MATHEMATICAL MODELS OF GENE COPY NUMBER EVOLUTION

INAUGURAL-DISSERTATION

zur Erlangung des Doktorgrades
der Mathematisch-Naturwissenschaftlichen Fakultät
der Universität zu Köln

vorgelegt von

MORITZ LUKAS OTTO
aus Emden

Köln, 2023

Berichterstatter: Prof. Dr. Thomas Wiehe
Prof. Dr. Joachim Krug

Tag der letzten mündlichen Prüfung: 26. Januar 2024

Eidesstattliche Erklärung

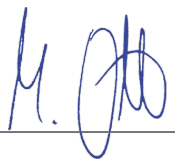
Hiermit versichere ich an Eides statt, dass ich die vorliegende Dissertation selbstständig und ohne die Benutzung anderer als der angegebenen Hilfsmittel und Literatur angefertigt habe. Alle Stellen, die wörtlich oder sinngemäß aus veröffentlichten und nicht veröffentlichten Werken dem Wortlaut oder dem Sinn nach entnommen wurden, sind als solche kenntlich gemacht. Ich versichere an Eides statt, dass diese Dissertation noch keiner anderen Fakultät oder Universität zur Prüfung vorgelegen hat; dass sie – abgesehen von unten angegebenen Teilpublikationen und eingebundenen Artikeln und Manuskripten – noch nicht veröffentlicht worden ist sowie, dass ich eine Veröffentlichung der Dissertation vor Abschluss der Promotion nicht ohne Genehmigung des Promotionsausschusses vornehmen werde. Die Bestimmungen dieser Ordnung sind mir bekannt. Darüber hinaus erkläre ich hiermit, dass ich die Ordnung zur Sicherung guter wissenschaftlicher Praxis und zum Umgang mit wissenschaftlichem Fehlverhalten der Universität zu Köln gelesen und sie bei der Durchführung der Dissertation zugrundeliegenden Arbeiten und der schriftlich verfassten Dissertation beachtet habe und verpflichte mich hiermit, die dort genannten Vorgaben bei allen wissenschaftlichen Tätigkeiten zu beachten und umzusetzen. Ich versichere, dass die eingereichte elektronische Fassung der eingereichten Druckfassung vollständig entspricht.

Teilpublikationen

Otto, Moritz, Yichen Zheng and Thomas Wiehe. *"Recombination, selection, and the evolution of tandem gene arrays"*. Genetics 221.3 (2022): iyac052.
DOI: https://doi.org/10.1093/genetics/iyac052

Otto, Moritz and Thomas Wiehe. *"The structured coalescent in the context of gene copy number variation"*. Theoretical Population Biology 154 (2023): 67-78.
DOI: https://doi.org/10.1016/j.tpb.2023.08.001

Otto, Moritz, Yichen Zheng, Paul Grablowitz and Thomas Wiehe. *"Distinguishing the roles of adaptation and demography in gene copy number changes in human populations"*. bioRxiv (2023): 2023-08.
DOI: https://doi.org/10.1101/2023.08.14.553171

Datum: 30.10.2023 Unterschrift: 

Declaration of individual contributions

Recombination, selection and the evolution of tandem gene arrays. The first and second author contributed equally to this work. The initial idea and the computer simulations were developed by Dr. Yichen Zheng. The mathematical analysis was conducted by the first author.

The structured coalescent in the context of gene copy number variation The first author conceptualized the question of the project, performed simulations, did the mathematical analysis and wrote the first draft, which was improved by the second author.

Distinguishing the roles of demography and adaptation in gene copy number changes in human populations The first and second author contributed equally to this work. Simulations were developed by Dr. Yichen Zheng. Analysis, concept and writing was conducted by both the first and second author. Paul Grablowitz was implementing an estimation procedure.

Acknowledgements

First, I thank my supervisor Thomas Wiehe for the great support throughout the last years. With the same background in mathematics it was a great inspiration to get in touch with population genetics. I am thankful for the opportunities to attend and present at international conferences from the very beginning as a part of the research group.

I would also like to thank my thesis advisors Ann-Marie Waldvogel and Joachim Krug for their insightful suggestions.

I also thank Yichen Zheng for his great ideas and the cooperative work and Johannes Wirtz for his first guidance into a new academic field.

Many thanks go to Yannick Schäfer, who made the time with scientific and especially non-scientific discusssions and meetings after work very enjoyable.

I also owe a huge thank you to my wonderful girlfriend Annika for always being supportive, attending my presentations and patiently listening to all my practice talks.

Last but certainly not least, I want to thank my family for enabling this opportunity and without whom I would have never come this far.

Abstract

One of the major interests in population genetics is to analyze the evolutionary forces, such as genetic drift, natural selection and recombination, that shape genetic variation between and within populations. Genetic variability ranges from single nucleotide mutations to structural variants such as gene duplications. In some gene families, as for instance immune or receptor genes, the variation of gene copy number is considered to play a crucial role in adapting to changing environmental conditions.

In this work, we developed a model that explores the interplay of unequal recombination and selection in the evolution of gene families. By analyzing data from the 1,000 Genomes Project, we were able to estimate selection and recombination parameters for selected candidate genes in different populations. Using analytical calculations and computer simulations, we tested whether changes in gene copy number distribution in different populations are effects of demography or a signal of adaptation. Furthermore, we introduced a new interpretation of the structured coalescent to examine genetic variation in gene families. This concept assumes gene copies to change their position within a gene array due to unequal recombination.

Zusammenfassung

Eine der Hauptaufgaben in der Populationsgenetik besteht darin, evolutionäre Kräfte wie natürliche Selektion, genetische Drift und Rekombination zu analysieren, welche die genetische Variabilität zwischen und innerhalb von Populationen gestalten. Diese genetische Vielfalt reicht von Einzelnukleotidmutationen bis hin zu strukturellen Varianten wie Genduplikationen. In einigen Genfamilien, wie beispielsweise bei Immun- oder Rezeptorgenen, spielt die Variabilität der Genkopienzahl eine entscheidende Rolle in der Anpassung an sich verändernde Umweltbedingungen.

In dieser Arbeit wurde ein Modell entwickelt, welches die Wechselwirkung von ungleicher Rekombination und Selektion in der Evolution von Genfamilien untersucht. Durch die Analyse von Daten aus dem 1.000 Genomprojekt konnten Selektions- und Rekombinationsparameter für ausgewählte Gene in verschiedenen Populationen geschätzt werden. Mit analytischen Berechnungen und Computersimulationen wurde getestet, ob Unterschiede in der Verteilung der Genkopienzahl in verschiedenen Populationen auf demografische Effekte zurückzuführen sind oder ein Signal der Anpassung darstellen. Darüber hinaus stellt diese Arbeit eine neue Interpretation der strukturierten Koaleszenztheorie vor, mit welcher die genetische Variation in Genfamilien untersucht wird. Dieses Konzept beschreibt die Positionsveränderungen von Genkopien innerhalb eines Genarrays aufgrund ungleicher Rekombination.

Contents

1 Introduction

1.1 Population genetic analysis

In Biology, the term *evolution* refers to the process of gradual and continuous change in living organisms over time. Some of these evolutionary changes happen rapidly within a few generations, such as the development of insecticide resistance (Raymond et al., 2001), others are extremely slow and occur on a timescale of hundreds of millions to billions of years, like the emergence of multicellular organisms from single-celled ones (Ridley, 2013). The field of population genetics deals with genetic variations that occur on a relatively short timescale, typically within the lifespan of populations or among closely related species. Its aim is to explain these changes based on genetics with a particular focus on the genetic information contained in DNA. Using mathematical models, population geneticists study how genetic variations are distributed and evolve within populations. Although those models rely on simplified representations of the real-world situation to be mathematically tractable, they support the understanding of inheritance and of the way genetic diversity of a population evolves through mechanisms such as genetic mutations, gene flow, recombination, natural selection and genetic drift.

The groundwork of mathematical population genetics was laid by the pioneering work of John Burdon Sanderson Haldane, Ronald Aylmer Fisher and Sewall Wright in the late 1920s (Haldane, 1927; Fisher, 1930; Wright, 1931). Wright developed the concept of *genetic drift*, emphasizing the role of random changes of gene frequencies within small populations. Fisher provided crucial insights into the role of *natural selection* in shaping populations and developed the mathematical framework to model the evolution of alleles that confer a fitness advantage and how they increase in frequency over time. Haldane made significant contributions to understanding the genetic basis of *adaptation*, described genetic linkage in mammals and analyzed the role of *recombination* in generating genetic diversity.

Back then, population genetics analyzed and quantified differences in phenotypes, i.e. the observable characteristics of an organism. With the development of molecular tools in the 1960s it was possible to analyze the genotype of individuals. One of those methods is the laboratory technique of *electrophoresis*, which can separate molecules such as DNA, RNA or proteins. This method relies on the principle that charged particles will move through a medium, usually a gel or a solution, when an electric current is applied. Based on their size, charge, and mobility the molecules move at different rates, such that larger DNA fragments travel slower (S-type) through the gel than fast, short ones (F-type).

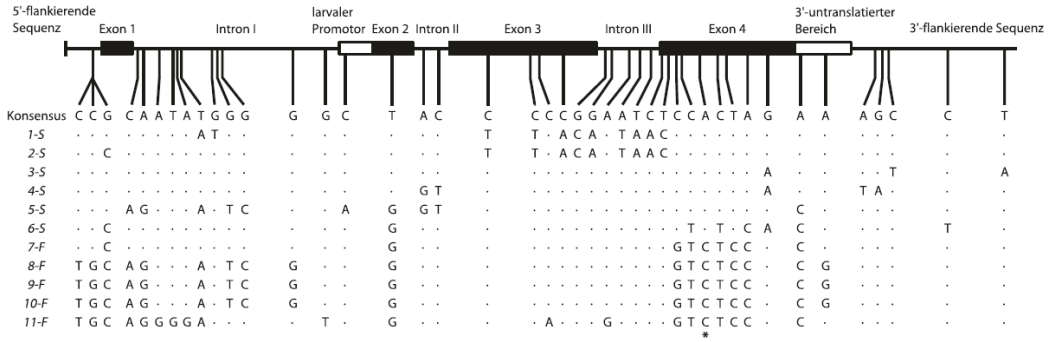


Figure 1.1: Kreitman's sequencing study on eleven sequences of the alcohol dehydrogenase gene (`Adh`) of *Drosophila melanogaster*. Figure shows the 43 polymorphic nucleotide positions. Dots indicate no deviation from the consensus sequence. The asterisk in exon 4 indicates the site where the lysine (encoded by AAG) of the S-allele is replaced by threonine (ACG) of the F-allele, causing the electrophoretic difference between the F and S alleles. Figure taken from Stephan and Hörger (2019), Figure 1.1.

The molecular techniques were improved further and with principles such as Sanger sequencing or the Maxam-Gilbert method it became possible to directly examine genetic variability at the DNA level (Sanger et al., 1977; Maxam and Gilbert, 1977). The first systematic sequencing study of multiple alleles at a locus was conducted by Kreitman (1983), even before the invention of polymerase chain reaction (PCR), which most of the modern sequencing techniques are based on. He isolated eleven different clones of the alcohol dehydrogenase gene (`Adh`) from a worldwide collection of the fruit fly *Drosophila melanogaster* and sequenced them (Figure 1.1). The nucleotide variability was surprisingly high, 43 out of the 2379 aligned loci were variable. Furthermore, 42 of the single nucleotide polymorphisms (SNPs) were *silent* or *synonymous*, i.e. they did not change the resulting amino acid. Only one SNP in exon 4, i.e. protein coding area of the gene, was non-synonymous and indeed, this nucleotide was responsible for the observed differences between the electrophoretic F and S variants.

In 1990 the Human Genome Project (HGP) was initiated, which aimed to map and sequence all genes in the human genome. It was a worldwide collaboration involving research groups across the United States, United Kingdom, France, Germany, Japan and China and it took approximately 13 years and $ 2.7 billion to complete the project[1]. Since then, new sequencing techniques have been developed which are more efficient and more affordable. In comparison, in March 2021 Euan Ashley and his *Ultra-Rapid Genome Team* from Stanford University were awarded with the GUINNESS WORLD RECORD^{TM} for sequencing a whole human genome within 5 hours and 2 minutes[2].

[1] https://www.genome.gov/human-genome-project
[2] https://www.guinnessworldrecords.com/world-records/675050-fastest%C2%A0dna-sequencing%C2%A0technique

To generate a detailed catalogue of human genetic variation from populations all around the world, the 1,000 Genomes Project was launched in 2008. With advanced sequencing methods it took only 4 years in phase 1 to collect and sequence genomes from 1,092 individuals belonging to 14 populations (1000 Genomes Project Consortium et al., 2012). After the final phase of the project (phase 3) they sequenced 2,504 human genomes from 26 populations across 5 continents, see Figure 1.2 (Sudmant et al., 2015b). Recently, the *human pangenome* was published, which is a reference genome containing 47 phased, diploid assemblies of the 1,000 Genomes Project and includes nucleotide variants, insertions / deletions and structural variants (Liao et al., 2023).

The amount of genetic data is continuously increasing and new models and theories are developed to explain the patterns found within. One objective is to identify genes which are involved in adaptation processes and have evolved under selective pressure. It is assumed that fast and extensive morphological and functional differentiation might have relied on gene duplication events with subsequent neofunctionalization (Magadum et al., 2013; Ohno, 1970). Indeed, a large portion of eukaryotic genomes is considered to be duplicated[1] and several studies indicate that multi-copy gene families are involved in adaptive processes and in maintaining genetic diversity (Perry et al., 2007; Brahmachary et al., 2014; Pajic et al., 2019; Manczinger et al., 2019).

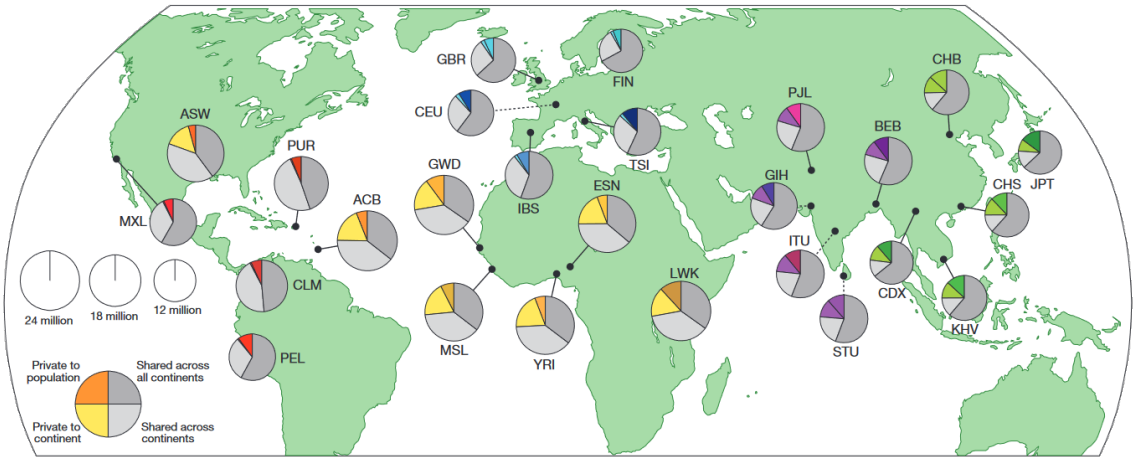


Figure 1.2: Genetic variation within the sampled populations of the 1,000 Genomes Project. The area of each pie chart is proportional to the number of polymorphisms within a population. Pies are divided into four slices, representing variants private to a population (darker colour unique to population), private to a continental area (lighter colour shared across continental group), shared across continental areas (light grey), and shared across all continents (dark grey). Dashed lines indicate populations sampled outside of their ancestral continental region. Figure taken from 1000 Genomes Project Consortium et al. (2015), Figure 1a.

[1] Around 40% in human (Zhang, 2003) and on average 65% in plants (Panchy et al., 2016).

1.2 Structure of the thesis

The aim of this thesis is to introduce a population genetic model that describes the evolution of multi-copy gene families. In the theoretical model we analyze the effects of selection strength, recombination rate and demographic patterns on the copy number and nucleotide diversity within gene families. Applied to empirical data from the 1,000 Genomes Project, we can detect gene families with copy number distributions that are likely to be involved in adaptation processes. This thesis includes three manuscripts and is organised as follows:

Chapter 2 gives a short introduction into population genetics. There are several books and review articles that provide a good overview on these topics, all with a slightly different focus – see for instance Okazaki et al. (2020); Stephan and Hörger (2019); Charlesworth and Charlesworth (2016); Wakeley (2016); Durrett (2008); Hartl and Clark (2007). However, to make this thesis self consistent, we are going to highlight and explain the basic concepts in our words.

Chapter 3 introduces the recombination model of copy number changes. The interplay of unequal recombination and selective pressure that favours genetic diversity is analyzed to understand its effect on gene copy number distribution within a population. Using empirical data from the 1,000 Genomes Project, we estimate recombination and selection parameters for three human genes. This work was published in the journal *Genetics* (Otto et al., 2022).

Chapter 4 relies on the same model extended with the demographic history of the human population to detect whether differences in African, European and Asian populations can be explained purely by demography and the out of Africa expansion, or whether shifts in the distribution are signatures of adaptation. At the time of submission, this work is available on bioRxiv and under review in the journal *G3: Genes/Genomes/Genetics* (Otto et al., 2023).

Chapter 5 gives a new interpretation of the *structured coalescent*, which is a population genetic model that includes migration. Intuitively, organisms become genetically more divergent, when they are geographically seperated. Instead of individuals travelling around, we consider gene copies to change their position along the genome according to unequal recombination and expect a present day sample of gene copies located at different positions to have a greater genetic variation than those at the same genetic position. This work was published in the journal *Theoretical Population Biology* (Otto and Wiehe, 2023).

Chapter 6 closes with a summary of the results and suggestions of possible future research questions.

2 Modelling evolution

"All models are wrong, but some are useful." George Box.

2.1 The first evolutionists

The foundations of evolutionary theory date back to Charles Darwin's famous work *"On the origin of species"* (Darwin, 1859). During his travels in the 1830s he observed several finch species on the Galapagos islands, which exhibited variations in their beak shapes, body sizes, and feeding behaviors (see Figure 2.1). Most importantly, he observed that the characteristics of finches on different islands were closely related to their specific ecological niches and dietary preferences. Darwin proposed the theory of *natural selection*, stating that beneficial traits which improve an individual's ability to survive and reproduce will become frequent in a population with time.

One of the first statistically evaluated evolutionary studies was conducted by Gregor Mendel between 1856 and 1863 (Mendel, 1865). He performed hybridization experiments on selected pea plants with distinct and easily recognizable traits, such as flower colour, seed colour, and seed texture. In his experiments he described the actions of invisible "factors", today known as "genes", and formulated the famous *laws of Mendelian inheritance*. The significance of his work was not recognized until 1900, when three scientists independently rediscovered and verified the Mendelian laws, which provided the genetic basis for understanding how traits are passed from one generation to the next.

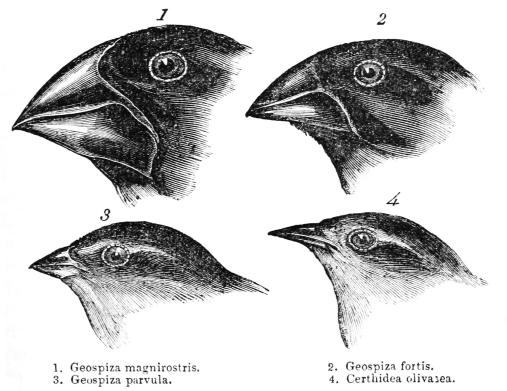


Figure 2.1: Darwin's finches or Galapagos finches. Figure taken from Darwin (1845) and also available at public domain: https://commons.wikimedia.org/wiki/File:Darwin%27s_finches_by_Gould.jpg?uselang=en#Licensing.

In 1908, Godfrey Harold Hardy and Wilhelm Weinberg independently derived one of the first mathematical descriptions of evolution, the *Hardy-Weinberg equilibrium*, which describes how allele frequencies remain constant in a sexually reproducing population (Hardy, 1908; Weinberg, 1908). It is a theoretical concept that relies on a hypothetical and idealised population satisfying the following assumptions:

- infinitely large, diploid population
- sexual reproduction
- random mating
- equal sex ratio
- non-overlapping generations
- no mutation, migration, recombination or selection

If these assumptions are fulfilled, the equilibrium describes the evolution of a single locus with two alleles (e.g., A and a). All diploid individuals carry two copies of this locus, one inherited from each parent. Denote the total frequency of A by p and respectively the frequency of a by q, such that $1 = p + q$. Then, the alleles will be distributed in the next generation as

		Females	
		A (p)	a (q)
Males	A (p)	AA (p^2)	Aa (pq)
	a (q)	Aa (pq)	aa (q^2)

(2.1)

Therefore, the total allele frequencies in the next generation are given by

$$\begin{aligned} p' &= p^2 + pq = p(p+q) = p \\ q' &= q^2 + pq = q(p+q) = q \end{aligned} \tag{2.2}$$

i.e. they remain constant and hence, genetic diversity is maintained.

Deviations from this equilibrium can therefore be interpreted as violations of the assumptions. For instance, subpopulation structure decreases the heterozygosity in a population. This is also known as the *Wahlund effect* (Wahlund, 1928). Consider a population with allele frequencies p and q, such that the expected heterozygosity is $2pq$. If the population is divided into two subpopulations with allele frequencies p_1, q_1 and p_2, q_2, it holds that

$$\frac{1}{2}(2p_1q_1 + 2p_2q_2) < 2pq, \quad \text{if } p_1 \neq p_2. \tag{2.3}$$

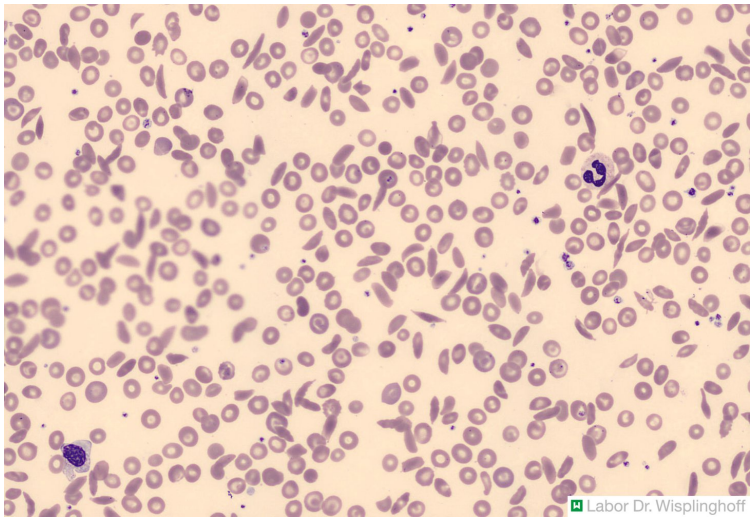


Figure 2.2: Blood smear of a patient with sickle-cell disease. Picture available at https://www.amboss.com/de/wissen/Sichelzellkrankheit.

