C Davis, and Dmitri A Petrov. “Genomewide spatial correspondence between nonsynonymous divergence and neutral polymorphism reveals extensive adaptation in Drosophila.” Genetics 177, no. 4 (December 1, 2007): 2083–2099.

[44] McDonald, J H, and Martin Kreitman. “Adaptive protein evolution at the Adh locus in Drosophila.” Nature 351, no. 6328 (June 20, 1991): 652–654.

[45] McVean, GAT, and Brian Charlesworth. “The effects of Hill-Robertson interference between weakly selected mutations on patterns of molecular evolution and variation.” Genetics 155, no. 2 (2000): 929–944.

[46] Muller, H. “Some Genetic Aspects of Sex.” The American Naturalist 66, no. 703 (1932): 118–138.

[47] Muller, HJ. “Our load of mutations.” American journal of human genetics 2, no. 2 (June 1, 1950): 111–176.

[48] Mustonen, Ville, and Michael Lässig. “Adaptations to fluctuating selection in Drosophila.” Proceedings of the National Academy of Sciences of the United States of America 104, no. 7 (February 13, 2007): 2277–2282.

[49] Mustonen, Ville, and Michael Lässig. “Molecular evolution under fitness fluctuations..” Physical Review Letters 100, no. 10 (March 14, 2008): 108101.

[50] Mustonen, Ville, and Michael Lässig. “From fitness landscapes to seascapes: non-equilibrium dynamics of selection and adaptation.” Trends in genetics : TIG 25, no. 3 (March 1, 2009): 111–119.

[51] Mustonen, Ville, and Michael Lässig. “Fitness flux and ubiquity of adaptive evolution..” Proceedings of the National Academy of Sciences of the United States of America 107, no. 9 (March 2, 2010): 4248–4253.

[52] Neher, Richard A, and Thomas Leitner. “Recombination rate and selection strength in HIV intra-patient evolution.” PLoS Computational Biology 6, no. 1 (2010): e1000660.

[53] Neher, Richard A, and Boris I Shraiman. “Competition between recombination and epistasis can cause a transition from allele to genotype selection.” Proceedings of the National Academy of Sciences of the United States of America 106, no. 16 (April 21, 2009): 6866–6871.

[54] Nelder, JA, and R Mead. “A simplex method for function minimization.” The computer journal (1965).

[55] Orr, H Allen. “The rate of adaptation in asexuals.” Genetics 155, no. 2 (June 1, 2000): 961–968.

[56] Otto, S P, and M C Whitlock. “The probability of fixation in populations of changing size.” Genetics 146, no. 2 (June 1, 1997): 723–733.

[57] Park, Su-Chan, and Joachim Krug. “Clonal interference in large populations.” Proceedings of the National Academy of Sciences of the United States of America 104, no. 46 (November 13, 2007): 18135–18140.

[58] Park, Su-Chan, Damien Simon, and Joachim Krug. “The Speed of Evolution in Large Asexual Populations.” Journal Of Statistical Physics 138, no. 1 (2010): 381–410.

[59] Peck, J R. “A ruby in the rubbish: beneficial mutations, deleterious mutations and the evolution of sex.” Genetics 137, no. 2 (June 1, 1994): 597–606.

[60] Perfeito, Lilia, Lisete Fernandes, Catarina Mota, and Isabel Gordo. “Adaptive mutations in bacteria: high rate and small effects.” Science (New York, NY) 317, no. 5839 (August 10, 2007): 813–815.

[61] Press, William H., Saul A. Teukolsky, William T. Vetterling, and Brian P. Flannery. “Numerical recipes, Third edition - The Art of Scientific Computing.” Cambridge University Press (2007).

[62] Rambaut, Andrew, Oliver G Pybus, Martha I Nelson, Cecile Viboud, Jeffery K Taubenberger, and Edward C Holmes. “The genomic and epidemiological dynamics of human influenza A virus.” Nature 453, no. 7195 (May 29, 2008): 615–619.