Whereas the Wahlund-effect decreases heterozygosity, there are also cases of a selective pressure in which heterozygosity is beneficial and hence might be increased. A well-studied case of heterozygote excess is that of *sickle cell anemia* in humans. It is a hereditary disease that leads to the production of misshapen red blood cells, which take on a crescent or "sickle" shape instead of the normal round shape (see Figure 2.2). This disease is caused by a genetic mutation in the hemoglobin gene – more precisely, the inheritance of two copies of the `HgbS` allele, one from each parent (Allison, 1956). Individuals with this genotype experience significant sensitivity of their red blood cell hemoglobin to oxygen deprivation, leading to a reduced life expectancy. Heterozygous individuals inheriting one `HgbS` allele and one normal hemoglobin allele (`HgbA`) from their parents may encounter occasional health issues but typically enjoy a normal life expectancy. However, an intriguing aspect of heterozygous individuals is their resistance to malaria. This illustrates a phenomenon known as *balancing selection*, where two opposing selective forces operate. On the one hand, there is strong selection against homozygous individuals with sickle cell anemia due to their health challenges. On the other hand, malaria exerts selective pressure on individuals with standard `HgbA` alleles. As a result, heterozygous individuals exhibit a permanent advantage, or higher fitness, in regions where malaria is prevalent.

2.2 Genetic drift

The assumptions of the Hardy-Weinberg equilibrium enable derivations of mathematical results, yet they are oversimplified or even unrealistic, as for instance an infinitely large population size. The role of random fluctuations in allele frequencies due to finite population sizes, which is called *genetic drift*, was extensively studied by Wright (1931). To analyze this effect and to get a first feeling of the mathematical models in population genetics, consider a short example of a simple *Wright-Fisher* model on a finite and haploid population of constant size N (Wright, 1931; Fisher, 1930). The model progresses in discrete, non-overlapping generations, such that at each time step the entire population is replaced by N new individuals. We assume a bi-allelic single locus model and denote the two alleles as A and a and the allele frequencies as p and q, such that $1 = p + q$. Let X_t be the frequency of individuals of type A at generation t. Then, the next generation is generated according to a *binomial sampling*, that is

$$\text{Prob}\left[X_{t+1} = \frac{k}{N} \;\middle|\; X_t = p\right] = \binom{N}{k} p^k \; q^{N-k}. \tag{2.4}$$

Conceptually, $(X_t)_t$ is a *Markov process* on the state space

$$\mathcal{S} = \left\{0, \frac{1}{N}, \frac{2}{N}, ..., \frac{N-1}{N}, 1\right\}.$$

The Markov property indicates that the allele frequency in the next generation X_{t+1} only depends on the present frequency X_t and is independent of the history $X_0, ..., X_{t-1}$. In a diploid population, the population size is replaced by $2N$.

Let us have a short excursion into stochastic processes. We try to give an intuitive approach here and refer to the books of Webel and Wied (2016) and Durrett (2008) for a more detailed and precise introduction. A *stochastic process* is a sequence of random variables, where the index of the sequence is usually interpreted as time. The most well known example is the discrete *simple random walk*. Consider independent and identically distributed random variables $(Y_i)_{i\in\mathbb{N}}$ such that

$$\text{Prob}[Y_i = +1] = \text{Prob}[Y_i = -1] = \frac{1}{2},$$

and let

$$X_n = \sum_{i=1}^{n} Y_i = X_{n-1} + Y_n$$

be the sum of steps Y_i and let $X_0 = 0$ be the initial starting point. Then, X_n takes values in $\mathbb{Z}$ and in each time step, one flips a fair coin to go either one step to the left or one step to the right. The position at step $n + 1$ only depends on the position at step n and is independent of the past trajectory.

The continuous version of the simple random walk is a *Brownian motion* $(B_t)_{t\geq 0}$. Time is measured in $\mathbb{R}_+$, it starts at $B_0 = 0$ and takes values in $\mathbb{R}$. The increments are independent and Gaussian distributed, such that for any $0 \leq s < t$

$$B_t - B_s \sim \mathcal{N}(0, t-s). \tag{2.5}$$

An intuitive approach to Brownian motions is *Donsker's invariance principle* (Donsker, 1951), which states that, properly rescaled in time and space, the simple random walk converges to a Brownian motion, see Figure 2.3. It can be interpreted as *"random noise"* around 0.

The more general class is that of *diffusion processes*, which are defined by stochastic differential equations. More precisely, given a standard Brownian motion $(B_s)_{s\geq 0}$, we call the $\mathbb{R}$-valued stochastic process $(X_t)_{t\geq 0}$ a diffusion process, if it satisfies for given Lipschitz-continuous functions $b(x)$ and $\sigma(x)$

$$X_t - X_0 = \int_0^t b(X_s)\,\mathrm{d}s + \int_0^t \sigma(X_s)\,\mathrm{d}Bs. \tag{2.6}$$

The functions $b(x), \sigma^2(x)$ are the *infinitesimal mean* and *infinitesimal variance* of the process. As an example, let $\sigma(x) = 0$ and $b(x) = b$ constant. Then

$$X_t - X_0 = \int_0^t b(X_s)\,\mathrm{d}s = b \cdot t,$$

so X_t is just a linear function with slope b. Vice versa, let $b(t) = 0$ and $\sigma(t) = \sigma$ constant. Then

$$X_t - X_0 = \int_0^t \sigma(X_s)\,\mathrm{d}Bs = \sigma \cdot (B_t - B_s),$$

i.e. X_t is a Brownian motion with variance σ^2. Combined, a diffusion process is a stochastic process with deterministic mean function $b(x)$ and random fluctuations according to $\sigma(x)$.

Getting back to the Wright-Fisher process, Kimura (1964) derived the diffusion limit of the allele frequency changes in a neutral population. According to the binomial transition probabilities (2.4) the allele frequency is expected to remain constant but changes randomly due to drift. This can be approximated by a diffusion process with infinitesimal mean 0 and variance

$$\sigma^2(x) = \frac{1}{N} \cdot x(1-x). \tag{2.7}$$

In a diploid population, N is replaced by $2N$. If $N \to \infty$, the effect of random fluctuations become smaller and the allele frequencies remain constant as in the Hardy-Weinberg equilibrium. In finite populations, alleles may get fixed or lost due to drift and hence genetic

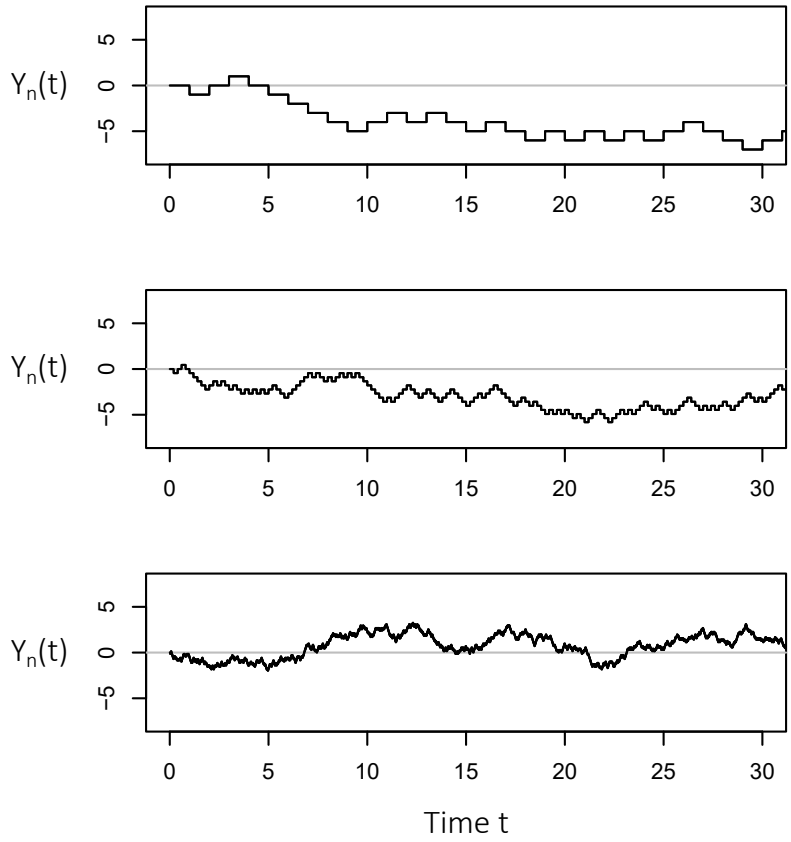


Figure 2.3: Donsker's Theorem. Let $Y_n(t) = \frac{1}{\sqrt{n}} \sum_{i=0}^{\lfloor tn \rfloor} X_i$ and $n \to \infty$, then the process converges to a Brownian motion. Figure shows a simple random walk with scaling $n = 1, 5, 50$ (top to bottom).

variability decreases. To quantify the loss of genetic diversity, consider a diploid population of size N. The probability that two randomly chosen genes have the same ancestor in the previous generation and are hence "identical by descent" (IBD) is

$$\text{Prob[IBD]} = \frac{1}{2N}.$$

Denote the degree of genetic diversity in the population at time $t = 0$ as H_0, which is the complement of being identical by descent. Then, after one generation at time $t = 1$ one finds

$$H_1 = H_0 \left(1 - \frac{1}{2N}\right), \tag{2.8}$$

and iterating gives

$$H_t = H_0 \left(1 - \frac{1}{2N}\right)^t \approx H_0 \cdot e^{-\frac{t}{2N}}. \tag{2.9}$$

Therefore, the genetic diversity decays exponentially with parameter $1/(2N)$. As a calculation example, consider a population of size $N = 10{,}000$. Then, the initial diversity is reduced by 50% after

$$t = 2N \cdot \log(2) \approx 13{,}863 \text{ generations.}$$

Wright (1931) also introduced the concept of *effective population size* N_e, which represents the idealized size of a population that would experience the same amount of genetic drift as the actual population. For instance, a sudden reduction in population size increases the effect of drift, as genetic diversity decreases faster in a smaller population. Even if the population size is recovered shortly after, the effective population size N_e can be heavily decreased by such a *bottleneck*.

For example, consider a population of size $N = 10{,}000$ over a time period of 200 generations, that experienced a population size reduction to 100 individuals between generation 50 and 60. Without mutation, genetic diversity is expected to be reduced to

$$H_{200}/H_0 = \prod_i \left(1 - \frac{1}{2N_i}\right) = \left(1 - \frac{1}{20{,}000}\right)^{190} \cdot \left(1 - \frac{1}{200}\right)^{10} \approx 0.942$$

A population of constant size N_e that would experience the same genetic drift can be derived by

$$\left(1 - \frac{1}{2N_e}\right)^t = \prod_i \left(1 - \frac{1}{2N_i}\right), \tag{2.10}$$

which solves to

$$N_e \approx \frac{t}{\sum_i 1/N_i} \approx 1{,}680.$$

The effective population size N_e of a non-constant population can therefore be calculated by its *harmonic mean* over time. Even though this bottleneck lasted for only 10 generations, it reduced the N_e from 10,000 to 1,680. Also, non-random mating, migration or an unequal sex ratio may affect the effective population size N_e.

2.3 Infinite alleles

While genetic drift reduces the diversity of a population, mutations introduce new alleles into the population, thereby increasing the genetic variability. Consider mutations to occur at rate μ, such that any mutation event generates a completely new and unique allele. We call this concept the *infinite alleles model* (Kimura and Crow, 1964).

There are now two forces shaping genetic variation: mutations, which introduce new alleles into the population and increase genetic diversity, and drift, which decreases diversity, since some alleles get lost by chance. Using the probability to not be identical by descent H_t as measure of diversity, we find analogue to equation (2.8)

$$H_1 = H_0 \left(1 - \frac{1}{2N}\right) + (1 - H_0)2\mu. \tag{2.11}$$

The question arises, whether there is an equilibrium of these two forces, i.e. $\bar{H}$, such that

$$\bar{H} = \bar{H} \left(1 - \frac{1}{2N}\right) + (1 - \bar{H})2\mu, \tag{2.12}$$

and indeed, solving this equation gives

$$\bar{H} = \frac{4N\mu}{4N\mu + 1} = \frac{\theta}{\theta + 1}, \tag{2.13}$$

where $\theta = 4N\mu$ is the *population scaled mutation rate* ($\theta = 2N\mu$ in a haploid population). In a finite population, this is not a static but dynamic equilibrium, since allele frequencies fluctuate. However, over a long time some alleles are lost or fixed and new alleles appear, contributing to the long term equilibrium. As an example, consider a population of size $N = 1{,}000$ over a time period of 2,000 generations with mutation rate $\mu = 0.0006$. Without mutation, one expects a loss of heterozygosity according to (2.9), whereas with mutation we expect the population to fluctuate around the equilibrium (2.13). Figure 2.4A shows mean heterozygosity of 50 replicates of a population evolving forward in time according to a Wright-Fisher process. We also included a bottleneck at generation 500 to see the increased effect of drift in a small population (see Figure 2.4B).

In the infinite alleles model, the number of new alleles is not limited. Nevertheless, at any given time point there is only a finite, typically small, number of alleles segregating in a population. The probability distribution of the number of distinct alleles at a particular genetic locus in a finite population was derived by Ewens (1972). Denote by a_i the number of alleles which are represented i times in a sample of size n, so that a_1 counts the number of unique alleles, a_2 those that occur twice etc. Then, with scaled mutation rate $\theta = 4N\mu$ we find

$$P_{\theta,n}(a_1, ..., a_n) = \frac{n!}{\theta(\theta + 1) \cdots (\theta + n - 1)} \prod_{j=1}^{n} \frac{(\theta/j)^{a_j}}{a_j!}. \tag{2.14}$$

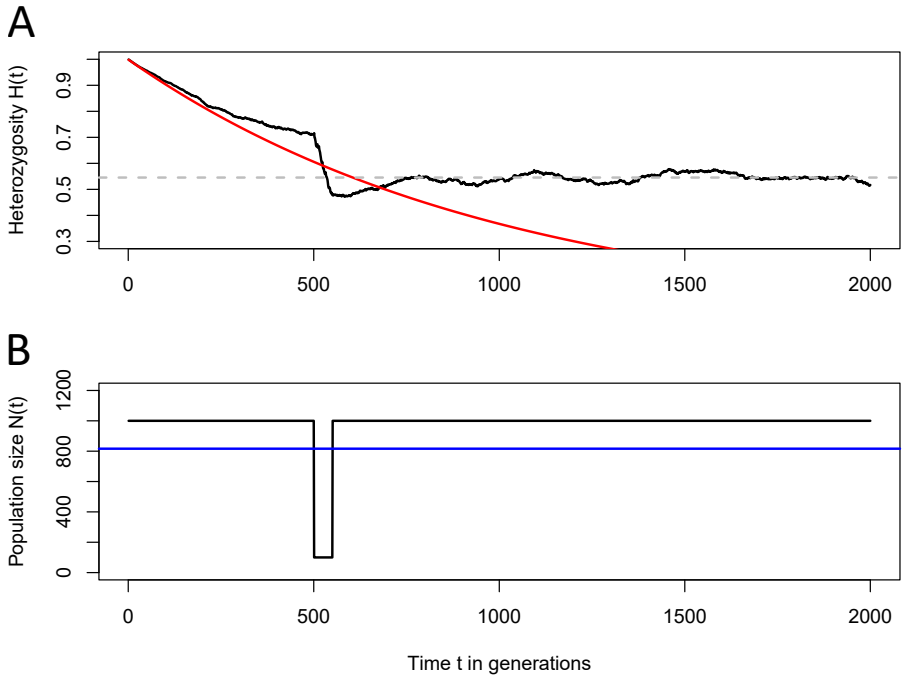


Figure 2.4: **A**. Simulated mean heterozygosity of 50 replicates of a population over time with mutation $\mu = 0.0006$ and population size $N(t)$, as shown below. Red line shows expected exponential decay of heterozygosity without mutation (2.9) and dotted line the mutation drift equilibrium (2.13). **B**. Population size $N(t)$ over time, with population size reduction from 1,000 to 100 during generation 500-550. Blue line shows effectice population size $N_e \approx 816$.

As an example, consider a sample of $n = 20$, with the following allele counts:

A	B	C	D	E
4	8	4	2	2

This gives the partition $(a_1, a_2, ..., a_n) = (0, 2, 0, 2, 0, 0, 0, 1, 0, ..., 0)$, since the two alleles D, E occur twice, A, C four times and B eight times. Using *Ewen's sampling formula* (ESF), we calculate for different values of θ the probability to sample such a partition, resulting in a likelihood function of θ. A low mutation rate would result in only a few (unique) alleles, whereas a high mutation rate would result in many unique alleles. For this particular example, we find the most likely θ at ≈ 1.8.

2.4 Coalescence theory

While the Wright-Fisher model describes evolution of a whole population of size N forward in time, Kingman (1982a,b) introduced a model in which the ancestry of a present day sample of size $n \ll N$ is traced backward in time. Given two haploid individuals in a population of constant size N, they may be offspring of the same ancestor with probability $1/N$. Hence, considering a sample of $n \ll N$ individuals, the probability that two randomly chosen genes are identical by descent is $\binom{n}{2} \cdot 1/N$. Iteratively, all lineages of the sample can be traced back to their *most recent common ancestor* (MRCA), see Figure 2.5. This backward-in-time model of merging lineages is known as the *Kingman coalescent*. The time – measured in generations of units N – until a set of k lineages collapses to a set of $k-1$ lineages is denoted by T_k and $\text{Exp}(\binom{k}{2})$-distributed. Therefore, the expected time to the most recent common ancestor (T_{MRCA}) in generations is given by

$$E\left[T_{MRCA}\right] = \sum_{k=2}^{n} E[T_k] = \sum_{k=2}^{n} \frac{2}{k(k-1)} \cdot N = 2N\left(1 - \frac{1}{n}\right). \tag{2.15}$$

The expected total tree length, i.e. the sum of all branch lengths, is given by

$$E\left[\text{Tree-length}\right] = \sum_{k=2}^{n} kE[T_k] = \sum_{k=2}^{n} \frac{2}{k-1} \cdot N = 2Nh_{n-1}, \tag{2.16}$$

where h_n denotes the *n-th harmonic number*. In the case of diploid organisms, the probability of coalescence is $1/2N$, and hence, time is scaled with factor 2. As an example with parameters motivated from *Homo sapiens*, consider a diploid population of effective size[1] $N_e = 10{,}000$ with a generation time of 20 years and a sample of $n = 10$. The expected time to the most recent common ancestor of these 10 individuals is therefore given by

$$E[T_{MRCA}] = 4N\left(1 - \frac{1}{n}\right) \cdot 20 = 800{,}000 \cdot 0.90 = 720{,}000 \text{ years}.$$

Throughout this time, mutations in the DNA increase the genetic diversity. While the infinite alleles model arose at a time when indirect methods had to be used to infer differences between individuals, the technical tools to sequence DNA in the 1960s enabled analysis of nucleotide differences and Kimura (1969, 1971) introduced the *infinite sites model*. When analyzing the sequenced data of a genetic locus, it assumes that any mutation changes one nucleotide and occurs on a unique site that did not mutate before. As an example, consider the pattern in Figure 2.5. Mutations are marked as red dots along the tree and change one particular nucleotide in the sequence. The differences in the sequence sample are called *single nucleotide polymorphisms* (SNPs). If the *ancestral state* is known, the information of the mutations can be represented in a SNP-matrix (see Figure 2.5), where 1 represents a

[1] Note, that the effective population size N_e reflects the population size of an idealized population, that would experience the same amount of genetic drift as the population under consideration. Due to several different factors, such as subpopulation structure, population size changes, migration, unequal sex ratio etc., N_e is significantly smaller than the true population size.

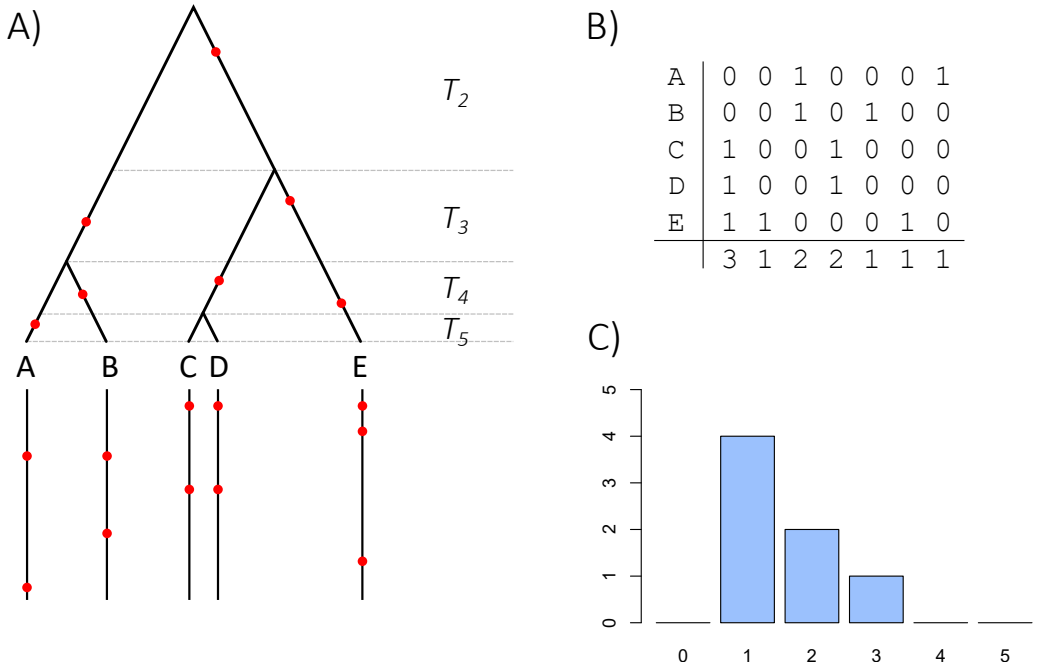


Figure 2.5: **A**. Illustration of the coalescence process for $n = 5$. Coalescence times are indicated as T_2, T_3, T_4, T_5, mutations are marked as a red dot. **B**. SNP-matrix. If the ancestral state is known, mutations are indicated as a 1, the ancestral state as 0. **C**. Site frequency spectrum. Four mutations occur once, two twice and one mutation three times.

mutation and 0 the ancestral state. In this example, there are 7 polymorphic sites, distributed on a sample of 5 haploid individuals. Four of these mutations affect only one individual and are therefore called *singletons*. The frequency of singletons, doubletons etc. is denoted by $\xi_1, \xi_2, ...$ and called *site frequency spectrum* (SFS).
Using coalescence theory, Fu (1995) derived the expected SFS

$$E[\xi_i] = \theta \frac{1}{i}. \tag{2.17}$$

Note, that often the ancestral state is unknown and hence one can not distinguish, which of the two variants of a nucleotide results from the mutation event. In the *folded site frequency spectrum* the minor alleles are counted and the i-classes and the $n - i$-classes are combined, i.e.

$$f_k = \xi_k + \xi_{n-k}.$$

However, in the following models we assume the ancestral state to be known.

Since the total tree length of a sample of n individuals is given by $4Nh_{n-1}$, one expects a total number of $4Nh_{n-1}\mu$ segregating sites in the sample. Using this property, Watterson (1975) introduced an estimator of the mutation rate θ as

$$\hat{\theta}_W = \frac{S}{h_{n-1}}, \tag{2.18}$$

which is today known as *Watterson's estimator*, where S denotes the total number of segregating sites in the sequenced sample. Another estimator was introduced by Tajima (1989), which is based on pairwise nucleotide differences. Since the pairwise T_{MRCA} is $2N$ generations, one expects $4N\mu = \theta$ mutation events to happen along the branches that separate two individuals. Therefore, the average nucleotide diversity π is defined as

$$\pi = \frac{2}{n(n-1)} \sum_{i<j} d_{ij}, \tag{2.19}$$

where d_{ij} is the number of differences between the ith and the jth sequence. The expected pairwise nucleotide diversity is given by $E[\pi] = \theta$, and hence Tajima introduced the estimator

$$\hat{\theta}_\pi = \frac{2}{n(n-1)} \sum_{i<j} d_{ij}. \tag{2.20}$$

Considering again the example of Figure 2.5, we can use three different approaches to estimate θ: the maximum likelihood estimate of Ewen's sampling formula (considering the infinite alleles model), Watterson's estimator $\hat{\theta}_W$ using the number of segregating sites S and Tajima's nucleotide diversity estimator $\hat{\theta}_\pi$. The infinite alleles model can only decide, whether two alleles are identical or not. In this example, we find three single-type alleles and one that occurs two times, leading to a partition $(a_1, a_2, a_3, a_4, a_5) = (3, 1, 0, 0, 0,)$. Analyzing the DNA sequences we find $S = 7$ positions in which they differ and a mean nucleotide diversity in the sample of $\pi = 4.2$. This leads to

$$\begin{aligned}
\hat{\theta}_\pi &= \frac{2}{n(n-1)} \sum_{i<j} d_{ij} = 4.2 \\
\hat{\theta}_W &= S/h_{n-1} = 7/2.083 = 3.36 \\
\hat{\theta}_{ESF} &= \text{argmax}_\theta \left(P_{\theta,n}(a_1, ..., a_n) \right) = 7.1
\end{aligned}$$

There are several other frequency spectrum based estimators of θ, as for instance Fu and Li's estimator, which just considers the singleton class $\hat{\theta}_{FL} = \xi_1$ (see equation (2.17)) or Fay and Wu's estimator, that gives most weight to high-frequency classes

$$\hat{\theta}_H = \binom{n}{2} \sum i^2 \xi_i.$$

Under neutrality, one expects the estimators to coincide. Hence, deviations may be indicators of violations of the underlying assumptions in the model. This gave rise to test statistics that measure the differences of the estimators. The most commonly known are *Tajima's D* (Tajima, 1989), defined as

$$D = \frac{\hat{\theta}_\pi - \hat{\theta}_W}{\sigma_D}, \tag{2.21}$$

and *Fay and Wu's H*, (Fay and Wu, 2000) defines as

$$H = \frac{\hat{\theta}_\pi - \hat{\theta}_H}{\sigma_H}. \tag{2.22}$$

For example, a value of $D > 0$ indicates a higher heterozygosity than the number of segregating sites would predict, which results in a lack of rare alleles in the SFS. A negative value on the other hand indicates a low heterozygosity and many rare alleles. It is one of the greatest tasks in population genetics to identify such regions that deviate from neutral theory and to understand which evolutionary forces are responsible for these differences.

2.5 Selection

While Wright considered drift to be the primary driving force in shaping genetic diversity, Fisher (1930) and Haldane (1927) argued that *natural selection* plays a more significant role. In the context of a Wright-Fisher-model, considering the type A as beneficial compared to a, one may model selection with weighted sampling. The allele A with higher fitness is given a sampling weight of $1+s$ with fitness parameter $s \geq 0$, whereas a is sampled with weight 1.

Intuitively, a beneficial or *positively selected* mutation becomes fixed in a population with higher probability and on average faster than a neutral one. Considering a selective factor $s > 0$ and an infinitely large population size, the frequency of a beneficial allele X_t can be described by a logistic growth differential equation

$$\frac{\mathrm{d}}{\mathrm{d}t} X_t = s X_t (1 - X_t), \tag{2.23}$$

which solves to

$$X_t = \frac{X_0}{X_0 + (1 - X_0) e^{-st}}, \tag{2.24}$$

where X_0 denotes the initial allele frequency at time $t = 0$. Kimura (1962, 1964) showed, that in a finite population of size N the probability of fixation of such a mutation is given by

$$\frac{1 - e^{-4N \cdot s \cdot X_0}}{1 - e^{-4N \cdot s}}. \tag{2.25}$$

Hence, a newly arising allele $X_0 = 1/2N$ with a relative selective benefit of $s = 0.01$ becomes fixed with probability $\approx 2\%$ in a diploid population of $N = 10{,}000$. The time to fixation (say $X_t = 1 - 1/2N$) can be calculated by solving

$$1 - X_0 = \frac{X_0}{X_0 + (1 - X_0) e^{-st}}, \tag{2.26}$$

which gives

$$t = \frac{2}{s} \log(2N - 1) \approx 1,981 \text{ generations}. \tag{2.27}$$

In contrast, a neutral allele reaches fixation due to drift with probability $1/2N = 0.00005$ and the time to fixation is on average $4N = 40{,}000$ generations.

A famous example of such a beneficial mutation is a SNP in the gene encoding lactase (`LCT`), which is associated with the ability to digest milk as adults (lactase persistence) in Europeans (Tishkoff et al., 2006). It is hypothesized to be a beneficial trait which increased the fitness of human populations that traditionally practiced cattle domestication, since it is only found in 1% in non-pastoralist Asian and African populations, but up to 90% in Swedes and Danes (Swallow, 2003; Hollox, 2004).

Pairing the two allele variants in a diploid population leads to three different genotypes, that may all have different fitness values. Considering type A as beneficial compared to a, define fitness with the *dominance factor* h as

Genotype	fitness
AA	$1+2s$
Aa	$1+2hs$
aa	1

If $h = 0$, the allele is *recessive* and the fitness benefit only sets in if an individual carries both beneficial alleles. In contrast, if $h = 1$, one allele is already sufficient to produce the full selective benefit and hence called *dominant*. If $h = 1/2$, it is called *co-dominant*. If $0 < h < 1$, the allele frequency follows a *directional selection*, since the evolution is directed towards the fixation of allele A. An interesting case occurs if $h > 1$ and aa also has fitness of $1 + 2s$, same as AA. Then, heterozygotes have the greatest fitness, which is called *overdominance* or *balancing selection*. A famous example of this phenomenon is the previously mentioned sickle cell anemia, in which the heterozygote defeats both malaria and the sickle shaped red blood cells.

Those prime examples are rare and most of the mutations have either no effect at all or lead to the loss of gene function and are slightly or severely deleterious. Ohta (1973) introduced the *almost neutral theory*, in which she postulated that there are mutations with a small negative selection coefficient, also known as *purifying selection*, such that

$$|N_e s| < 1.$$

With such small s the effect of drift is still one of the main forces in evolution and even if a mutation is solely deleterious, it may reach fixation in the population due to drift. Consider a population of wild-types only, all with fitness 1. Mutations occur at rate μ and create alleles with lower fitness, such that an individual with k mutations compared to the wild type has fitness

$$(1-s)^k. \tag{2.28}$$

There are two forces counteracting on the population: selection removes alleles of low fitness from the population, whereas mutation generates new alleles with low fitness. Considering a population without drift, Haigh (1978) showed that the population reaches a mutation-selection equilibrium, where the relative frequency f_k of individuals with k mutations is given by

$$f_k = \frac{1}{k!}\left(\frac{\mu}{s}\right)^k e^{-\frac{\mu}{s}}, \tag{2.29}$$

i.e. the classes follow a Poisson distribution. However, in a finite population the fittest class that has no mutations might get lost due to drift. And since mutations only decrease fitness

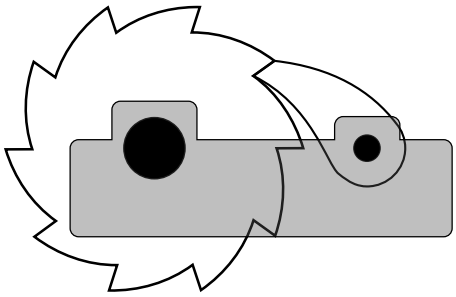


Figure 2.6: Illustration of Muller's ratchet. The pawl prevents the ratchet wheel from turning backwards. In the same way, the Muller ratchet can only be turned in the "direction" of additional, harmful gene mutations.

and back mutations are neglected, this class is indeed lost forever and can not be restored. But then, the class with one mutation becomes the relatively fittest in the population and the distributions of (2.29) shift by one. Again, the fittest class might get lost forever and the population fitness decreases over time, with no possibility to gain it back. This process is known as *Muller's ratchet* (Muller, 1964) and illustrated in Figure 2.6.

The time between clicks of the ratchet, i.e. the loss of the currently fittest class, was approximated by Haigh (1978) and is given by

$$E[\text{Time between clicks}] \approx 4N \cdot e^{-\frac{\mu}{s}} + 7\log(\frac{\mu}{s}) + \frac{2}{s} - 20. \tag{2.30}$$

This time is obviously quite long in large populations, but on evolutionary time scales this process would nevertheless lead to the *mutational meltdown*. Indeed, such a process was demonstrated in an evolutionary experiment in yeast (Zeyl et al., 2007). However, several factors could either greatly slow down or even stop the ratchet mechanism, e.g. compensatory mutations (Wagner and Gabriel, 1990), beneficial mutations (Rouzine et al., 2008) or synergistic epistasis (Kondrashov, 1994; Jain, 2008). Another effective mechanism to escape the ratchet is *recombination*, by which descendants may inherit fewer mutations than their parents if the parental genotypes have mutations at different loci.

2.6 Recombination

Recombination is the biological process in which genetic material is exchanged between two homologous chromosomes. In diploid eukaryotes, specific DNA segments break and recombine with their counterparts on another chromosome during meiosis. Therefore, we now consider a two loci model in a diploid population. In its simplest form we call them the $\mathcal{A}$ locus with alleles A and a and the $\mathcal{B}$ locus with B and b alleles. This results in four different haplotypes: AB, Ab, aB and ab. Given two haplotypes, a recombination event may happen with probability r and fuse the counterparts together (see Figure 2.7). Denote the frequencies of the haplotypes as f_{AB}, f_{Ab}, f_{aB} and f_{ab}, such that $1 = f_{AB} + f_{Ab} + f_{aB} + f_{ab}$ and the allele frequencies as $1 = f_A + f_a$ and $1 = f_B + f_b$. If we choose two haplotypes, we may get the allele AB in the next generation with probabilities

Recombine	AB	Ab	aB	ab
AB	1	$\frac{1}{2}$	$\frac{1}{2}$	$\frac{1}{2}$(1-r)
Ab	$\frac{1}{2}$	0	$\frac{1}{2}$r	0
aB	$\frac{1}{2}$	$\frac{1}{2}$r	0	0
ab	$\frac{1}{2}$(1-r)	0	0	0

Therefore, with recombination rate r and without drift the allele frequency of AB at time $t+1$ is given by

$$\begin{aligned} f_{AB}(t+1) &= f_{AB}(t) + \frac{1}{2} r f_{AB}(t) f_{Ab}(t) + ... + \frac{1}{2}(1-r) f_{ab}(t) f_{AB}(t) \\ &= f_{AB}(t) - r(f_{AB}(t) \cdot f_{ab}(t) - f_{Ab}(t) \cdot f_{aB}(t)) \\ &= f_{AB}(t) - r L_D(t), \end{aligned}$$

where we define L_D as the *linkage disequilibrium*:

$$L_D = f_{AB} \cdot f_{ab} - f_{Ab} \cdot f_{aB}.$$

If $L_D = 0$, we find that $f_{AB} = f_A \cdot f_B$ (and the analogous results for the other haplotype frequencies) and the two loci are in *linkage equilibrium*, which is the "horizontal" analogue of the Hardy-Weinberg equilibrium. Therefore, we can describe the haplotype frequencies over time as

$$f'_{AB}(t) = -r L_D(t), \quad f'_{ab}(t) = -r L_D(t), \quad f'_{Ab}(t) = r L_D(t), \quad f'_{aB}(t) = r L_D(t).$$

Consequently, we find that

$$L'_D(t) = ... = -r \cdot L_D(t),$$

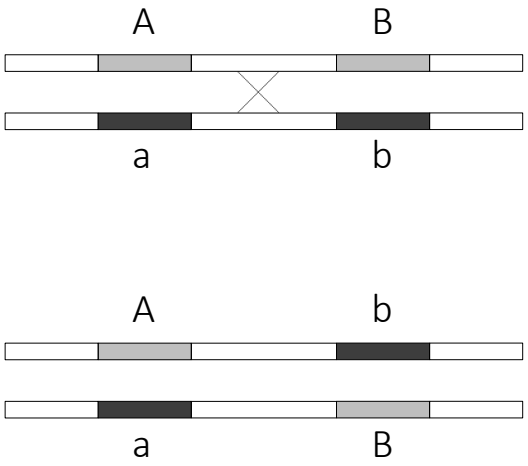


Figure 2.7: Illustration of the recombination process in a two loci model.

which means that for $L_D(0) = d_0$ the linkage disequilibrium decays exponentially over time with

$$L_D(t) = d_0 \cdot e^{-rt}. \tag{2.31}$$

It also decreases exponentially with genomic distance (implicitly given by the recombination rate r). Therefore, two loci that are at distant positions and evolve under neutrality over a long time should be in linkage equilibrium.

Linkage disequilibrium arises, for example, when a new allele with a fitness benefit has recently been fixed. Consider such a beneficial mutation B' on locus $\mathcal{B}$ and let it be under strong directional selection, i.e. large s and N_e, such that $N_e s \gg 1$. Then, the new allele B' will eventually get fixed in the population. Without recombination, the entire haplotype becomes fixed, which is either aB' or AB'. The neutral locus "hitchhikes" with the beneficial allele as if it was also under directional selection. This *genetic hitchhiking* leads to a loss of genetic variability at locus $\mathcal{A}$ (Smith and Haigh, 1974). The genomic signal of such a *selective sweep* is illustrated in Figure 2.8: The strength of this effect depends on the selective strength $N_e s$, the genomic distance between the loci and the time that has passed since the fixation of the allele.

As an example, consider two completely linked loci in a distance of 10,000 base-pairs (short: 10kb) in human. As a rule of thumb, the recombination rate per nucleotide per generation is about 10^{-8}, therefore the recombination rate between these two loci can be considered as $r = 0.0001$. Then, if we neglect the effect of drift, linkage disequilibrium decays exponentially according to (2.31) and decreases to 50% after

$$t = \frac{1}{r}\log(2) \approx 7{,}000 \text{ generations.}$$

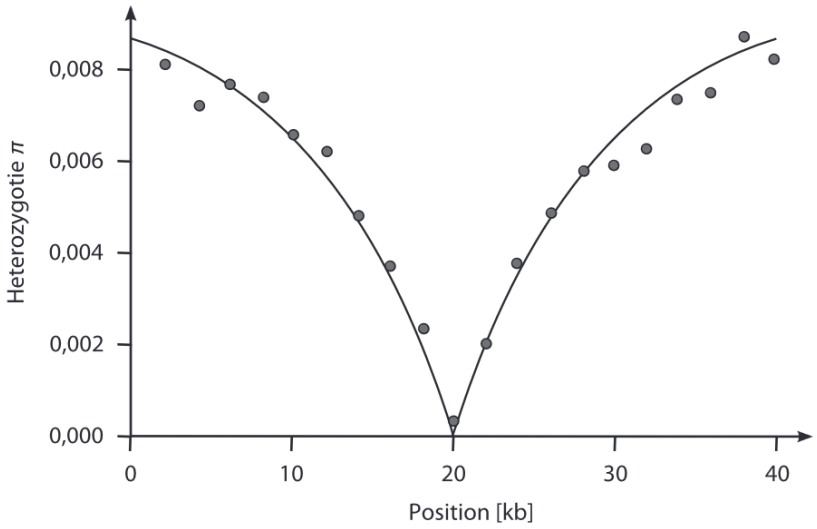


Figure 2.8: Simulated heterozygosity in a region of a beneficial selected genetic locus. Taken from Stephan and Hörger (2019), Figure 8.2 (originally Kim and Stephan (2002), Figure 2).

If we also apply the effect of drift, the linkage disequilibrium would decline even more rapidly and the signal of the recent selective sweep will vanish.

Let us now consider the backward in time process of two loci with recombination in a sample of size n in the coalescent process. If a recombination event splits the haplotype into two pieces, the number of lineages increases, since locus $\mathcal{A}$ now follows the maternal genealogy and $\mathcal{B}$ the paternal. The coalescence tree of $\mathcal{A}$ differs from the one of $\mathcal{B}$ and changes along the whole genome. Therefore, the entire ancestral process features coalescence and splitting events and we call this concept the *ancestral recombination graph* (ARG) (Hudson, 1983). An example is shown in Figure 2.9.

The ancestral recombination graph is an important extension to Kingman's classical coalescence theory. Without recombination, the following sample can not be explained with just one coalescence tree, if we consider the infinite sites model:

Individual	Locus 1	Locus 2
A	0	0
B	1	0
C	0	1
D	1	1

On the one hand, the mutation at locus 1 affects both individuals B and D. Therefore, there has to be a branch that connects these two and only these two. But on the other hand, the mutation at locus 2 affects C and D, so that these two have to share a common branch. Concluding, one has to either reject the assumption of the infinite sites model, or include the possibility of different coalescence trees along the genome, i.e. recombination.

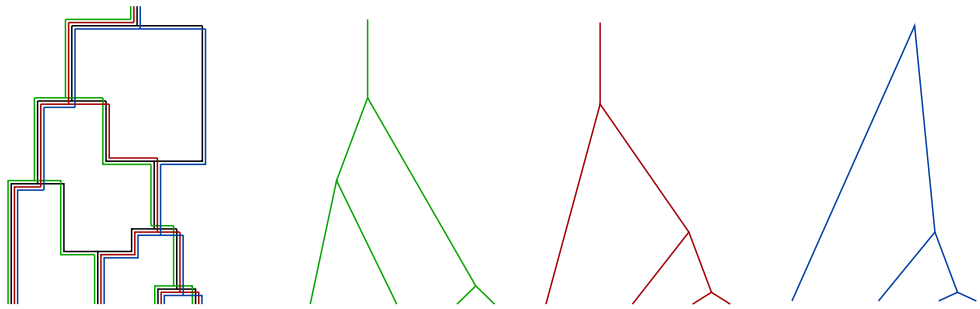


Figure 2.9: Ancestral recombination graph. A sample of $n = 4$ in a three locus model. Two recombination events result in three different genealogies in the haplotype.

The inference of all coalescence trees along the DNA is a complex problem. An approach is given by the *Sequential Markov Chain* (SMC) (McVean and Cardin, 2005; Rasmussen et al., 2014). The Markov approach considers changes not across time, but along the genome and assumes, that the tree at locus $x+1$ depends on the tree at position x. The transition process is given by the procedure of *prune and regrafting*, which means that if a recombination event has happened at that locus, the tree at the next position can be produced by choosing a "break point" on the tree and a "clipping" point to attach the branch back to the tree. This is only an approximation, but it provides an efficient way to simulate and infer the coalescence trees along the genome.

A famous application of this idea is the *Pairwise Sequentially Markov Coalescent* (PSMC) (Li and Durbin, 2011) and its successors *Multiple Sequentially Markov Coalescent* (MSMC and MSMC2) (Schiffels and Durbin, 2014; Schiffels and Wang, 2020) and *extended Sequential Markov Coalescent* (eSMC) (Wang et al., 2022). They use a *Hidden Markov Model* (HMM)[1] approach to infer the population size history from whole-genome data. Given the whole DNA of a single diploid individual, the genealogy and hence time to the most recent common ancestor (T_{MRCA}) changes along the genome, which is approximated by the Sequentially Markov Chain. If T_{MRCA} is large, one expects high genetic diversity in the present day sample. Vice versa, given the genetic variation in a present day sample, one can infer the most likely sequence of T_{MRCA} along the genome. And since the probability of coalescence is negatively correlated to the population size, one can estimate the population size history.

In its original version, Li and Durbin (2011) used PSMC on single genomes of humans from Africa, Asia and Europe and showed that the findings are in agreement with the founder event (bottleneck) in East Asian and European populations, associated with the out of Africa event $\approx$ 60,000 years ago. This is a fascinating example of how signatures of evolutionary events that date back far in the past can still be detected in present day DNA.

[1] A Hidden Markov Model considers a Markov process, in which the current state of the process can not be detected (i.e. it is hidden). However, in any time step, the state changes according to the transition matrix of the Markov model and a detectable signal is emitted. Here, we consider the true genealogy to be the hidden states and the DNA sequence as the emitted signal. See (Rabiner, 1989; Fink, 2008).