[63] Rouzine, Igor M, A Rodrigo, and J M Coffin. “Transition between stochastic evolution and deterministic evolution in the presence of selection: general theory and application to virology.” Microbiology and molecular biology reviews : MMBR 65, no. 1 (March 1, 2001): 151–185.

[64] Rouzine, Igor M, J Wakeley, and JM Coffin. “The solitary wave of asexual evolution.” Proceedings of the National Academy of Sciences of the United States of America 100, no. 2 (2003): 587–592.

[65] Rouzine, Igor M, Eric Brunet, and Claus O Wilke. “The traveling-wave approach to asexual evolution: Muller’s ratchet and speed of adaptation.” Theoretical population biology 73, no. 1 (February 1, 2008): 24–46.

[66] Rozen, Daniel E, J Arjan G M de Visser, and Philip J Gerrish. “Fitness effects of fixed beneficial mutations in microbial populations.” Current biology : CB 12, no. 12 (June 25, 2002): 1040–1045.

[67] Sattath, Shmuel, Eyal Elyashiv, Oren Kolodny, Yosef Rinott, and Guy Sella. “Pervasive adaptive protein evolution apparent in diversity patterns around amino acid substitutions in Drosophila simulans.” PLoS Genetics 7, no. 2 (2011): e1001302.

[68] Schiffels, Stephan, Gergely Szöllösi, Ville Mustonen, and Michael Lässig. “Emergent Neutrality in Adaptive Asexual Evolution..” Genetics, September 16, 2011.

[69] Sella, Guy, Dmitri A Petrov, Molly Przeworski, and Peter Andolfatto. “Pervasive natural selection in the Drosophila genome?.” PLoS Genetics 5, no. 6 (June 2009): e1000495.

[70] Silander, Olin K, Olivier Tenaillon, and Lin Chao. “Understanding the evolutionary fate of finite populations: the dynamics of mutational effects.” PLoS biology 5, no. 4 (April 1, 2007): e94.

[71] Singh, Nadia D, Peter F Arndt, and Dmitri A Petrov. “Genomic heterogeneity of background substitutional patterns in Drosophila melanogaster..” Genetics 169, no. 2 (February 2005): 709–722.

[72] Smith, J M, and J Haigh. “The hitch-hiking effect of a favourable gene.” Genetical Research 23, no. 1 (February 1, 1974): 23–35.

[73] Tajima, F. “Statistical method for testing the neutral mutation hypothesis by DNA polymorphism..” Genetics 123, no. 3 (November 1989): 585–595.

[74] de Visser, J Arjan G M, C W Zeyl, P J Gerrish, J L Blanchard, and Richard E Lenski. “Diminishing returns from mutation supply rate in asexual populations.” Science (New York, NY) 283, no. 5400 (January 15, 1999): 404–406.

[75] de Visser, J Arjan G M, and Daniel E Rozen. “Clonal interference and the periodic selection of new beneficial mutations in Escherichia coli.” Genetics 172, no. 4 (April 1, 2006): 2093–2100.