2.7 Population structure

The Hardy-Weinberg equilibrium assumes panmixia, i.e. the hypothetical situation with complete random mating within one large population. We already mentioned the *Wahlund-effect*, which states that the mean heterozygosity decreases if a population is divided into isolated subpopulations (2.3). Intuitively, if a panmictic population is divided into two isolated "island"-populations, they will evolve independently and become genetically differentiated over time. To measure the degree of reduction in heterozygosity, Wright (1951) introduced his hierarchical F-statistics. They measure the proportion of genetic variance among subpopulations relative to the total genetic variance in the entire population and range from 0 (no differentiation) to 1 (complete differentiation). There are three F-statistics:

- F_{ST}, the *fixation index*
- F_{IS}, the *kinship coefficient*
- F_{IT}, the *total inbreeding*.

Here, we only focus on F_{ST}, which measures the genetic differentiation among subpopulations. The S denotes the *subpopulation*, T the *total population* and I is the abbreviation of an *individual*. For alleles at a single locus, F_{ST} is defined in terms of probabilities of identity as

$$F_{ST} = \frac{f_S - f_T}{f_T}, \tag{2.32}$$

where f_S is the probability that two genes sampled at random from a single subpopulation are identical (carry the same allele), and f_T is the probability that two genes randomly chosen from the collection of subpopulations considered are identical.

In the context of population genetics, *migration* refers to the movement of individuals from one (sub-) population to another, with the potential to introduce new genetic material. This exchange of alleles between different populations of a species is also known as *gene flow*. One of the first and simplest models is the *continental island* introduced by Wright (1940), see Figure 2.10A. Motivated by the South-American continent and the Galapagos island, we consider an infinitely large population on the continent and a finite population of size N_e on the island. At each generation, M individuals migrate from the continent to the island and replace M individuals there. The migration rate is denoted by $m = M/N_e$. The migration process can be seen as the equivalent of the mutation in (2.13), since it introduces new alleles, whereas drift reduces genetic diversity. Therefore, analogue to (2.13) we find the equilibrium

$$\bar{F}_{ST} = \frac{1}{1 + 4N_e m}. \tag{2.33}$$

Another population configuration model is given by the *symmetric island model*, where one considers d islands, all with equal population sizes N_e and connected with each other with equal migration rate m (Figure 2.10B). This means that individuals from each island have

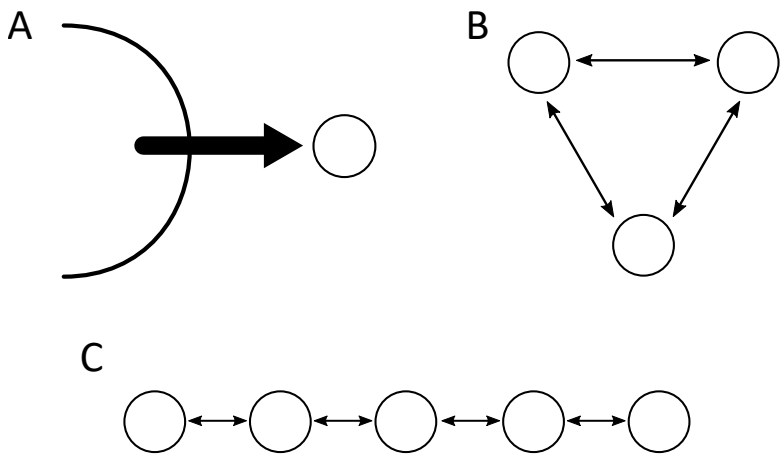


Figure 2.10: Illustration of the three migration models. A: Continental model. B: Symmetric island model with $d = 3$ islands. C: One-dimensional stepping stone model.

an equal chance of migrating to any other island. If the number of islands d is large, the equilibrium (2.33) also holds true (Hudson, 1998).

The *stepping stone model* takes into account the spatial arrangement of subpopulations (Figure 2.10C). It assumes that migration is more likely to occur between neighboring islands and therefore one expects a higher genetic differentiation in populations at higher geographical distance.

In general, any migration dynamics can be modeled, with migration rates m_{ij} from island i to island j. However, with increasing complexity the calculations of analytical results become more challenging.

While the equilibrium (2.33) can be derived under the assumption of a forward in time neutral Wright-Fisher process, it is only natural to apply population structure to Kingman's coalescence model (Takahata, 1988). Going backward in time, two individuals can only be offspring of the same ancestor, if they are located at the same island. Otherwise, one or mulitple migration events have to have happened in the past. We describe this process in more detail in chapter 5.2. As an example, consider a symmetric $d = 2$ island model (islands A and B) with equal migration rates m and equal population size N. Taking a sample of $n = 2$ individuals, they can be either located at the same or different islands. We denote the time to their most recent common ancestor as $T_{AA} = T_{BB}$ and T_{AB}. Then, it can be shown (see for example Wilkinson-Herbots (1998)) that

$$E[T_{AA}] = E[T_{BB}] = 2N, \qquad E[T_{AB}] = 2N + \frac{1}{m}N. \tag{2.34}$$

Therefore, with high migration rate m the coalescence times are close to the standard panmictic case, whereas with low migration and hence high isolation the coalescence time of two individuals on different islands can be severely increased.

2.8 Gene duplications

When measuring genetic diversity, we considered up to this point only single nucleotide polymorphisms (SNPs). However, structural variants such as inversions, deletions, insertions and duplications contribute significantly to genetic variability. One of the first reported observations of duplicated materials was the bar eye locus in *Drosophila melanogaster*, which exhibited extreme reduction in eye size (Bridges, 1936). In many eukaryotes a large portion of the genome is considered to be duplicated material. For instance, in human around 40% of the genes were identified to be duplicates (Zhang, 2003). Gene duplications segregate in high numbers in natural populations, and some cause disease (Singleton et al., 2003) or confer an adaptive advantage (Perry et al., 2007).

A duplication initially arises in a single individual and may be lost by drift or be propagated in the following generations, just like a nucleotide mutation. Suppose that a new duplicated gene pair (A-A) arises in a diploid population of size N, in which all individuals initially have single copies of gene A. If the new arising copy has neither selective advantage nor disadvantage, the A-A type will be fixed in the population with probability $1/2N$ and the fixation process takes on average $4N$ generations. During that time, the duplicated genes can accumulate mutations independently, potentially leading to the emergence of new functions or the refinement of existing ones, possibly resulting in adaptation and a selective benefit. There are several theories on the selection scheme of duplicated genes (see for example Innan and Kondrashov (2010) and Magadum et al. (2013) as overview) and we are going to present three models here, which are *positive dosage*, *neofunctionalization* and *diversifying selection*, illustrated in Figure 2.11.

1. The idea of a *beneficial dosage* is straightforward: If an increase of dosage of a particular gene is beneficial, then a duplication of this gene may be fixed with positive selection. In a variable environment, selection for increased dosage may be followed by selection against it, leading to a cycle of gene duplication and loss. Taken further, this implies an optimal copy number that provides the ideal dosage. This concept may be applicable to three categories of genes:

    - Genes mediating interactions between the organism and the environment, including stress response genes, sensory genes, transport genes and those involved in metabolic functions (Kondrashov et al., 2002).
    - Genes with dosage-sensitive functions due to their products' properties regarding protein-protein interactions or their role in metabolic pathways (Kondrashov and Koonin, 2004; Veitia, 2005).
    - Genes with products generally required in large quantities, such as ribosomal or histone genes (Sugino and Innan, 2006).

2. Ohno (1970) argued that the presence of a single gene copy is sufficient to carry out the gene's function. If an extra, redundant copy becomes fixed within the population due to genetic drift, the original copy will maintain its ancestral role, while the new copy may acquire a *new function*. However, since the majority of newly arisen mutations are likely to be detrimental, pseudogenization is the most probable fate for the newly formed gene copy.

3. In cases where natural selection favours *genetic diversity*, gene duplications are beneficial as they offer a larger platform for genetic mutation and selection. As an example, consider the major histocompatibility (MHC) genes, which are subject to overdominant selection and heterozygous individuals reach the maximum fitness value. Therefore, the gene under selection accumulates several alleles with distinct functions.

In general, there are four molecular mechanisms of gene duplication: unequal crossing over, retrotransposition, duplicated DNA transposition and polyploidization (Magadum et al., 2013). Unequal crossing over occurs during meiosis when homologous chromosomes misalign, leading to the duplication or deletion of genetic material and often creates tandem repeats or gene families in the genome. Retrotransposition involves the reverse transcription of an RNA molecule into a DNA sequence that is then inserted back into the genome. Duplicated

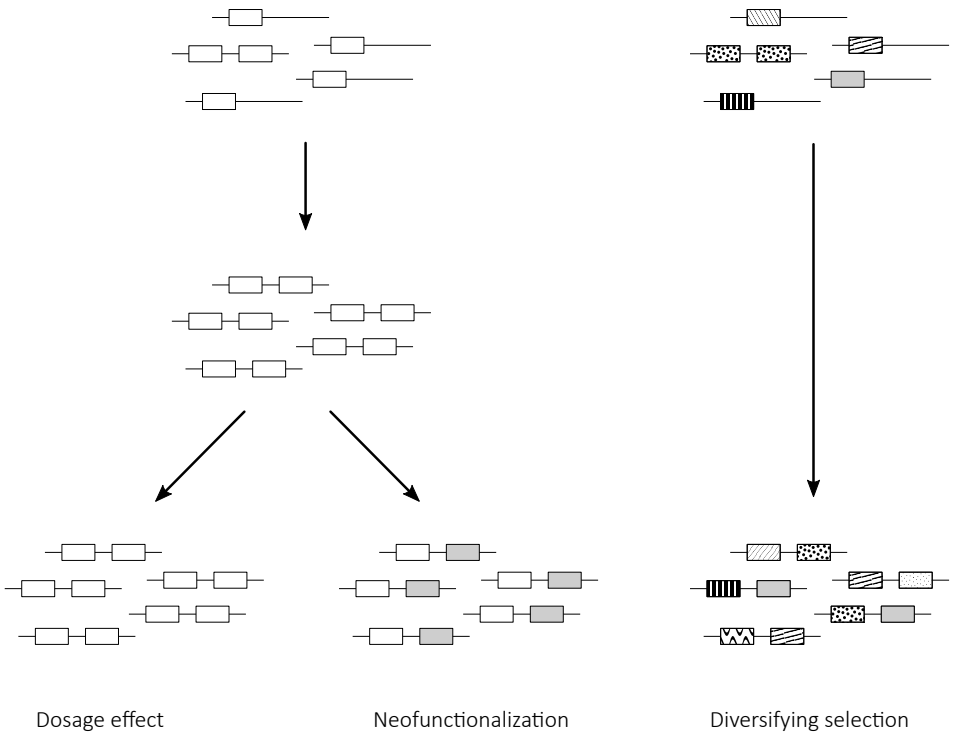


Figure 2.11: Illustration of dosage effect, neofunctionalization and diversifying selection. The different colours and shapes denote different functions of the genes.

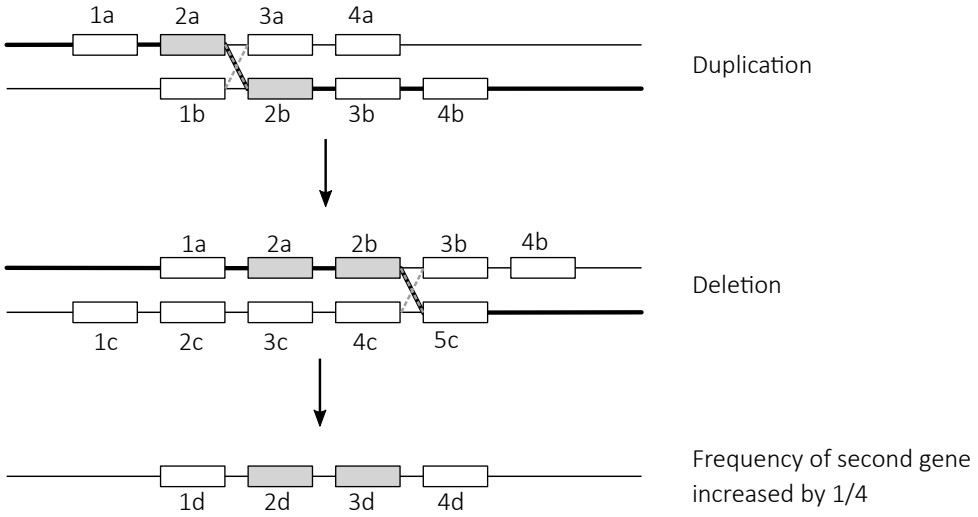


Figure 2.12: Ohta's cycle model. Unequal recombination happens by shifts of one unit. Duplication and deletion phases are alternating to keep the copy number constant. Here, the frequency of the lineage of the second gene increases by 1/4.

DNA transposition involves the movement of DNA segments from one genomic location to another, resulting in the duplication of genetic material. This process is typically mediated by transposons, which are mobile genetic elements that can cut and paste themselves into different regions of the genome. When transposons carry host genes during this process, it can lead to gene duplication. Polyploidization is a whole-genome duplication event in which an organism ends up with multiple copies of its entire genome. This process can occur through various mechanisms, such as errors during meiosis or hybridization between closely related species.

In the following, we focus on the process of unequal recombination. There are two statistics to explore: the genetic variation of the gene copies and the variation in copy number. We shortly present two models that adress these questions. A model of genetic variation in large gene families is given by the *cycle model* of Ohta (1976), in which she uses Kimura's diffusion approach for nucleotide mutations to derive the fixation of gene lineages. Consider a gene family of n gene units and that initially each unit is represented by a different gene lineage. Unequal recombination is modeled by shifts of exactly one unit, such that one copy is either added or lost. These duplication and deletion processes occur alternately, keeping the total number constant (see Figure 2.12). Denote the frequency of a particular lineage as x. Then, in one cycle it may either increase by $1/n$, decrease by $1/n$ or remain constant. The mean $M_{\Delta x}$ and variance of frequency change $V_{\Delta x}$ per time unit (i.e. per cycle) are then given by

$$\begin{aligned} M_{\Delta x} &= \frac{1}{n}x(1-x) - \frac{1}{n}x(1-x) = 0 \\ V_{\Delta x} &= \frac{1}{n^2}x(1-x) + \frac{1}{n^2}x(1-x) = \frac{2}{n^2}x(1-x). \end{aligned} \tag{2.35}$$

Therefore, it is possible to apply Kimura's diffusion approach. With $M_{\Delta x} = 0$, this setting corresponds to the allele frequency of a neutral variant. The drift parameter in the neutral diffusion process of a diploid population is $x(1-x) \cdot 1/(2N_e)$, see equation (2.7), which means that the genealogy of n gene units under unequal recombination with shifts of one unit as described in equation (2.35) can be seen as an analogy of a neutrally evolving mutation in a population of size $N_e = n^2/4$. Therefore, the time to fixation of one lineage is given by

$$t_1(p) = -\frac{1}{p}\left(n^2(1-p)\log(1-p)\right) \approx n^2,$$

where p denotes the initial frequency of the gene lineage (which is $1/n$) and the approximation holds true for large n or small p.

A general model to study the evolution of copy number variation in a population was introduced by Takahata (1981). Let r_{jk} denote the rate of unequal crossing over of two chromosomes with j and k repeated genes and $p_{i,j+k}$ the probability that a chromosome with i copies is generated, if a j- and k-chromosome recombine. Then, without drift, the frequency f_i of chromosomes with i copies evolves according to

$$f_i(t+1) = \sum_{j=0}^{\infty}(1-r_{ij})f_{ij}(t) + \sum_{j=0}^{\infty}\sum_{k=0}^{\infty} p_{i,j+k} \cdot 1_{\{0 \le i \le j+k\}} r_{jk} f_{jk}. \tag{2.36}$$

This concept can be extended with selection towards an optimal copy number and even with sister chromatid exchange. Under this process the distribution of gene copy numbers converges to an equilibrium as time proceeds. However, Takahata (1981) could only use computer simulations to derive the mean and variance of the limiting distribution for different parameters and recombination schemes.

2.9 Motivation of a new model

In the following we combine the concepts and models mentioned in this chapter. We introduce a biologically motivated recombination model of copy number changes based on the process of Takahata (1981) equipped with diversifying selection in an infinite alleles model. Under neutrality and in the absence of drift, we derive the equilibrium distribution of gene copy numbers. Using Wright-Fisher simulations we analyze the effect of selection strength, recombination rate and demographic parameters on the population fitness. For given empirical copy number data from the 1,000 Genomes Project, we can estimate recombination and selection parameters. When comparing copy number distributions of African, Asian and European populations, we answer the question, whether differences can be explained purely by human demography and the out of Africa expansion, or whether shifts in the distribution are signatures of adaptation. Finally, we introduce a new interpretation of the structured coalescent to estimate genetic variation in multi-copy gene families. Instead of individuals migrating in an island model, we consider gene copies to change their position within a gene array according to unequal recombination. Hence, two copies can only be offspring of the same ancestor, if located at the same genetic position. This shows, that standard test-statistics as Tajima's D may lead to misinterpretations when analyzing sequencing data of gene families.

3 Recombination, selection and the evolution of tandem gene arrays

Authors: Moritz Otto, Yichen Zheng and Thomas Wiehe
Status: Published in Genetics, July 2022

DOI: https://doi.org/10.1093/genetics/iyac052

Abstract

Multi-gene families – immunity genes or sensory receptors, for instance – are often subject to diversifying selection. Allelic diversity may be favoured not only through balancing or frequency dependent selection at individual loci, but also by associating different alleles in multi copy gene families. Using a combination of analytical calculations and simulations, we explored a population genetic model of epistatic selection and unequal recombination, where a trade-off exists between the benefit of allelic diversity and the cost of copy abundance. Starting from the neutral case, where we showed that gene copy number is Gamma-distributed at equilibrium, we derived also mean and shape of the limiting distribution under selection. Considering a more general model which includes variable population size and population substructure, we explored by simulations mean fitness and some summary statistics of the copy number distribution. We determined the relative effects of selection, recombination and demographic parameters in maintaining allelic diversity and shaping mean fitness of a population. One way to control the variance of copy number is by lowering the rate of unequal recombination. Indeed, when encoding recombination by a rate modifier locus, we observe exactly this prediction. Finally, we analyzed the empirical copy number distribution of three genes in human and estimated recombination and selection parameters of our model.

3.1 Introduction

Multi-gene families occur in most, if not all, genomes of eukaryotes – in metazoans as well as in plants. They may be conserved across large evolutionary distances, such as the histones or tRNA gene families, or rapidly diversify in single species, such as the NLR-genes in *Danio rerio* (Howe et al., 2016) or the LRR-genes in *Arabidopsis thaliana* (de Weyer et al., 2019).

Interspecies comparison of gene families derived from whole genome duplication has been used, for instance, to estimate relative rates of gene loss and functional divegence (Nadeau and Sankoff, 1997). On a shorter time scale, segmental duplication and unequal recombination are perhaps the more important mechanisms to explain gene family size differences between species, populations and individuals. Modeling gene family evolution has a quite long history (Smith, 1974; Innan, 2009; Demuth and Hahn, 2009; Liu et al., 2011). The roadmap in a population genetic framework was laid out in a series of contributions by Ohta (1976, 1979, 1984, 1987, 1988, 2000). These models typically include forces such as selection and unequal recombination or gene conversion. To describe the dynamics of copy number variation (CNV) generated by unequal recombination Takahata (1981) introduced a general model based on the work of Krüger and Vogel (1975). Fostered especially by the human genome diversity projects, leading to the realization that structural variation is more than abundant in human populations and observing genome size differences between individuals of 100Mb and more (Tuzun et al., 2005; Redon et al., 2006; Eichler, 2008), we are witnessing revived interest in modeling and analyzing the evolution of gene families and of the forces and mechanisms driving copy number polymorphisms.

Tandem gene duplication may happen due to some form of replication error, mis-pairing or segregation anomaly, notably unequal or – less frequently – non-homologous recombination (Silver, 2001). A duplicated gene initially arises in a single individual, very much like a base mutation, and may be lost by drift or be propagated to the offspring in subsequent generations. On its way to fixation, or loss, such a duplication manifests itself as copy number variation (CNV) in a given population and – given sufficiently large populations – is sensed by the filter of natural selection. When beneficial, directional selection will accelerate its fixation and subsequent purifying selection will prevent it from loss. Alternatively, when beneficial only in conjunction with other alleles or other copies, balancing selection may force it to remain at intermediate frequency. The best known examples are perhaps the alleles of pathogen receptors and immune genes, such as those of the MHC complex in vertebrates. Balancing selection, however, comes with a fitness cost in terms of segregation load.

Haldane (1937) had suggested that this effect may be alleviated or avoided when over-dominant alleles are arrayed in tandem on the same chromosome rather than be combined on homologous chromosomes. Only recently, this fundamental idea has been experimentally tested – and confirmed – in populations of the mosquito *Culex pipiens* (Milesi et al., 2017).

Here, we designed a model of tandemly arrayed genes whose evolution is driven by unequal recombination together with a mixture of diversifying and negative selection. More precisely, negative selection will keep copy number in check, while allelic diversity is positively selected.

We implement this via a product of two multiplicative fitness components: one of them is decreasing with copy number and the other one is increasing with allele number (see equation (3.1)). In its structure this fitness function is an old acquaintant. Very similar versions feature in the classical model of Muller's ratchet (Haigh, 1978) and its epistatic relatives (Kondrashov, 1982; Chao, 1988).

We discovered the following: first, in the absence of selection, i.e. when diversity of alleles does not confer any fitness benefit and additional copies do not provide any cost, the distribution of copy numbers can be analytically expressed. It is a Gamma distribution with shape $\alpha = 4$ and with a scale which depends only on the mean copy number of the initial distribution. With selection, the limiting distribution is still well approximated by a Gamma distribution, but depends on the combination of selection coefficients and recombination rate, and *not* on the initial distribution. Second, population size can have a stronger effect on mean fitness and allelic diversity than the strength of selection itself. Third, low recombination rates may be favourable to maintain allelic diversity. Consistent with this, when recombination rates are coded as alleles at a modifier locus and are allowed to evolve over time, we observe a tendency towards recombination rate reduction.

Taken together, our model captures essential aspects of a multi gene family driven by a force of increasing allelic diversity and, at the same time, an opposing force of maintaining genome and chromosome integrity and of limiting both segregation and recombination loads.

Based on the empirical copy number distribution in a set of three exemplary gene families in human we estimated the strengths of selection and (unequal) recombination rates in a natural population.

3.2 Methods

3.2.1 Model

We consider a *compound* model in which the number of copies (y) of a certain gene per individual, as well as the number of alleles (x), are variable. When alleles are all considered distinct (but without labeling their identities) and copy numbers remain variable, we call this the *y-only model*.

In a diploid population of effective size $N \leq \infty$ let individual i, $1 \leq i \leq N$, carry $y_i = y_i^{m\prime} + y_i^{\prime p}$ copies of a particular gene on its maternal (m) and paternal (p) chromosomes. We use the notation y' for the number of copies per chromosome when neither the individual nor the parental status matter. If copies are distinguishable, we call them *alleles* and let x', $1 \leq x' \leq y'$, be the number of different alleles on a chromosome with y' copies. By extension, individual i has $x_i \leq x_i^{m\prime} + x_i^{\prime p}$ alleles (Figure 3.1C, alleles indicated by different colours). Fitness ω_i of individual i is determined by both copy and allele numbers: $\omega_i = \omega_i(x_i, y_i)$. We assume that increasing the number of copies incurs a fitness *cost*, representing adverse effects to genomic structure and integrity, while increasing the number of alleles incurs a fitness *benefit*, representing improved function such as recognition of a wider range of pathogens or stimuli. To fix ideas, we consider the following fitness function

$$\omega_i = \omega(x_i, y_i) = (1 + s_x)^{(\sum_{k=0}^{x_i - 1} \beta_x^k)} \times (1 - s_y)^{(\sum_{k=0}^{y_i - 3} \beta_y^k)} . \tag{3.1}$$

The cost is only counted from the third copy, since the ground state is a single copy gene with exactly one copy on each chromosome. The selection coefficients $0 < s_x, s_y \ll 1$ are positive and the epistasis parameters $0 < \beta_x \leq 1 \leq \beta_y$ are independent of i. In the following, we omit index i unless required for clarity. The way we define epistasis reflects the classical concepts of diminishing returns (β_x) and synergistic epistasis (β_y): the benefit of adding new alleles decreases with the number of already existing alleles. Think of the physiological limit preventing perfect recognition of an infinite number of possible pathogens or sensory stimuli in nature. In contrast, the cost of adding more copies increases with the number of already existing copies. This reflects the growing threat to genome integrity by inserting more and more copies.

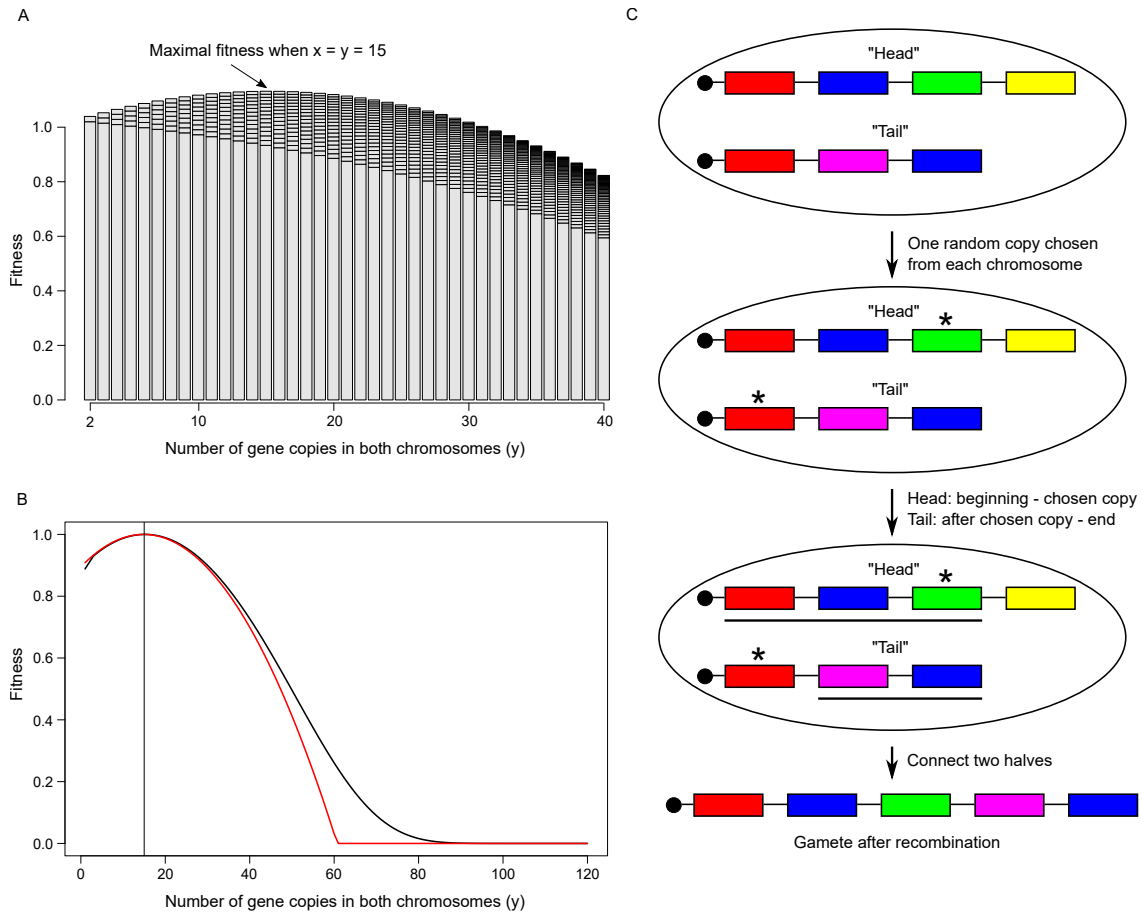


Figure 3.1: **A.** Fitness of an individual as a function of x (stacks) and y (bars). Parameters: $s_x = 0.02$, $s_y = 0.005$, $\beta_x = 0.95$, $\beta_y = 1.05$. Each bar represents one value of y with stacked fitness "layers" for $x = 1$ to $x = y$. **B.** Normalised fitness of an individual in the y-only model. Parameters: $s_x = 0.02$, $s_y = 0.005$, $\varepsilon = 0.05$ (black) and its Taylor-approximated version $\mathcal{T}(y) = 1 - \tilde{s}(y - y^*)^2$, with $\tilde{s} \approx 0.00047$ (red). The vertical line marks $y^* \approx 14.86$. **C.** Illustration of individual genotype unequal recombination. Recombination occurs in an individual with $y = 7 = 4 + 3$ gene copies and $x = 5 < 4 + 3$ different alleles (colours). The black bullet on each chromosome represents the RRM locus (see text).

For any fixed copy number y, fitness is maximized when $x = y$, i.e., when every copy is a different allele (which is an assumption in the y-only model). Whether fitness is maximized for small or for large y depends on the relative magnitudes of s_x and s_y: assuming $x = y$ and $s_x \leq s_y$, maximum fitness is achieved at the lowest possible copy number, $y = 2$. Arguably, this situation represents the standard scenario for single copy genes in nature: the cost of adding copies would outweigh its benefit. In contrast, when $s_x > s_y$, maximum fitness may be attained at values $y > 2$. Without epistasis, and as a function of y, fitness is monotonically increasing, with lowest fitness at $y = 2$. With epistasis, fitness has a non-trivial maximum at y^* (Figure 3.1A). In this case, we have (see Appendix)

$$y^* = \frac{1}{\ln(\frac{\beta_y}{\beta_x})}\left(2\ln(\beta_y) + \ln\left(\frac{\beta_y - 1}{1 - \beta_x}\right) + \ln\left(-\frac{\ln(1+s_x)}{\ln(1-s_y)}\right) + \ln\left(-\frac{\ln(\beta_x)}{\ln(\beta_y)}\right)\right). \tag{3.2}$$

Assuming further $\beta_x = 1 - \varepsilon$ and $\beta_y = 1 + \varepsilon$ for small $\varepsilon > 0$, and using $\ln(1 + \varepsilon) \approx \varepsilon$, y^* simplifies to

$$y^* \approx 1 + \frac{\ln(s_x) - \ln(s_y)}{2\varepsilon} = 1 + \frac{\ln(\frac{s_x}{s_y})}{2\varepsilon}\,. \tag{3.3}$$

In finite populations alleles are lost by drift. Although new alleles are introduced by mutation, one generally has $x < y$ at mutation-drift equilibrium. We employ an infinite alleles model: mutation occurs with rate μ per copy per individual per generation and turns a given allele into a new, previously non-existing one. The more copies an individual has, the more likely a new allele will be generated. Note that mutation does not change y or y', but it may increase x and x'. The y-only model can be interpreted as the limiting scenario for large mutation rates such that any two copies are different. Therefore, mutation is explicitly required only in the simulations of the compound model, but not for the analytical results of the y-only model.

In both the compound and the y-only models recombination may be non-homologous, or *unequal*. As a consequence, copy number may change across generations. It is implemented as follows (Figure 3.1C): first, choose a pair of chromosomes and decide whether recombination occurs (probability r) or not $(1 - r)$. In the first case, randomly mark a gene copy on both chromosomes. Then, the "upstream" fragment *in*cluding the marked copy of chromosome m ("head"), say, is fused with the "downstream" fragment *ex*cluding the marked copy of chromosome p ("tail"). For simplicity we assume recombination break points to lie outside of genes and exclude the possibility that genes may be disrupted by recombination. Only one recombination product is considered further. If the last copy was marked on the tail chromosome, no copy is added to the head fragment. Starting from two chromosomes with $y^{m\prime}$ and y'^{p} copies, copy number in the offspring gamete can range between 1 and $y^{m\prime} + y'^{p} - 1$. More precisely, copy number in the offspring chromosome is a sum of uniform random variables

with

$$Y' = B_1 + B_2 - 1\,,$$

where $B_1 \sim \mathrm{U}(y^{m\prime})$, $B_2 \sim \mathrm{U}(y'^p)$ are uniform on the integers $\{1, ..., y^{m\prime}\}$ and $\{1, ..., y'^p\}$, respectively. The sum Y' is trapezoidal with

$$P[Y' = y' \mid y^{m\prime}, y'^p] = T(y', y^{m\prime}, y'^p)$$
$$= \frac{1}{y^{m\prime} \cdot y'^p} \begin{cases} 0, & y' \leq 0 \\ y', & 1 \leq y' \leq (y^{m\prime} \wedge y'^p) \\ (y^{m\prime} \wedge y'^p), & (y^{m\prime} \wedge y'^p) \leq y' \leq (y^{m\prime} \vee y'^p) \\ y^{m\prime} + y'^p - y', & (y^{m\prime} \vee y'^p) \leq y' \leq y^{m\prime} + y'^p - 1 \\ 0, & y' \geq y^{m\prime} + y'^p \end{cases},$$

where $\wedge$ denotes the minimum and $\vee$ the maximum. When no recombination occurs, only one of the two parental chromosomes is propagated.

We also consider a version with recombination rate variation: assume that each chromosome carries a recombination rate modifier (RRM) locus which encodes a chromosome-specific recombination rate. For a pair of chromosomes m and p, a recombination event occurs with rate $r = r_o\sqrt{(\rho_m\,\rho_p)}$ for modifier "alleles" ρ_m, $\rho_p > 0$ which are multipliers of the base recombination rate r_o. The modifier allele inherited to the recombination product is the geometric mean $\sqrt{\rho_m\,\rho_p}$. Note that selection, operating on the genotype, exerts an indirect force on the recombination rate. Symbolically, the modifier locus is represented by a black bullet in Figure 3.1C. It is itself not subject to recombination, but attached to the first gene copy. We set $r_o = 0.01$ in all simulations.

3.2.2 Simulations

For all simulations we used an in-house developed R programme (https://github.com/y-zheng/Recombination-gene-family) implementing a Wright-Fisher-type model with discrete generations and multinomial sampling of gametes. Simulation raw data can be downloaded from the same repository. Simulations consisted of a burn-in phase and an observation phase in which the statistics shown in Table 3.1 were recorded at certain time intervals. We considered four basic scenarios:

(a) single population with constant size N;

(b) single population with bottleneck;

(c) two sub-populations with reciprocal migration;

(d) single population of constant size with RRM.

Simulations for scenario (a) were started with $y = 10$ and $x = 1$ for all i and run for an initial burn-in phase of $20,000$ generations. A run was re-started in case it entered the (absorbing) state $y = 2$ during burn-in, i.e. when all individuals have only a single copy on each chromosome. To start simulations in scenarios (b)-(d), we used the final state which was reached at the end of scenario (a). To reduce standard error of the mean of this final sampling point, we we ran 500 replicates for scenario (a) and 200 replicates for scenarios (b)-(d). For the simulations we selected parameter ranges which we considered realistic and which turned out to be compatible with the estimates for s_x, s_y and r and the mean copy number obtained from empirical data (see below). The parameters used in the different scenarios are listed in Table 3.3 in the Appendix.

Table 3.1: Summary statistics recorded in simulations.

Individual statistics				
	Mean[1]	Std. Dev.	Min.	Max.
Copies	$\bar{\mathbf{y}} = (\sum_i y_i)/N_e$	$\sigma_{\mathbf{y}}$	$\min_{\mathbf{y}}$	$\max_{\mathbf{y}}$
Alleles	$\bar{\mathbf{x}} = (\sum_i x_i)/N_e$	$\sigma_{\mathbf{x}}$	$\min_{\mathbf{x}}$	$\max_{\mathbf{x}}$
Ratio	$\overline{\mathbf{x}/\mathbf{y}} = (\sum_i \frac{x_i}{y_i})/N_e$	$\sigma_{\mathbf{x}/\mathbf{y}}$	$\min_{\mathbf{x}/\mathbf{y}}$	$\max_{\mathbf{x}/\mathbf{y}}$
Fitness	$\bar{\boldsymbol{\omega}} = (\sum_i \omega_i)/N_e$	$\sigma_{\boldsymbol{\omega}}$	$\min_{\boldsymbol{\omega}}$	$\max_{\boldsymbol{\omega}}$

Population statistics	
Total number of copies in population[b]	$\lvert\mathbf{y}\rvert$
Total number of *different* alleles[2]	$\lvert\mathbf{x}\rvert$
Absolute frequency of alleles[3]	$m_j,\ j = 1, ..., \lvert\mathbf{x}\rvert$
Relative frequency of alleles	$\xi_j = \frac{m_j}{2N_e},\ j = 1, ..., \lvert\mathbf{x}\rvert$
Effective number of alleles[4]	$\lvert\mathbf{x}\rvert_{\text{eff}} = \left(\sum_{j=1}^{\lvert\mathbf{x}\rvert} \left(\frac{m_j}{\lvert\mathbf{y}\rvert}\right)^2\right)^{-1}$

3.2.3 Empirical data

Based on data from the 1,000 Genomes project, Brahmachary et al. (2014) analyzed copy number variation in 193 gene families in three human populations (CEU, CHB, YRI). We chose three candidates (`PSG3`, `MUC12` and `PRR20A`) which satisfied the following criteria:

- genes tandemly arrayed
- genes autosomal
- mean copy number between 10 and 20
- one example each with small, intermediate and large copy number variance.

`PSG3` (Pregnancy specific glycoprotein 3) is located on the long arm of the particularly gene rich chromosome 19 (Grimwood et al., 2004). It is a member of the carcinoembryonic antigen gene family and of the immunoglobulin superfamily and is involved in pregnancy maintenance. `MUC12` (Mucin 12) is a membrane glycoprotein of the mucin family. Mucins are involved in mucous protection, epithelial cell differentiation and intracellular signalling and have been recognized having similar evolutionary features as HLA genes (Vahdati and Wagner, 2016). `PRR20A` (Proline-rich protein 20A) is a predicted gene located on the long arm of chromosome 13. It has low Uniprot annotation score with experimental evidence only at transcript level[1].

The available empirical data from this data set can be analyzed in the context of the y-only model. To estimate the underlying parameters (s_x, s_y, r) of the y-only model that best describe the empirical copy number distribution we implemented an *EM*-like grid search as follows: we use the data from the African (YRI) population, assuming that it is closest to recombination-selection-drift equilibrium and least affected by a recent population bottleneck (e.g., Schiffels and Durbin (2014); Rafajlović et al. (2014); Spence and Song (2019)). Individual copy numbers are derived from the data published by Brahmachary et al. (2014) and calculated by dividing the individual read ("nanostring", in the authors' terminology) counts by the average read count per copy[2]. This way, we found for `MUC12`, `PSG3` and `PRR20A` mean numbers of, respectively, $11.85, 14.94, 19.85$ copies per individual in the YRI population (diploid sample size $n = 60$). To compare these results with our model, we uniformly sampled $5,000$ parameter combinations of independently chosen s_x, s_y and r from the product of initial intervals $[1e{-}6, 5e{-}2]^3$. For each parameter combination we calculate the Gamma-approximation of the equilibrium distribution of the y-only model (see Results) and use the Kolmogorov-Smirnov (KS) test to calculate the probability that the data are sampled from this distribution. We choose the top 100 (= 2%) parameter combinations to define the range of the new parameter intervals to sample from. In each iteration parameter intervals are shrinking and we terminate this process after 10 iterations to obtain a possibly small range of the final parameter combinations with highest KS p-value. We then chose the best parameter combinations for further analysis.

[1] `https://www.uniprot.org/uniprot/P86496`
[2] `https://github.com/y-zheng/Recombination-gene-family`

3.3 Results

3.3.1 y-only model

Consider first the y-only model. Each copy is considered a unique and distinct allele. Therefore, at any time, $x_i = y_i \ \forall i$, and fitness of an individual is a function only of y:

$$\omega = \omega(y_i) = (1+s_x)^{(\sum_{k=0}^{y_i-1} \beta_x^k)} \times (1-s_y)^{(\sum_{k=0}^{y_i-3} \beta_y^k)}\,.$$

for all individuals i.

Let y' be the number of gene copies on a single chromosome, without regard of parental status, and let $p_t(y')$ be the frequency of chromosomes with y' copies in an infinitely large population in generation t.

Choosing parental chromosomes according to their fitness $\omega(y = y^{m\prime} + y'^p)$, the frequency of y' changes to

$$p_{t+1}(y') = (1-r)\sum_{y'^p} q_t(y', y'^p) + r \sum_{y^{m\prime}, y'^p} q_t(y^{m\prime}, y'^p) T(y', y^{m\prime}, y'^p), \tag{3.4}$$

where T denotes the trapezoidal distribution and

$$q_t(y^{m\prime}, y'^p) = \frac{p_t(y^{m\prime}) p_t(y'^p) \cdot \omega(y^{m\prime} + y'^p)}{\bar{\omega}_t}$$

is the frequency of the pair $(y^{m\prime}, y'^p)$ after selection. In the last equation $\bar{\omega}_t$ is mean population fitness at time t, i.e.

$$\bar{\omega}_t = \sum_{y^{m\prime}, y'^p} p_t(y^{m\prime}) p_t(y'^p) \cdot \omega(y^{m\prime} + y'^p),$$

where the sum runs over all possible pairs $(y^{m\prime}, y'^p) \in \mathbb{N} \times \mathbb{N}$. Therefore, this process can be thought of as an irreducible aperiodic Markov chain on the state space $\{1, 2, ...\}$, which converges to its unique stationary distribution. Under neutrality ($\omega \equiv 1$), this simplifies to

Proposition 3.3.1. *Under (unequal) recombination and under neutrality it holds that*

- *the expected value of copy number remains constant over time, i.e.* $\forall t$

$$\sum_{y'=1}^{\infty} y' \cdot p_{t+1}(y') = \sum_{y'=1}^{\infty} y' \cdot p_t(y') = ... = \sum_{y'=1}^{\infty} y' \cdot p_0(y') =: E_{Y'}$$

- *the stationary distribution is given by the discrete kernel of the Gamma-distribution with shape parameter* $\alpha = 2$ *and expected value* $E_{Y'}$*, i.e.*

$$p_{\text{stat}}(y') = y' \cdot \exp\left\{-\frac{2}{E_{Y'}} y'\right\} \cdot \frac{1}{Z}\,, \tag{3.5}$$

where Z is the normalisation constant given by

$$Z = \sum_{y'} y' \cdot \exp\left\{-\frac{2}{E_{Y'}} y'\right\} = \frac{\exp\left\{\frac{2}{E_{Y'}}\right\}}{\left(\exp\left\{\frac{2}{E_{Y'}}\right\} - 1\right)^2}.$$

The proof is given in the Appendix.
Hence, the neutral equilibrium distribution of copy numbers on individuals is given by the convolution

$$\begin{aligned}\tilde{p}_{stat}(y) &= \sum_{y_1'+y_2'=y} p_{stat}(y_1') p_{stat}(y_2') \\ &= \frac{1}{6}(y^3 - y) \exp\left\{-\frac{1}{E_Y} y\right\} \cdot \frac{1}{Z^2},\end{aligned}$$

which is the discrete kernel of the Gamma-distribution with shape parameter $\alpha = 4$ and expected value $E_Y = 2E_{Y'}$.

Adding selection to the process makes the analysis less straightforward. We note that the process described by equation (3.4) is still an irreducible Markov chain which has a stationary distribution. However, determining a closed formula of p_{stat} is not easily feasible and we resorted to the following approximation.