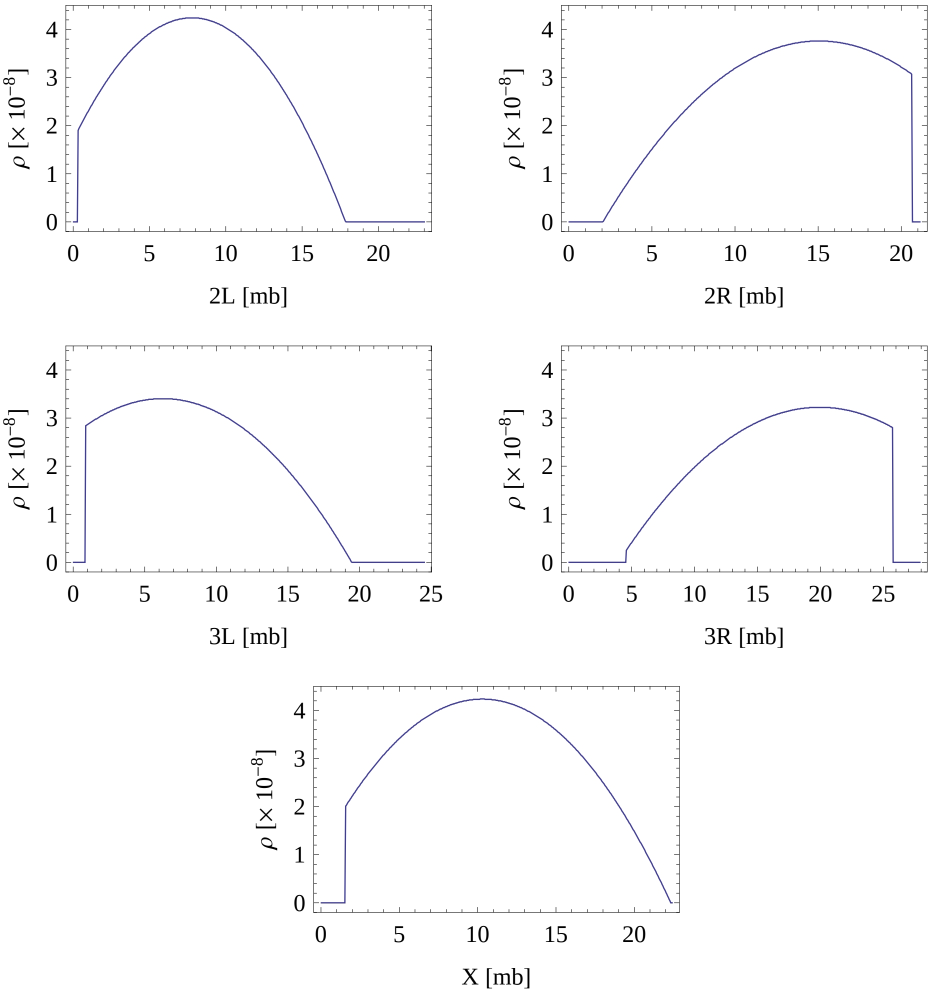
[76] Wiehe, T H, and W Stephan. “Analysis of a genetic hitchhiking model, and its application to DNA polymorphism data from Drosophila melanogaster.” Molecular Biology and Evolution 10, no. 4 (July 1, 1993): 842–854.

[77] Wilke, Claus O. “The speed of adaptation in large asexual populations.” Genetics 167, no. 4 (August 1, 2004): 2045–2053.

Appendix

Data

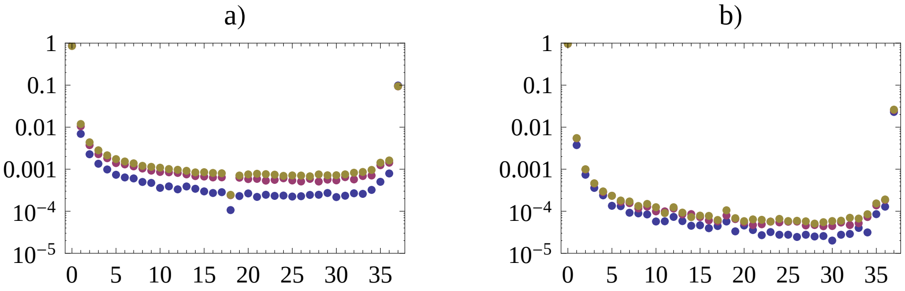
We focus exclusively on the 5 large chromosome arms of the drosophila genome. The recombination rate varies across these chromosomes, as shown in Figure 39. The recombination rate estimates used here are obtained via the *Drosophila melanogaster recombination rate calculator* from Fiston-Lavier et al. [26,71].



▲ **Figure 39. Recombination rates in Drosophila.** Data obtained from Fiston-Lavier et al. [26].