We choose ω as defined in equation (3.1), assume that $|\mathbf{x}| = |\mathbf{y}|$ (y-only model) and that $\beta_x = 1 - \varepsilon$ and $\beta_y = 1 + \varepsilon$ for some $\varepsilon > 0$. Thus, the fitness function simplifies to

$$\begin{aligned}&\omega(y) = \exp\{f(y)\}, \text{ where} \\ &f(y) = \frac{s_x + s_y}{\varepsilon} - \frac{s_x}{\varepsilon} \cdot e^{-\varepsilon y} - \frac{s_y}{\varepsilon} \cdot e^{\varepsilon(y-2)}.\end{aligned} \tag{3.6}$$

The Taylor expansion up to order 2, evaluated at y^* and scaled with $\omega(y^*)^{-1}$ is

$$\begin{aligned}\mathcal{T}(f(y)) &= \frac{1}{\omega(y^*)}\left(\omega(y^*) + \frac{\mathrm{d}}{\mathrm{d}y}\omega(y^*)(y - y^*) + \frac{1}{2}\frac{\mathrm{d}^2}{\mathrm{d}y^2}\omega(y^*)(y - y^*)^2\right) \\ &= 1 + \frac{1}{2}\frac{\mathrm{d}^2 f}{\mathrm{d}y^2}(y^*) \cdot (y - y^*)^2 \\ &= 1 - \varepsilon e^{-\varepsilon}\sqrt{s_x s_y} \cdot (y - y^*)^2.\end{aligned}$$

Note, that this coincides with the fitness function introduced by Krüger and Vogel (1975)

$$\tilde{\omega}(y) = 1 - \tilde{s}(y - y^*)^2, \tag{3.7}$$

when substituting

$$\tilde{s} = \varepsilon e^{-\varepsilon}\sqrt{s_x s_y}\,.$$

Hence, the quadratic distance of y from the optimal copy number y^* determines fitness. It fits

well with our definition of synergistic epistasis when y is not too far from y^* (see Figure 3.1B) and yields a threshold $y^o = y^* + 1/\sqrt{\tilde{s}}$ with $\mathcal{T}(f(y)) < 0$ for $y > y^o$.

Therefore, with this quadratic approximation of the fitness function, equation (3.4) becomes a finite system of equations, which can be numerically solved with standard iteration algorithms. Starting with an arbitrary initial distribution we iterate until

$$\|p_{t+1} - p_t\|_{\mathrm{TV}} := \sum_{y'} |p_{t+1}(y') - p_t(y')| < 0.001,$$

where $\|\cdot\|_{\mathrm{TV}}$ denotes the *total variation* and the sum runs from 1 to the maximal y' given by y^o. After convergence, we calculate the copy number distribution on *individuals* as convolution of the copy number distribution on chromosomes.

For fixed parameters the process converges to the same limiting distribution, independently of initial conditions. Varying the recombination rate leads to different limiting distributions: it is close to the neutral stationary distribution when r is large; it is sharply peaked, and centered at y^*, when r is small. The variance is almost vanishing when $r < 0.01\ s_x$. Increasing selection shifts $\bar{y}$ towards y^*. Generally, the stationary distribution is determined by a balance of recombination and selection and the relative magnitudes of r, s_x and s_y. Visual inspection of the limiting distribution for various parameter choices suggests that it is well approximated by a Gamma distribution also in the non-neutral case (see, for instance, the three examples shown in Figure 3.3, lines in blue). We estimate its parameters as follows:

We numerically solved the system of equations (equation (3.4)) for about $50,000$ random parameter combinations. We kept $\varepsilon = 0.05$ constant and chose $r \in [0, 0.01]$, $s_x \in [0, 0.05]$ and s_y such that $s_x/s_y \in [2.5, 18]$, producing an optimal copy number y^* between 10 and 30. Then, we calculated mean and variance of the equilibrium distribution for all parameter combinations. Assuming that the expectation (E_Y) of the limiting Gamma distribution is determined by equation (3.3), we set

$$\hat{E}_Y = y^* = \frac{\ln(s_x) - \ln(s_y)}{2 \cdot 0.05} + 1\,.$$

Assuming $r > 0.01 \cdot s_x$ and that its standard deviation scaled by the mean (σ/E_Y) depends on recombination-selection balance, $\ln(r/s_x)$, we obtain by linear fitting (Figure 3.2A):

$$\hat{\sigma} = y^* \cdot (0.046 \cdot \ln(r/s_x) + 0.26)\,. \tag{3.8}$$

Furthermore, $(E_Y/\sigma)^2$ converges towards the shape parameter ($\alpha = 4$) of the Gamma distribution under neutrality, when selection becomes small or recombination becomes large (Figure 3.2B). Therefore, for given parameters r, s_x, s_y and $\varepsilon = 0.05$ we use the discrete kernel of the Gamma-distribution with shape parameter $\alpha = (y^*/\hat{\sigma})^2$ and expected value y^* as an approximation of the equilibrium distribution of the y-only process with selection. Note that the distribution is uniquely determined by its shape and mean.

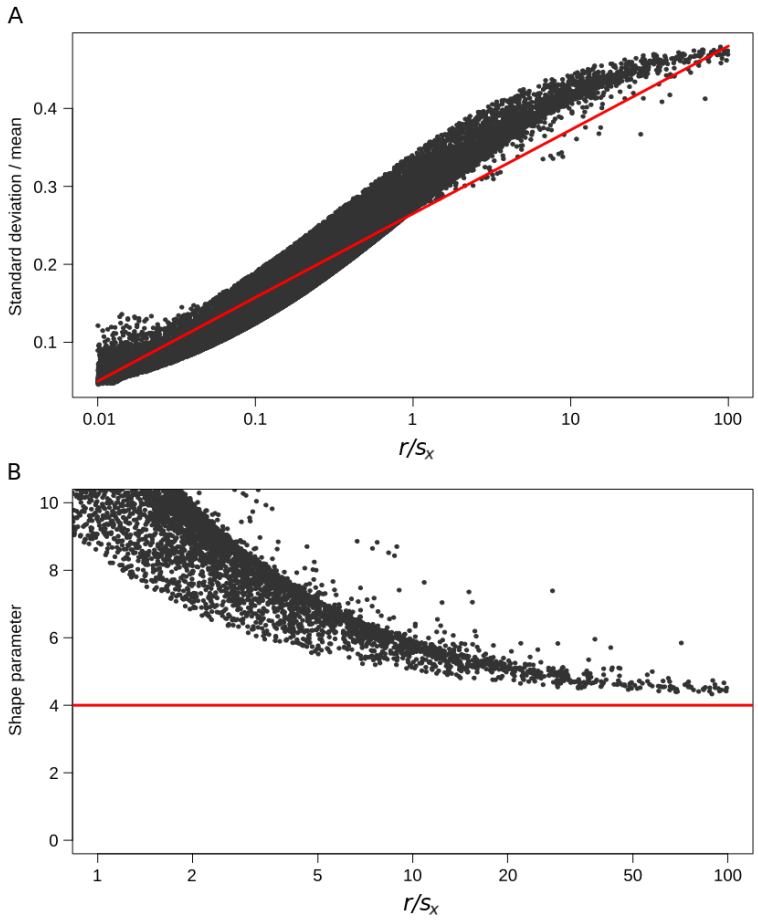


Figure 3.2: **A** Linear fit of σ/E_Y on $\ln(r/s_x)$ (for details see text). Note the strong correlation of $\ln(r/s_x)$ and σ/E_Y, with a Pearson correlation coefficient of $\rho = 0.97$. The estimated regression line $\sigma/y^* = 0.046 \cdot \ln(r/s_x) + 0.26$ is shown in red. **B** Convergence of the Gamma shape parameter $\alpha = (E_Y/\sigma)^2$ towards the value $\alpha = 4$, expected under neutrality, when r is increasing or s_x is decreasing.

Application of the y-only model to empirical data

To estimate selection coefficients and rates of unequal recombination for the three gene families `PSG3`, `MUC12` and `PRR20A` we used the *EM*-like grid search described above. We calculated the KS-test p-value for three distributions: (1) a neutral equilibrium distribution $\tilde{p}_{stat}$ with mean value given by the arithmetic mean of the data, (2) one of the best-fitting Gamma-distributions with parameters given by the *EM*-like grid search and (3) the equilibrium distribution of the y-only process with the same recombination and selection coefficients as obtained from the grid search. Sufficiently small p-values indicate a significant difference from any of the three models, whereas a p-value close to one can be interpreted as a good approximation of the data. The results are given in Table 3.2 and Figure 3.3. Distributions

Table 3.2: Parameter estimates for empirical data obtained by *EM* grid search, with fixed $\varepsilon = 0.05$, that returned the best KS p-value for the Gamma approximation.

Gene family	Estimated parameters	p-value of KS-test		
		Neutral	Gamma	y-only
PSG3	r = 0.001 $s_x = 0.04$ $s_y = 0.01$	1.4e−9	0.99	0.82
MUC12	r = 0.008 $s_x = 0.017$ $s_y = 0.006$	0.0012	0.99	0.98
PRR20A	r = 0.008 $s_x = 0.001$ $s_y = 0.00028$	0.217	0.98	0.98

of the 100 best parameter combinations for each gene are shown in Figure **S1**[1]. For PSG3 the empirical distribution of copy numbers (histogram in Figure 3.3, top) is well approximated by a Gamma distribution (red line) yielding a KS-test p-value of 0.99. The limiting distribution under the y-only model still fits fairly well with $p = 0.82$ (blue line). In contrast, the hypothesis of neutrality can be clearly rejected: the neutral Gamma distribution (equation (3.5)) produces a p-value of 1.4e−9 (black line). The parameter estimates suggest a small recombination rate of about 0.1% per generation per gamete and strong selection ($s_x = 0.04$ and $s_y = 0.01$), maintaining copy number close to its optimal value. Although the gene family PRR20A is much more variable than MUC12 (Figure 3.3, middle and bottom) we estimate the same recombination rate of about 0.8% for both families. However, the difference in their distributions can be explained by different selection strengths. The estimates in MUC12 are $s_x = 0.017$ and $s_y = 0.006$ – about half as strong as in PSG3. In contrast, the estimates in PRR20A are $s_x = 0.001$ and $s_y = 0.00028$, lower by roughly a factor of 40 than in PSG3. While neutrality can still be clearly rejected in MUC12 ($p = 0.0012$), it cannot be rejected in PRR20A. Still, also for this gene family pure neutrality has a much lower explanatory power than do have models with selection ($p = 0.217$ *vs.* $p = 0.98$). One should keep in mind, however, that the above estimates depend on our choice of the epistasis parameter $\varepsilon = 0.05$. From equation (3.3) it is clear that the ratio s_x/s_y and ε are inversely related. In work dedicated to data analysis, rather than model development, one may want to include ε (or even β_x and β_y separately) among the parameters to be estimated.

[1]Supplementary material available at https://doi.org/10.1093/genetics/iyac052.

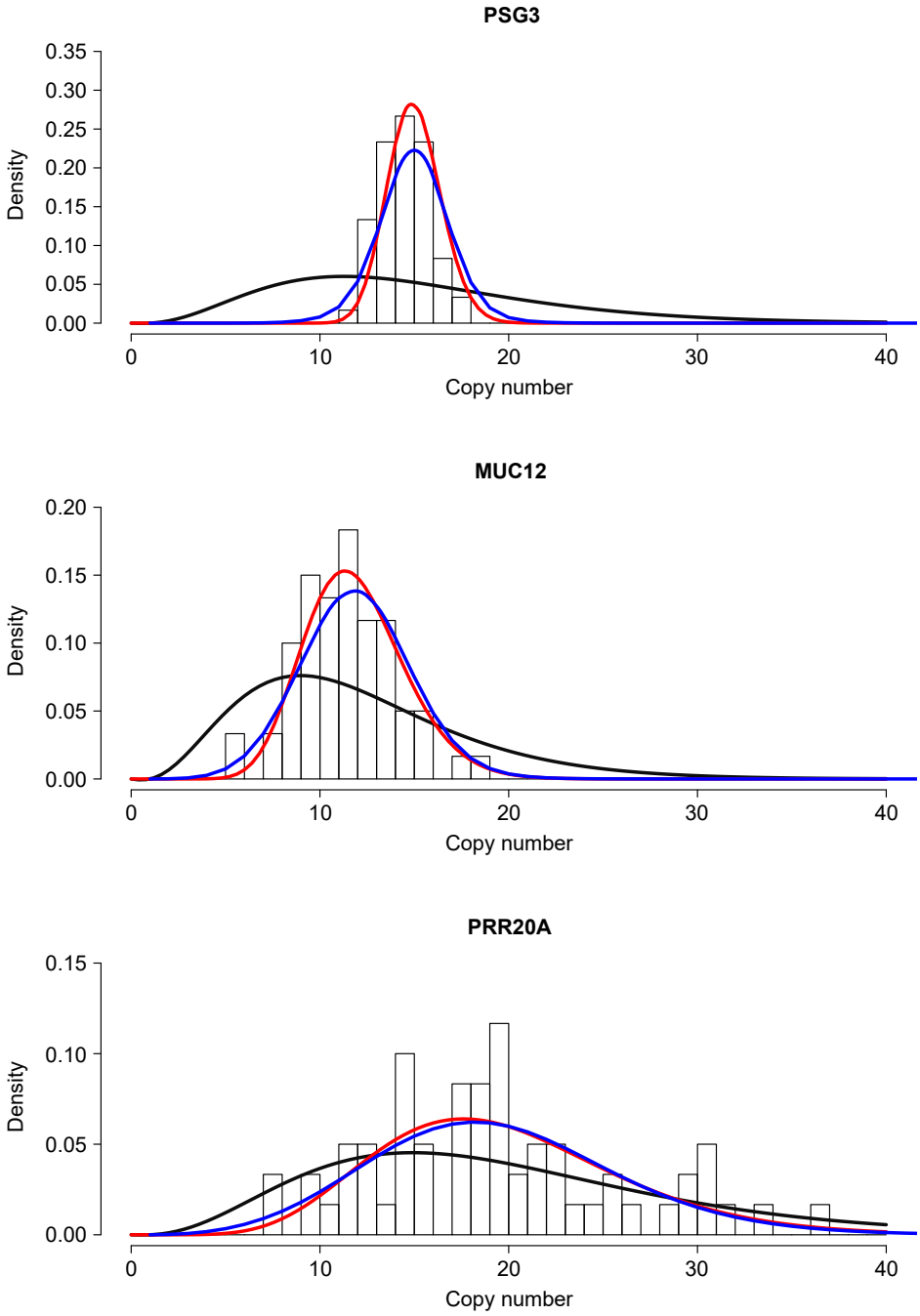


Figure 3.3: Copy number distribution of three different human genes and their approximations. Black: Copy number distribution under neutrality $\tilde{p}_{stat}$ with E_Y =14.94, 11.85 and 19.85 for `PSG3`, `MUC12` and `PRR20A`. Red: Gamma-distribution with parameters given in Table 3.2, resulting in best KS-test p-value. Blue: Equilibrium distribution of the y-only model generated from equation (3.4) with parameters as in Table 3.2.

3.3.2 Simulation results of the compound model

In **scenario (a)** we analyzed the effect of different population sizes, selection strengths **(a1)** and recombination rates **(a2)** on the statistics of Table 3.1 at equilibrium. In **scenario (a1)** we used $s_x = 0.01, 0.02, 0.04$ (weak, medium and strong selection), with $s_x/s_y = 4$ and $\varepsilon = 0.05$. These parameters were chosen such that the optimal genotype for an individual is $x = y = 15$ in all three selection regimes. Population size varied from $N_e = 500$, 1000, 2000 to 4000 and recombination rate was kept constant at $r = 0.01$. Results are shown in Figure 3.4.

Both larger population sizes and stronger selection lead to an increase in population means $\bar{\mathbf{x}}$ and $\bar{\mathbf{y}}$ (Figure 3.4A and B). Note, that the demographic effect (decrease of drift by increase of population size) on these quantities is much stronger than the effect by increasing selection. Both $\bar{\mathbf{x}}$ and $\bar{\mathbf{y}}$ are always below the optimal value of 15. However, doubling N_e has a stronger effect than doubling selection strength in bringing the population closer to the optimal value. Essentially the same pattern is observed for the ratio $\overline{\mathbf{x}/\mathbf{y}}$ (Figure 3.4C). For example, $N_e = 1000, 2000, 4000$ with low selection leads to a higher ratio $\overline{\mathbf{x}/\mathbf{y}}$ than $N_e = 500, 1000, 2000$ with intermediate selection. The total (Figure 3.4E) and the effective (Figure 3.4F) number of alleles scale roughly linearly with N_e. Again, both quantities depend more strongly on population size than on selection strength. This effect is more pronounced in the total number of alleles than in $|\mathbf{x}|_{\text{eff}}$, which is explained by drift: alleles at low frequency, in particular newly generated alleles ($N_e\mu\bar{\mathbf{y}}$ per generation), are prone to loss when drift is strong. They count for the total number, but contribute little to $|\mathbf{x}|_{\text{eff}}$. In contrast, mean fitness is more affected by the strength of selection than by N_e. This is because mean fitness depends on two ingredients: the equilibrium distribution y itself and the weights ω_i of its components. Both are altered by selection. Finally, the frequencies of the most common alleles (Figure **S2**) are negatively correlated both with N_e and s_x. In summary, allelic diversity at population scale appears to be driven mainly by N_e.

In **scenario (a2)** we kept selection at intermediate level ($s_x = 0.02, s_y = 0.005$) and varied the rate of (unequal) recombination from $r = 0.002$ to 0.05. Results are shown in Figure 3.5. Increasing recombination decreases $\bar{\mathbf{x}}$ and $\bar{\mathbf{y}}$, as well as the ratio $\overline{\mathbf{x}/\mathbf{y}}$. Therefore, it also decreases mean fitness $\bar{\omega}$. Recombination acts here in a similar way as drift: doubling the recombination rate has the same effect on fitness as halving the population size. This observation can be interpreted as a recombination load: frequent recombination can generate chromosomes whose copy number is far away from the optimum. Deviation from the optimal copy number has an asymmetric effect because of epistasis: a surplus of copies is more harmful than a deficit (Figure 3.1B), explaining the somewhat counter-intuitive effect that increasing the recombination rate decreases both total and effective number of alleles in the population.

In **scenario (b)** we explored the impact of a single instantaneous and short bottleneck. Starting with an equilibrated panmictic population of constant size $N = 2000$, population size was reduced to 1% (= 20) for 5, 10, or 20 generations, then restored to its original value N and the generation counter reset to $t = 0$. After that, simulations are carried on for

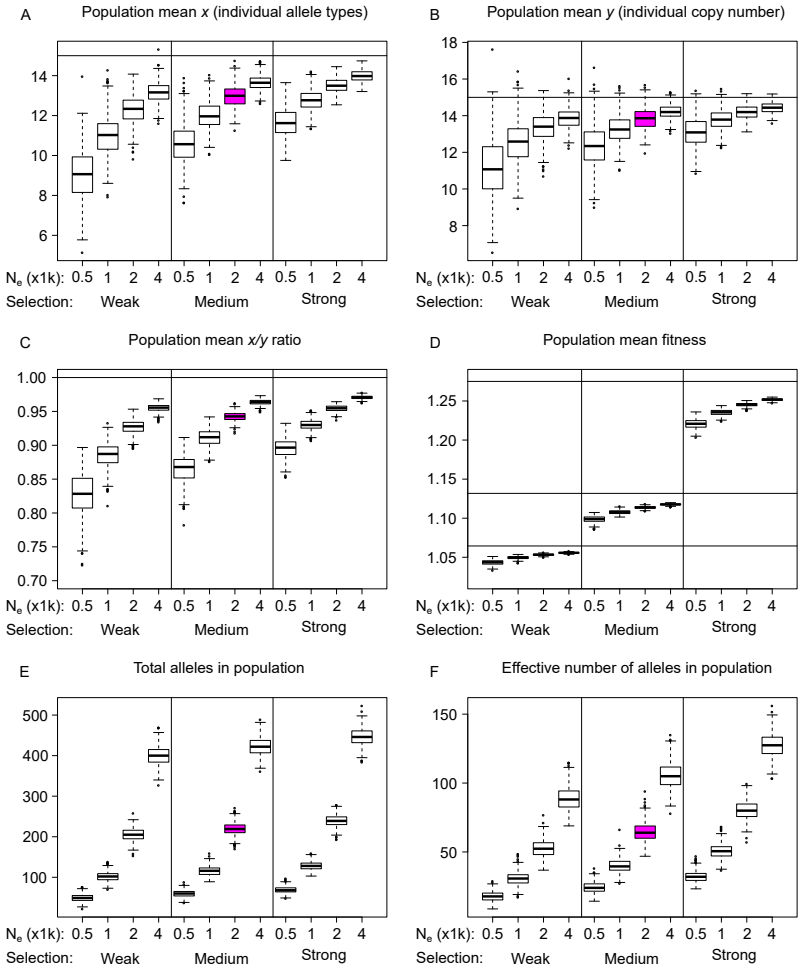


Figure 3.4: **Scenario (a1)** – constant population size. Population statistics at equilibrium: population mean $\bar{\mathbf{x}}$ (A); population mean $\bar{\mathbf{y}}$ (B); $\overline{\mathbf{x}/\mathbf{y}}$ ratio (C); population mean fitness (D); total number (E) and effective number of alleles $|\mathbf{x}|_{\text{eff}}$ (F). Varying parameters: population size N_e and selection coefficient s_x. Mutation ($\mu = 0.0005$) and recombination rate ($r = 0.01$) are kept fixed. Boxplots based on 500 independent replicates. Box coloured in purple indicates a parameter combination ($N_e = 2000$, $r = 0.01$, $s_x = 0.02$, $s_y = 0.005$) shared by scenarios (a), (b), (c) and (d). Horizontal lines in A-C indicate the optimal copy number in the y-only model. Horizontal lines in D indicate optimal fitness.

another $10,000$ generations during which the recovery process of the six summary statistics mentioned above is recorded. Results for different selection strengths are summarized in Figure 3.6. A longer period of population size reduction results in populations with lower $\bar{\mathbf{x}}$ and lower $\bar{\omega}$. In contrast, length of the reduction period hardly affects $\bar{\mathbf{y}}$. Recovery time correlates positively with the length of the reduction period.

We observed that $\bar{\mathbf{y}}$ and, to a lesser extent, $\bar{\mathbf{x}}$ experience a decrease *after* the restoration of population size, and before it returns to its constant equilibrium value. Furthermore, the total number of alleles recovers much faster than the effective number. The reason is that

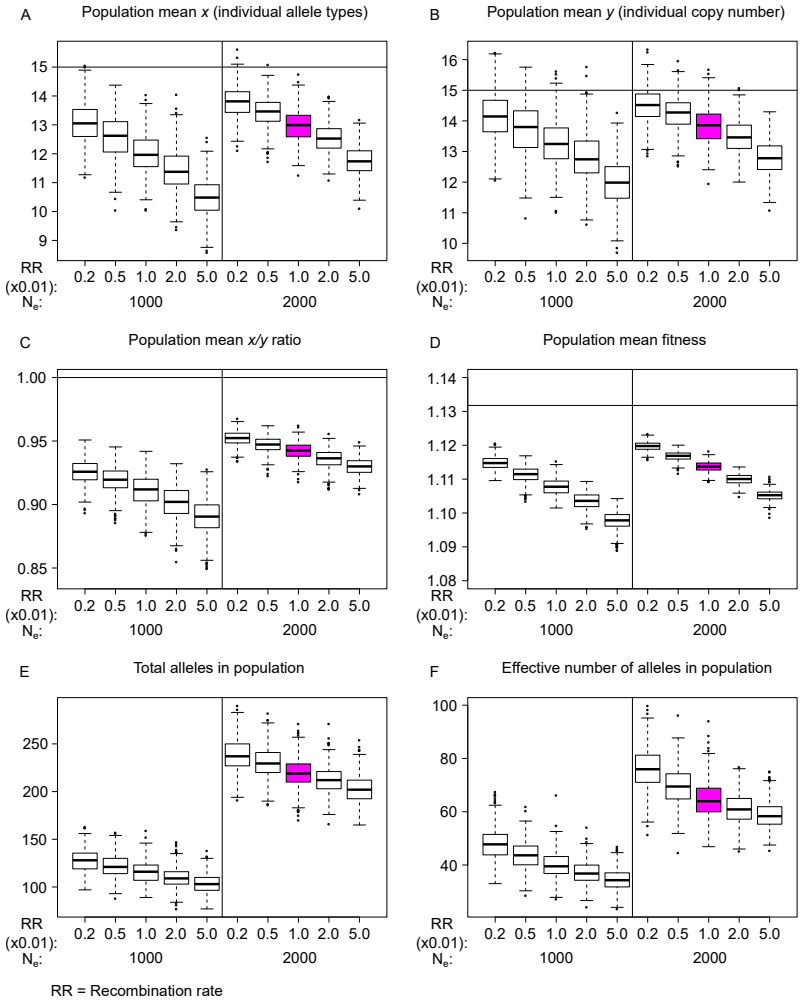


Figure 3.5: **Scenario (a2)** – constant population size. Population statistics at equilibrium: population mean $\bar{\mathbf{x}}$ (A); population mean $\bar{\mathbf{y}}$ (B); $\overline{\mathbf{x}/\mathbf{y}}$ ratio (C); population mean fitness (D); total (E) and effective number $|\mathbf{x}|_{\text{eff}}$ (F) of alleles. Varying parameters: population size $N_e = 1000, 2000$ and recombination rate ($r = 0.01$ times the factor indicated on the abscissa). Mutation rate $\mu = 0.0005$ and selection strength $(s_x, s_y) = (0.02, 0.005)$ are kept fixed. Boxplots based on 500 independent replicates. Box coloured in purple indicates the parameter combination (see Figure 3.4) shared by scenarios (a), (b), (c) and (d). Horizontal lines as explained in Figure 3.4.

new alleles are quickly created by mutation, but – while rare – they continue to bias the effective number of alleles, before equilibrium frequencies are restored. By segmental regression we found that mean fitness recovers faster than $|\mathbf{x}|_{\text{eff}}$ (Figure **S3** A and B). Furthermore, populations under stronger selection recover faster. The variation of these statistics among replicates is shown in Figure **S4**. Except for total and effective number of alleles, all other statistics show little among-replicate-variation after about 500 to 1000 generations after the bottleneck. Variation of the total number of alleles reaches a plateau and then gradually decreases, while among-replicate-variation of $|\mathbf{x}|_{\text{eff}}$ is generally small.

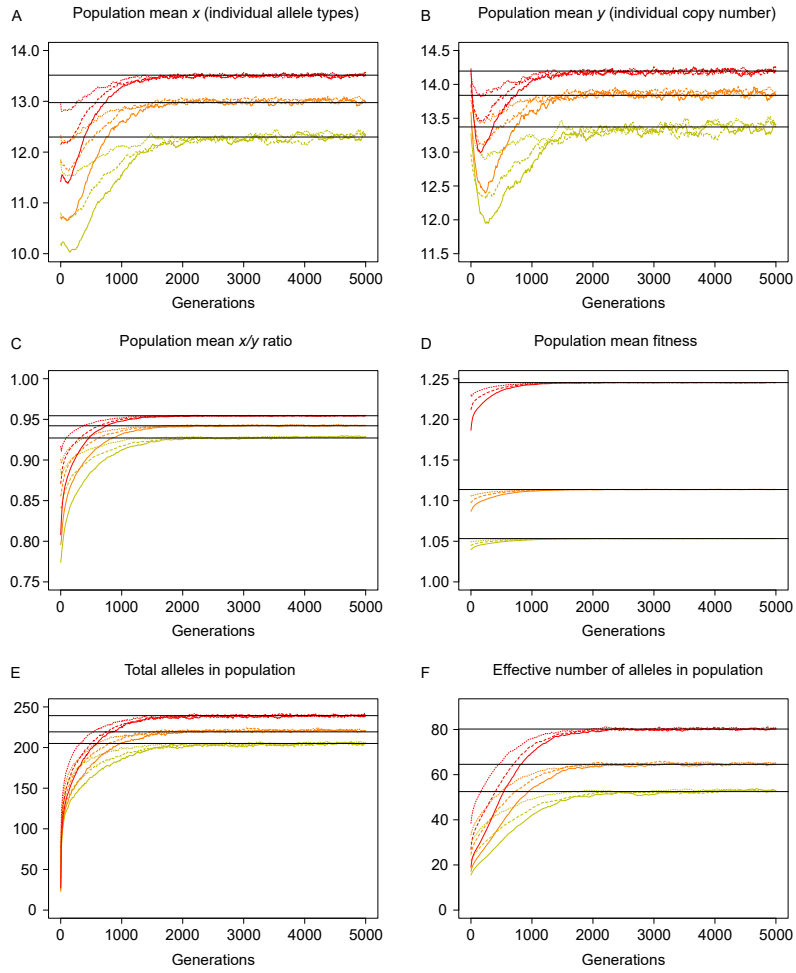


Figure 3.6: **Scenario (b)** – recovery after a bottleneck. Equilibrium populations with $N = 2000$ are reduced to $N_{\text{red}} = 20$ for a period of 5, 10 or 20 generations and then restored. During recovery six statistics are traced. A: population mean $\bar{\mathbf{x}}$; B: population mean $\bar{\mathbf{y}}$; C: ratio $\overline{\mathbf{x}/\mathbf{y}}$; D: mean fitness $\bar{\omega}$; E: total number of alleles; F: $|\mathbf{x}|_{\text{eff}}$. Red, orange and yellow indicate strong, intermediate and weak selection. Solid, dashed and dotted lines indicate bottleneck durations of 5, 10 and 20 generations. Each curve is an average across 200 replicates. Horizontal black lines are equilibria under constant population size.

In **scenario (c)**, we studied the effect of population subdivision and migration. We simulated reciprocal migration with two sub-populations of equal size, small ($N = 500$) and intermediate ($N = 1000$), starting from pairs of independent equilibrated replicates from scenario (a). Then, time was reset to $t = 0$ and migration was turned on with rates $Nm = 0.1$, 1 or 10 individuals per generation per direction. Summary statistics $\bar{\mathbf{x}}$, $\bar{\mathbf{y}}$, mean fitness $\bar{\omega}$, total number of alleles and $|\mathbf{x}|_{\text{eff}}$ in the combined super-population were recorded over time. After about 1500 to 2000 generations, these statistics approached a migration-drift-selection equilibrium, which is between the means for the panmictic populations of size $N_e = 1000$ and $N_e = 2000$. While the scenario with high migration ($Nm = 10$) is almost indistinguishable from the panmictic population with respect to $\bar{\mathbf{x}}$, $\bar{\mathbf{y}}$ and $\bar{\omega}$ (Figure 3.7A-

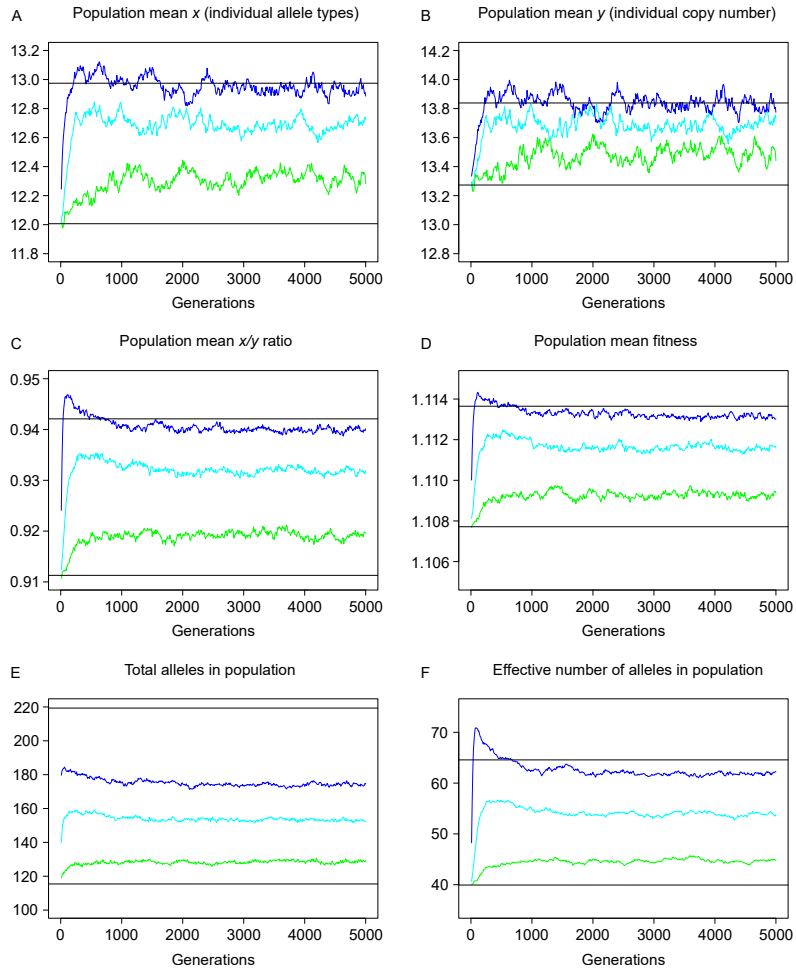


Figure 3.7: **Scenario (c)** – migration. Two separated and equilibrated sub-populations of size $N = 1000$ start to exchange migrants at time $t = 0$. Medium strength of selection ($s_x = 0.02, s_y = 0.005$). Migration rate: $2Nm = 0.1$ (green), 1 (cyan) or 10 (blue) migrants per generation in each direction. (A) population mean $\bar{\mathbf{x}}$; (B) population mean $\bar{\mathbf{y}}$; (C) ratio $\overline{\mathbf{x}/\mathbf{y}}$; (D) population mean fitness $\bar{\omega}$; (E) total and (F) effective number of alleles in the combined super-population. Shown are mean values across 100 replicates. Black lines indicate mean values (across 500 replicates) in panmictic populations of size $N_e = 1000$ (lower line) and $N_e = 2000$ (upper line).

D), there is still a clear deficit in the total and effective number of alleles compared to the panmictic population, even when the migration rate is high (Figure 3.7E,F). Note also in this case the initial overshooting of the panmictic equilibrium in the statistics $\overline{\mathbf{x}/\mathbf{y}}$, $\bar{\omega}$ and $|\mathbf{x}|_{\text{eff}}$ at about 100-200 generations, which is reminiscent of transient "hybrid vigour". Variation of these statistics among population replicates does not change appreciably with time (Figure **S5**). Similar results are observed for small populations $N_e = 500$ (Figure **S6** and **S7**).

In scenario (a2) we observed that lower recombination rates lead to an equilibrium of $\bar{\mathbf{x}}$ and $\bar{\mathbf{y}}$ which are closer to the optimum. A natural question to ask is whether the recombination

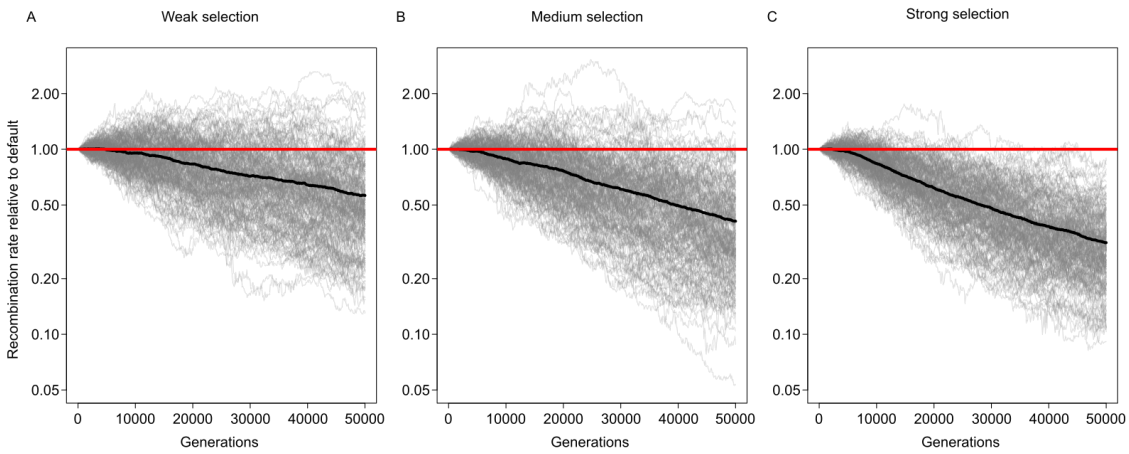


Figure 3.8: **Scenario (d)** – RRM: recombination rate modification. Populations, which have reached equilibrium without RRM, are carried on for 50,000 generations during which the recombination rate, encoded at a modifier locus, may change under the influence of selection. For all iterations: $N_e = 2000$, $r = 0.01$. Left: weak $(s_x, s_y) = (0.01, 0.0025)$; middle: intermediate $(0.02, 0.005)$; right: strong selection $(0.04, 0.01)$. Shown are trajectories of the recombination rate (in percent of its original value $r = 0.01$) for 200 replicates each. The mean across all 200 replicates is shown as a black line.

rate itself maybe subject to selection. Therefore, in **scenario (d)** a recombination rate modifier was added to the simple model. Given an equilibrated population which was reached with $r = 0.01$ as described in scenario (a), recombination rate modification was switched on, and time reset to $t = 0$. Recombination rate was coded by an RRM allele, which can increase or decrease the current recombination rate by a factor $e^{\pm 0.05}$ when mutated. Modification happens per chromosome per generation each with probability $p = 0.002$ for increase or for decrease. The RRM locus is thought to reside on the tip of a chromosome without itself being affected by recombination (Figure 3.1). Simulations were carried on for 50,000 generations and runs for each parameter setting of (s_x, s_y) were replicated 200 times. The results show that the mean recombination rate (average across all RRM alleles in the population) is continuously decreasing (Figure 3.8). It decreases more and faster when selection (s_x and s_y) is strong. When simulations terminated, the recombination rate was reduced – on average – to 56%, 41% and 31% of its original value ($r = 0.01$) and it showed a strongly negative correlation with population mean fitness (Pearsons's $r = -0.75, -0.83, -0.78$) for weak, intermediate and strong selection, respectively.

3.4 Discussion

We considered here a model in which two mechanisms, unequal recombination and mutation, may generate chromosomal diversity. While mutation leads to genetic diversity *sensu strictu*, by unequal recombination a chromosome may receive additional, or lose existing gene copies. Therefore, it is similar, but not identical, to segmental duplication or loss: copies gained by unequal recombination have their origin in a pairing haplotype, hence may be genetically diverse upon arrival, while those gained by duplication have their origin in the same haplotype, hence are genetically identical upon arrival. However, this distinction is negligible, since a single mutation event already suffices to make two identical copies distinct from each other when considering the infinite alleles model. Another feature of our model are the two overlaid components of the fitness function: it decreases with copy number, but increases with allele number, entailing a subtle and very interesting interaction of recombination and selection.

To gain some analytical insight into copy number dynamics under recombination, we first considered the neutral case in an infinitely large population. We find copy number of individuals to be distributed according to the discrete kernel of a Gamma distribution with an equilibrium mean which is identical to the initial mean at time $t = 0$ and remains constant over time. The limiting shape parameter is $\alpha = 4$, which is identical for all initial configurations. These two properties together uniquely determine the limiting distribution, which is independent of the shape of the initial distribution and of the recombination and mutation rates.

Adding selection changes the game. The limiting distribution becomes dependent on both the recombination rate and the strength of selection, but independent from the initial configuration. Still, it is well approximated by a Gamma distribution. The distribution that results from low selection strength or high recombination converges to the neutral equilibrium.

We inferred selection and recombination parameters for three different human genes, under the assumption of fixed epistasis $\varepsilon = 0.05$. Our analysis shows that observed copy number distributions can be well approximated within the framework of our model. Different means and variances of the distributions can be explained in terms of higher or lower recombination rates and stronger or weaker selection.

Note, that compound fitness, in which allele diversity is credited, contains a component of balancing selection: an individual which is heterozygous at any given locus has a higher fitness than one which is homozygous at the same locus. An important difference between the model considred here and one-locus models of balancing selection is the existence of gene copy number variation and unequal recombination. Note, that allelic diversity in the population can be stably maintained even in the case of allele fixation at single loci. The possibility to maintain allelic diversity through gene duplication, or unequal recombination, has been suggested by (Haldane, 1937). It is somewhat surprising that Haldane's idea has received only little attention in classical population genetics theory nor in experimental work. To our knowledge, tests confirming Haldane's hypothesis were conducted only a few years ago (Milesi et al., 2017).

We have shown that a high recombination rate has a negative effect on allelic diversity and resultant mean fitness. The reason is twofold: (1) a higher rate of unequal recombination produces individuals with much higher or lower copy number than the optimum, which have reduced fitness; (2) low recombination increases the likelihood for highly unfit homozygotes to appear, thus improving the efficiency of selection.

Populations which experienced strong bottlenecks are at risk of inbreeding depression, and loci under balancing selection are particularly affected (Frankham et al., 2014). Random loss of alleles increases homozygosity and consequently reduces fitness. This can affect and delay the recovery of genetic diversity even after population size has recovered (Miller and Lambert, 2004). In this study, we explored the effect of some parameters on the speed and process of bottleneck recovery at loci under diversifying selection. Both selection strength and bottleneck length influence the process. Relatively longer bottlenecks produce a temporary reduction in $\bar{\mathbf{x}}$, $\bar{\mathbf{y}}$ and mean fitness. The most likely reason is that high homozygosity results in selection towards haplotypes with fewer copies. Selection is more powerful after, than during, the bottleneck,when population size has recovered, but copy number recovery may lag behind. However, this somewhat paradoxical effect of fitness reduction at the initial phase of bottleneck recovery is only a short term effect, and – at least in part – due to the instantaneous, rather than gradual, restoration of population size in our model. Compared to fitness, $|\mathbf{x}|_{\text{eff}}$ is recovering even more slowly: for fitness to recover it suffices that new alleles appear and survive. But $|\mathbf{x}|_{\text{eff}}$ has recovered only when allele frequencies have reached their equilibrium values. Therefore, $|\mathbf{x}|_{\text{eff}}$ is a more sensitive statistic to test for deviation from equilibrium.

Simulations of scenario (c) show that fitness under population subdivision with moderate migration reaches an equilibrium which is intermediate between those under panmixis on the one hand and complete isolation on the other. While a short boost of hybrid vigour exists, we do not see a positive effect from limiting migration compared to panmixis. An earlier simulation study (Schierup et al., 2000) showed that the allelic diversity is largely insensitive to migration rates, but low-migration scenarios result in alleles with more divergent sequences. Additionally, balancing selection in the form of heterosis could increase the effective migration rate because migrant haplotypes are more likely to be successful in this case than under neutrality (Ingvarsson and Whitlock, 2000). Diversifying selection on MHC alleles has been shown to increase divergence between subpopulations, while diversity within subpopulations is still mostly governed by drift (Herdegen et al., 2014). MHC alleles and genes are also known to be shared among species through introgression, leading to restoration of diversity previously lost by drift (Dudek et al., 2019). In addition to generic balancing selection also local adaptation, i.e. the fixation of alleles which are adapted to specific subpopulations, may increase allelic diversity between populations (Ekblom et al., 2007). However, this effect is not considered in the model presented here, where selection operates only on the number of distinct alleles.

When the recombination rate is allowed to change over time we observe a trend towards lower rates. It is driven by selection and happens on a realistic population genetic timescale of some thousand generations. However, there is little empirical knowledge about (unequal) recombination rates in multi gene families. For example, in the human MHC locus the recombination rate is only about a third of the average genomic background rate (de Bakker et al., 2006; Traherne, 2008). On the other hand, studies on bovids (Schaschl et al., 2006) and horse (Beeson et al., 2019) show the opposite: high recombination in the MHC and olfactory receptor loci. In contrast again, the values reported for chicken seem to depend on mapping methodology (Fulton et al., 2016). Results from sheep (Petit et al., 2017) suggest a high "historical" (estimated from population data), but a low "meiotic" (from pedigree data) recombination rate, which suggests a recent change in time. From humans again, it is well known that recombination hotspots have a very fast turn-over time and are distinct in different subpopulations (Lam et al., 2013). Also, recombination rates may substantially differ in females and males – one example is the long arm of human chromosome 19 (Grimwood et al., 2004). Additionally, the presence of gene conversion makes the estimation of (reciprocal) recombination rates difficult (Martinsohn et al., 1999; Hosomichi et al., 2008). Anyway, current experimental results do not reveal a consistent picture as to whether there is a benefit, or trend, to suppress recombination in large multi gene families.

Caveats and future direction

While our model has incorporated multiple genetic processes, it is likely still far away from the details of how multi gene families evolve in real-life populations. One issue, not considered here, is gene conversion where an allele, or a fragment thereof, overwrites another one in a pairing chromosome. For example, gene conversion is known to play an important role in maintaining MHC diversity (Martinsohn et al., 1999; Högstrand and Böhme, 1999; Wiehe et al., 2000; Bahr and Wilson, 2012).

Also, our selection model assumes time-independent fitness and each allele provides the same selective benefit. This corresponds to an ideal situation where external factors are ubiquitous and stable. In practice, however, the selective benefits of certain alleles do change together with a changing environment. Evolving pathogens, for instance, lead to arbitrarily complex co-evolution dynamics (Ejsmond and Radwan, 2009; Tellier et al., 2014). Furthermore, population structure may interact with diversifying selection in a complex or even counter-intuitive way. In humans it is known that different populations harbour different MHC alleles, likely driven by pathogen diversity (Manczinger et al., 2019). A hypothesis is that multiple subpopulations act as reservoirs of alleles and backups for each other, allowing for quick response against new pathogens (Lenz et al., 2009; Linnenbrink et al., 2018). Interaction between population structure and local adaptation needs to take into account subpopulation sizes and migration networks. For instance it was shown that subpopulation sizes can affect local allelic diversity (Mason et al., 2009).