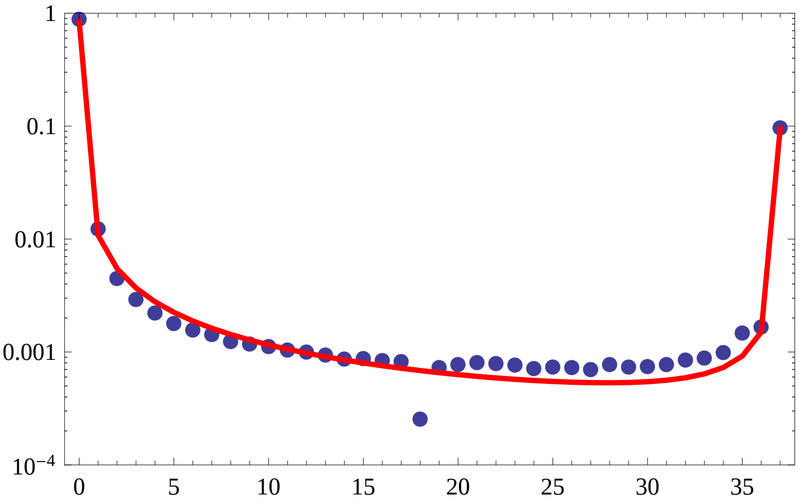
We expect linkage effects like background-selection and hitchhiking with driver mutations to be stronger in regions of low recombination rate (see section 4.4).

Indeed, the outgroup-directed allele frequency spectrum for synonymous sites show a clear grading with respect to high, medium and low recombination regions, see Figure 40.



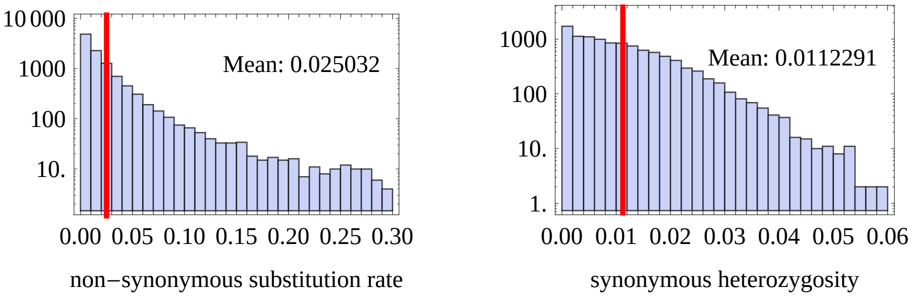
▲ **Figure 40. Synonymous and nonsymonymous polymorphism spectra.** a) synonymous polymorphisms, b) nonsynonymous polymorphisms. Blue: Low recombination ($\rho < 1\,\mathrm{cM/mb}$), Red: Medium recombination ($1 \leq \rho \leq 3$) Yellow: High recombination ($\rho > 3$). The outliers in bin 18 are numerical artifacts due to the discretization of the allele frequencies.

To obtain estimates for the mutation rate and the divergence time, we fit a neutral model to the synonymous sites data in highly recombining regions ($\rho > 3 \times 10^{-8}$). The model has been derived in section 4.2 (equation 101). For the neutral model fit, we set $\sigma = V = 0$ and fit only the parameters t and μ. Using the same maximization algorithm as for the gene inference in section 4.5, we obtain fit parameters $\tau_{\text{div}} / 2\,N = 3.928$ and $2\,N\,\mu = 0.01261$. As can be seen in Figure 41, a neutral model with these parameters fits the data very well.



▲ **Figure 41.** **Neutral model fit.** This fit has been obtained by maximizing the likelihood-function for outgroup-directed allele frequencies, equation 101, with fixed parameters $\sigma = V = 0$ and varying μ and τ_{div}. Fit Parameter values are $\tau_{\text{div}}/2N = 3.928$ and $2N\mu = 0.01261$.

To give an overview on the protein-coding data in drosophila, we show histograms of all protein-coding genes in *Drosophila* in Figure 42. Mean values are indicated by the red line. As expected, the average synonymous heterozygosity is very close to the population mutation rate $2N\mu = 0.01261$, as obtained from the allele-frequency spectrum. The mean fraction of non-synonymous substitutions is much lower than the neutral expectation $\sim\mu(2\tau_{\text{div}}) \approx 10\,\%$, indicating negative selection in the majority of genes. This is consistent with our likelihood model, where a fraction $(1-\eta)$ of non-synonymous sites evolves under static (and hence negative) selection pressure.



▲ **Figure 42.** **Divergence and diversity in protein-coding genes from *Drosophila.*** Here we show two histograms of the fraction of non-synonymous substitutions and the mean heterozygosity in protein-coding genes from Drosophila. The red lines indicate the mean values.

Erklärung

Ich versichere, dass ich die von mir vorgelegte Dissertation selbständig angefertigt, die benutzten Quellen und Hilfsmittel vollständig angegeben und die Stellen der Arbeit - einschließlich Tabellen, Karten und Abbildungen -, die anderen Werken im Wortlaut oder dem Sinn nach entnommen sind, in jedem Einzelfall als Entlehnung kenntlich gemacht habe; dass diese Dissertation noch keiner anderen Fakultät oder Universität zur Prüfung vorgelegen hat; dass sie - abgesehen von unten angegebenen Teilpublikationen - noch nicht veröffentlicht worden ist sowie, dass ich eine solche Veröffentlichung vor Abschluss des Promotionsverfahrens nicht vornehmen werde. Die Bestimmungen der Promotionsordnung sind mir bekannt. Die von mir vorgelegte Dissertation ist von Prof. Michael Lässig betreut worden.

Stephan Schiffels

Teilpublikation:
Schiffels, S., Szöllösi, G., Mustonen, V. und Lässig, M. (2011). *Emergent Neutrality in Adaptive Asexual Evolution,* Genetics.

Lebenslauf

Persönliche Information

- Name: Stephan Schiffels
- Geburtsdatum: 21. Dezember, 1980 in Aachen
- Staatsangehörigkeit: deutsch
- E-mail: stephan.schiffels@uni-koeln.de, Telefon: 0151-23503215

Studium

- Januar 2008 bis Dezember 2011: Promotionsstudium am Institut für theoretische Physik der Universität zu Köln. Finanzierung über ein Stipendium der Studienstiftung des deutschen Volkes und der *Bonn-Cologne Graduate School for Physics and Astronomy*.
- 2003 bis Dezember 2007: Diplomstudium Physik an der Universität zu Köln. Abschlussnote 1.0. Titel der Diplomarbeit “Dynamik genomischer Substitutionen”, Betreuer: Prof. Michael Lässig.
- 2001 bis 2003: Lehramtsstdium mit Fächern Musik und Physik an der Universität zu Köln und an der Hochschule für Musik, Köln.

Schulbildung

- 1991 bis 2000: Einhard-Gymnasium, Aachen. Abitur im Jahr 2000 mit Note 1.6.
- 1987 bis 1991: Katholische Grundschule Erzbergerallee, Aachen

Publikationen

- Schiffels, S., Szöllösi, G., Mustonen, V. und Lässig, M. (2011). *Emergent Neutrality in Adaptive Asexual Evolution*, Genetics.
- Illingworth, C. J. R., Parts, L., Schiffels, S., Liti, G., und Mustonen, V. (2011). *Quantifying selection acting on a complex trait using allele frequency time-series data*, Molecular Biology and Evolution

Konferenzen und Forschungsaufenthalte

- Februar/März 2011: Kavli Institute for Theoretical Physics (KITP), UCSB, Santa Barbara, CA, USA. Besuch des Workshops *Microbial and Viral Evolution.*
- Oktober 2010 bis Januar 2011: Wellcome Trust Sanger Institute, Hinxton, UK: Forschungsaufenthalt mit Ville Mustonen.
- Juni 2009: Bad Honnef, Haereus-Seminar *Physics of Biological Functions* am Physikzentrum Bad Honnef.
- November/Dezember 2008: KITP, UCSC, Santa Barbara, CA, USA: Besuch des Workshops *Population Genetics and Genomics.*
- Februar/März 2007, KITP, UCSC, Santa Barbara, CA, USA: Besuch des Workshops *Evolution of Molecular Networks.*
- September 2006, Max Planck Institut für Molekulare Genetik, Berlin: Besuch der Sommerschule *Otto Wartburg Summer School on Evolutionary Genomics.*