Finally, and perhaps most importantly, gene function decides on fitness. On population genetic time scales pseudogenization plays an important role for the evolution of multi gene families (Hess, 2000; Menashe et al., 2006). Although eventually removed by selection, pseudogenes can persist in real-life populations with high frequency. Conditions under which pseudogenes appear and persist can be identified in accordingly modified models. Structural and functional aspects being included together with gene conversion, temporally or locally varying selection strengths into theoretical models will help to address open questions, but remains to be considered in future work.

Appendix

Proof of (3.2). Using the closed form formula of the geometric series and the fact that $x = y$, we can write the fitness function $\omega = (1+s_x)^{(\sum_{i=0}^{x-1}\beta_x^i)} \times (1-s_y)^{(\sum_{j=0}^{y-3}\beta_y^j)}$ as a function of y that equals

$$f(y) = (1+s_x)^{\frac{1-\beta_x^y}{1-\beta_x}} \times (1-s_y)^{\frac{1-\beta_y^{y-2}}{1-\beta_y}}.$$

Defining

$$a := (1+s_x)^{\frac{1}{1-\beta_x}}, \quad b := (1-s_y)^{\frac{1}{1-\beta_y}},$$

we find that

$$f'(y) = -\left(\ln(a)\ln(\beta_x)\cdot\beta_x^y + \ln(b)\ln(\beta_y)\cdot\beta_y^{y-2}\right)\cdot a^{1-\beta_x^y}\cdot b^{1-\beta_y^{y-2}}.$$

Setting $f'(y^*) = 0$ leads us to

$$\begin{aligned} & \underbrace{-\ln(a)\ln(\beta_x)}_{:=p_1}\cdot\beta_x^{y^*} = \underbrace{\ln(b)\ln(\beta_y)\frac{1}{\beta_y^2}}_{:=p_2}\cdot\beta_y^{y^*} \\ \Rightarrow \quad & p_1\beta_x^{y^*} = p_2\beta_y^{y^*} \\ \Rightarrow \quad & y^* = \frac{\ln(p_1)-\ln(p_2)}{\ln(\beta_y)-\ln(\beta_x)}, \end{aligned}$$

and inserting the expressions for p_1, p_2, a, b gives the result.

□

Proof of Proposition 3.3.1. We note that the parental status of the chromosomes do not matter in the following calculations. Therefore, we use the notation $y'_{(\cdot)}$ instead of $y^{m\prime}$ and y'^p. Since the T describes the distribution of the sum of two uniform random variables, we observe that the expected value is given by

$$\sum_{y'} y'\cdot T(y', y'_1, y'_2) = \mathbb{E}\left[B_1 + B_2 - 1\right] = \frac{y'_1+1}{2} + \frac{y'_2+1}{2} - 1 = \frac{y'_1+y'_2}{2},$$

and therefore conclude that

$$\begin{aligned}
&\sum_{y'} y' \cdot p_{t+1}(y') \\
&= (1-r)\sum_{y'} y' \cdot p_t(y') + r\sum_{y'} y' \cdot \sum_{y_1',y_2'} p_t(y_1')p_t(y_2')T(y',y_1',y_2') \\
&= (1-r)\sum_{y'} y' \cdot p_t(y') + r\sum_{y_1',y_2'} p_t(y_1')p_t(y_2')\frac{y_1'+y_2'}{2} \\
&= (1-r)\sum_{y'} y' \cdot p_t(y') + r\sum_{y_1'} \frac{y_1'}{2} p_t(y_1') \underbrace{\sum_{y_2'} p_t(y_2')}_{=1} + r\sum_{y_2'} \frac{y_2'}{2} p_t(y_2') \underbrace{\sum_{y_1'} p_t(y_1')}_{=1} \\
&= \sum_{y'} y' \cdot p_t(y').
\end{aligned}$$

We define $a = 2/E_{Y'}$ and note that the stationary distribution is independent from the recombination rate $r > 0$, i.e.

$$\begin{aligned}
& p_{\text{stat}}(y') = (1-r)p_{\text{stat}}(y') + r \cdot \sum_{y_1',y_2'} p_{\text{stat}}(y_1')p_{\text{stat}}(y_2')T(y',y_1',y_2') \\
\Leftrightarrow \quad & p_{\text{stat}}(y') = \sum_{y_1',y_2'} p_{\text{stat}}(y_1')p_{\text{stat}}(y_2')T(y',y_1',y_2').
\end{aligned}$$

Therefore, we find that

$$\begin{aligned}
&\sum_{y_1',y_2'} p_{\text{stat}}(y_1')p_{\text{stat}}(y_2')T(y',y_1',y_2') \\
&= \left(\frac{1}{Z}\right)^2 \cdot \Bigg(\sum_{y_2'=1}^{y'} y_2' \cdot e^{-ay_2'} \sum_{y_1'=y'+1}^{\infty} e^{-ay_1'} + \sum_{y_2'=y'+1}^{\infty} e^{-ay_2'} \sum_{y_1'=1}^{y'} y_1' \cdot e^{-ay_1'} \\
&\quad + y' \cdot \sum_{y_2'=y'+1}^{\infty} e^{-ay_2'} \sum_{y_1'=y'+1}^{\infty} e^{-ay_1'} + \sum_{y_2'=1}^{y'} e^{-ay_2'} \sum_{y_1'=y'-y_2'}^{y_2'} (y_1'+y_2'-y')e^{-ay_1'} \Bigg) \\
&= \ldots (*) \ldots \\
&= y' \cdot e^{-ay'} \cdot \frac{(e^a-1)^2}{e^a} \\
&= p_{\text{stat}}(y'),
\end{aligned}$$

where the detailed calculations of $(*)$ are shown below. □

Proof of (*). Using the substitution $k = (y_1' + y_2' - y')$ we find that

$$
\begin{aligned}
& \sum_{y_1', y_2'} p_{stat}(y_1') p_{stat}(y_2') T(y', y_1', y_2') \\
= \; & \frac{(e^a - 1)^4}{e^{2a}} \cdot \Bigg[\sum_{y_2'=1}^{y'} y_2' \cdot e^{-a y_2'} \sum_{y_1'=y'+1}^{\infty} e^{-a y_1'} + \sum_{y_2'=y'+1}^{\infty} e^{-a y_2'} \sum_{y_1'=1}^{y'} y_1' \cdot e^{-a y_1'} \\
& + y' \cdot \sum_{y_2'=y'+1}^{\infty} e^{-a y_2'} \sum_{y_1'=y'+1}^{\infty} e^{-a y_1'} + \sum_{y_2'=1}^{y'} e^{-a y_2'} \sum_{y_1'=y'-y_2'}^{y_2'} (y_1' + y_2' - y') e^{-a y_1'} \Bigg] \\
= \; & \frac{(e^a - 1)^4}{e^{2a}} \cdot \Bigg[2 \frac{e^{-a(y'+1)}}{1 - e^{-a}} \cdot \left(-\frac{\partial}{\partial a} \right) \left(\frac{1 - e^{-a(y'+1)}}{1 - e^{-a}} - 1 \right) + y' \cdot \left(\frac{e^{-a(y'+1)}}{1 - e^{-a}} \right)^2 \\
& + \sum_{y_2'=1}^{y'} e^{-a y_2'} \sum_{k=0}^{y_2'} k \cdot e^{-a(k + y' - y_2')} \Bigg] \\
= \; & \frac{(e^a - 1)^4}{e^{2a}} \cdot \Bigg[2 \frac{e^{-a(y'+1)}}{1 - e^{-a}} \cdot \frac{e^{-a y'} \left(e^{a(y'+1)} - ((y'+1)e^a + y') \right)}{(e^a - 1)^2} + y' \cdot \left(\frac{e^{-a(y'+1)}}{1 - e^{-a}} \right)^2 \\
& + e^{-a y'} \sum_{y_2'=1}^{y'} \left(-\frac{\partial}{\partial a} \right) \left(\frac{1 - e^{-a(y_2'+1)}}{1 - e^{-a}} - 1 \right) \Bigg] \\
= \; & \frac{(e^a - 1)^4}{e^{2a}} \cdot \Bigg[\frac{e^{-2a y'} \left(2e^{a(y'+1)} - (y'+2)e^a + y' \right)}{(e^a - 1)^3} \\
& + e^{-a y'} \cdot \left(-\frac{\partial}{\partial a} \right) \left(\frac{y' - e^{-a} \left(\frac{1 - e^{-a(y'+1)}}{1 - e^{-a}} - 1 \right)}{1 - e^{-a}} \right) \Bigg] \\
= \; & \frac{(e^a - 1)^4}{e^{2a}} \cdot \Bigg[\frac{e^{-2a y'} \left(2e^{a(y'+1)} - (y'+2)e^a + y' \right)}{(e^a - 1)^3} \\
& + \frac{e^{-2a y'} \left(y' e^{a(y'+2) + (y'+2)e^a - (y'+2)e^{a(y'+1)} - y'} \right)}{(e^a - 1)^3} \Bigg] \\
= \; & \frac{(e^a - 1) \cdot e^{-2a y'} \cdot \left(y' e^{a(y'+2)} - y' e^{a(y'+1)} \right)}{e^{2a}} \\
= \; & y' \cdot e^{-a y'} \cdot \frac{(e^a - 1)^2}{e^a}.
\end{aligned}
$$

□

Table 3.3: Parameters used in simulations of the compound model.

Scenario (a) Single population of constant size N_e	
N_e:	$500, 1000, 2000, 4000$
μ:	0.0005
(a1) (s_x, s_y):	$(0.01, 0.0025)$, $(0.02, 0.005)$, $(0.04, 0.01)^\dagger$
(a1) r:	0.01
(a2) (s_x, s_y):	$(0.02, 0.005)$
(a2) r:	$0.002, 0.005, 0.01, 0.02, 0.05$
replicates:	500 per parameter combination
recording:	every 100-th for 20000 generations

†: the three levels of selection strengths are referred to as "weak", "intermediate" and "strong" in the text

Scenario (b) Instantaneous bottleneck	
$N_0^\dagger$:	$1000, 2000$
$N_b^\ddagger$:	20
duration :	$5, 10, 20$ generations
μ:	0.0005
(s_x, s_y):	$(0.01, 0.0025)$, $(0.02, 0.005)$, $(0.04, 0.01)$
r:	0.01
replicates:	200 per parameter combination
recording:	every 10-th for $5,000$ generations after bottleneck

†: population size before and after bottleneck;
‡: population size during bottleneck

Scenario (c) Two populations of constant size N_e with two-way migration†	
N_e:	$500, 1000$
$N_e\, m$:	$0.1,\ 1,\ 10$
μ:	0.0005
(s_x, s_y):	$(0.01, 0.0025)$, $(0.02, 0.005)$, $(0.04, 0.01)$
r:	0.01
replicates:	100 pairs per parameter combination
recording:	every 10-th for $2,000$ generations

†: at rate m per individual per generation per direction

Scenario (d) Single population of constant size N_e with recomb. rate modifier ρ	
N_e:	$1000, 2000$
μ:	0.0005
(s_x, s_y):	$(0.01, 0.0025)$, $(0.02, 0.005)$, $(0.04, 0.01)$
base rate r_o:	0.01
initial ρ_0:	1 for all chromosomes
modification†	of $r = r_o \cdot \rho$ according to $\begin{cases} \rho_{t+1} = \rho_t & (p = 0.996) \\ \rho_{t+1} = \rho_t \cdot e^{0.05} & (p = 0.002) \\ \rho_{t+1} = \rho_t \cdot e^{-0.05} & (p = 0.002) \end{cases}$
replicates:	200 per parameter combination
recording:	every 100-th for $50,000$ generations

†: ρ changes from ρ_t to ρ_{t+1} per generation per chromosome with probability p

4 Distinguishing the roles of adaptation and demography in gene copy number changes in human populations

Authors: Moritz Otto, Yichen Zheng, Paul Grablowitz and Thomas Wiehe
Status: Submitted to G3: Genes|Genomes|Genetics
Available on BioRxiv

DOI: https://doi.org/10.1101/2023.08.14.553171

Abstract

Genes with multiple copies are likely to be maintained by stabilizing selection that puts a bound to unlimited expansion of copy number. We designed a model where copy number variation is generated by unequal recombination, which fits well to a number of genes, surveyed in three human populations. Based on this theoretical model and on computer simulations, we were interested in answering the question whether gene copy number distribution in the derived European and Asian populations can be explained by a purely demographic scenario or whether shifts in the distribution are signatures of adaptation. Although copy number distribution in most of the analyzed gene clusters can be explained by a bottleneck as in the out of Africa expansion of homo sapiens 60-10kyrs ago, we identified several candidate genes, for instance `AMY1A` and `PGA3`, whose copy numbers are likely to be selected differently among African, Asian and European populations.

4.1 Introduction

Gene copy number variation (CNV) refers to the presence of multiple copies of a gene family within a genome, resulting from duplications, deletions, or rearrangements. Combined with their high mutation rate CNVs constitute a significant driver of genomic variability that allows for rapid adaptive evolution in response to environmental changes (Sudmant et al., 2015a; Brahmachary et al., 2014; Carvalho and Lupski, 2016; Iskow et al., 2012; Sebat et al., 2004).

A well studied example of CNV within human population is provided by the salivary amylase gene, whose variations in the number of copies are hypothesized to correlate with the extent of dietary starch consumption not only in human but also in other species (Pajic et al., 2019; Atkinson et al., 2018; Carpenter et al., 2015; Usher et al., 2015; Falchi et al., 2014; Perry et al., 2007).

In general, copy number variation may result from different evolutionary forces acting upon them. Demographic events, such as population migrations and expansions, can lead to changes in gene frequencies and distributions over time. Simultaneously, natural selection acts on genetic variations, favoring advantageous alleles and promoting their proliferation within populations.

It is known that both demographic effects and selection may produce similar patterns in single nucleotide as well as in structural variants, making it difficult to disentangle these forces (Lohmueller, 2014; Stajich, 2004). For SNP or allele frequency data, there have been well-developed statistics (e.g., Tajima (1989); Fu (1997)) that are "standardized " so that a genomic baseline can be established, from which loci under selection may be detected. However, such a genomic baseline is not available for gene copy number variation data. Therefore, we resorted to a more basic approach involving modelling and computer simulations.

We have recently examined the evolutionary dynamics of multi-copy gene families with respect to selective pressure and unequal recombination (Otto et al., 2022). This study focused on analyzing the impact of stabilizing selection on gene copy numbers, while considering the role of recombination as a randomizing mechanism that introduces variability within the population.

By expanding this model, we aimed to assess whether gene copy number alterations observed within human populations could be solely attributed to demographic events or whether selective pressures have played a role in shaping these variations.

In this study, we conducted extensive simulations under various scenarios of human demography and selective changes. By disentangling the effects of these two forces, we sought to gain a deeper understanding of the evolutionary processes driving gene copy number variation in human populations. Based on empirical data of human gene copy numbers we identified several candidate genes, whose copy numbers are likely to be selected differently among African, Asian and European populations.

4.2 Materials and Methods

4.2.1 Gene copy number variation in human

We started with the dataset provided by Brahmachary et al. (2014). Using Nanostring technology they estimated gene copy numbers of 180 gene-families in 165 individuals of three populations (60 African Yuroba - YRI, 60 Central Europe - CEU and 45 Asia - CHB) based on data collected in the framework of the 1,000 Genomes Project (Sudmant et al., 2010). While some of these loci showed copy numbers of > 100 copies (`DUX4` even up to 600), we focused on intermediate copy numbers and removed all satellite loci, genes on sex chromosomes, genes with minimum copy number below 2, and genes with mean copy number (in YRI) below 5 or above 60. For genes that have two primer sets, only one is taken. We used t-test and f-test statistics to select gene families with significant differences in mean and standard deviation between either YRI-CHB or YRI-CEU comparisons and removed those that showed no statistical evidence in any of these. This resulted in 42 gene families, see Table 4.1. An example of the copy number distribution of four gene families is shown in Figure 4.1.

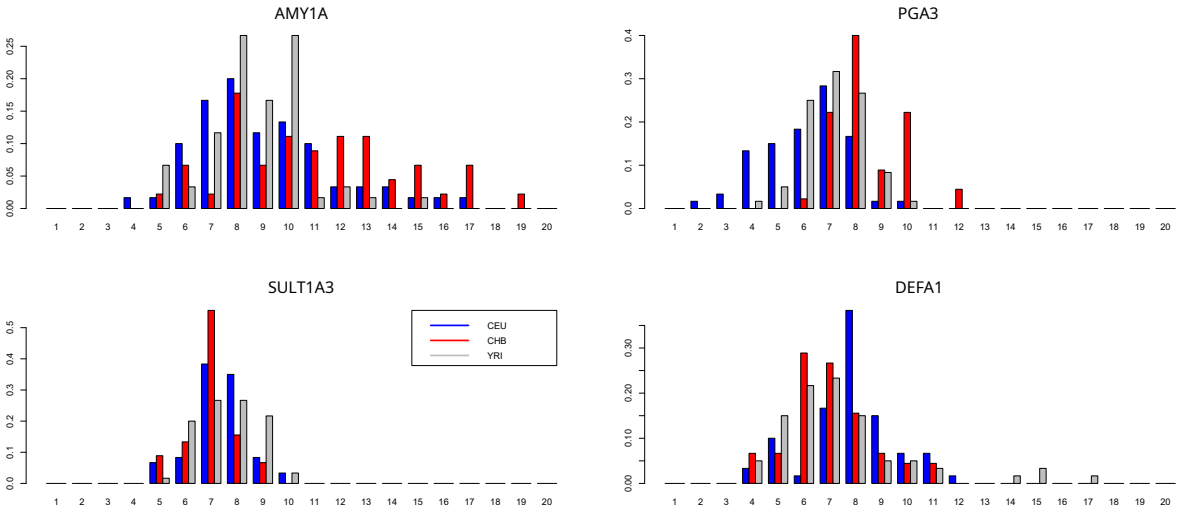


Figure 4.1: Gene copy number distribution in four exemplary gene famiies in three human populations, CEU, CHB, YRI. Data adapted from (Brahmachary et al., 2014).

Table 4.1: Differences in mean and variance between populations. 0 indicates no significant change, + a significant increase and – a significant decrease (t- and f-test, $\alpha = 0.05$).

Gene	YRI mean	CEU mean		CHB mean		YRI sd	CEU sd		CHB sd	
AMY1A	8.628	0	9.151	+	11.128	1.907	+	2.713	+	3.35
ANKRD20A3	26.414	+	29.148	+	29.904	1.804	0	1.718	0	1.698
BOLA2B	7.235	-	6.679	-	6.49	1.012	0	0.914	0	0.839
CBWD3	12.146	0	12.374	+	13.068	0.995	0	0.949	+	1.81
CDC37P1	14.941	+	19.977	0	16.582	4.11	+	5.619	+	5.579
CLEC18A	7.799	+	8.362	0	7.932	1.331	0	1.216	0	1.392
CSH	6.738	+	7.182	+	7.474	0.497	0	0.555	0	0.575
DEFA1	7.442	0	7.891	0	7.056	2.643	-	1.671	-	1.604
DEFB130	5.081	+	5.315	0	5.243	0.562	0	0.532	0	0.462
FAM72A	6.914	+	7.573	+	7.561	0.617	+	0.86	0	0.651
FAM75A1	11.859	0	11.972	+	13.362	1.473	0	1.391	+	2.019
FAM75A5	11.693	0	11.522	+	12.533	1.115	0	1.197	+	1.751
FCGBP	5.282	+	5.693	+	5.79	1.291	-	0.678	0	1.046
FOXD4L2	13.013	+	13.694	+	14.55	1.015	0	0.994	+	1.877
GOLGA6L9	27.683	0	28.586	+	29.181	2.615	0	2.532	0	2.59
GOLGA8G	29.209	+	31.641	+	30.37	3.065	0	2.783	0	2.35
GUSBP1	12.95	+	15.886	+	13.987	2.249	0	2.585	0	2.213
HIST2	8.436	+	8.709	+	8.894	0.528	0	0.673	0	0.644
LIMS3	5.829	-	5.408	-	5.661	0.346	0	0.354	0	0.39
LOC23117	50.194	0	50.304	-	48.639	3.685	0	2.963	0	2.789
LOC653606	6.56	0	6.403	-	5.999	0.486	0	0.621	+	0.917
MUC12	11.845	+	14.098	0	12.123	2.586	0	2.01	-	1.803
NBPF11	49.963	-	48.002	0	48.68	4.203	-	3.114	0	3.311
NBPF16	45.25	0	46.436	+	47.006	4.706	0	5.023	0	3.988
NPIP	51.171	-	49.488	-	48.938	2.16	0	2.327	0	2.224
PGA3	7.044	-	6.181	+	8.473	1.205	+	1.565	0	1.353
PPIAP21	43.141	+	48.632	+	49.493	3.765	0	4.315	0	3.881
PRAMEF14	10.516	+	11.835	+	11.888	1.295	+	2.246	+	1.937
PRAMEF20	7.253	0	7.415	+	7.576	0.566	0	0.723	+	0.924
PRAMEF5	17.844	-	16.475	-	15.804	1.721	+	2.386	+	2.578
PRAMEF8	5.919	0	5.787	0	5.842	0.652	+	1.281	0	0.819
PRR11	6.868	+	8.298	+	8.305	0.923	0	0.965	0	0.708
PRR20A	20.639	-	17.284	-	14.85	6.903	-	5.288	0	5.584
PSG3	14.943	+	15.624	0	15.087	1.314	0	1.238	+	1.843
RGPD1	13.959	0	14.037	0	14.151	0.791	+	1.309	+	1.266
SPDYE3	34.611	-	31.656	-	32.828	2.836	-	2.105	0	2.617
SULT1A3	7.627	0	7.406	-	7.017	1.197	0	1.087	0	0.904
TBC1D3	45.515	-	33.191	-	39.306	6.337	0	6.888	+	8.381
TCEB3C	33.02	-	28.574	-	25.895	7	0	7.383	0	6.299
TP53TG3	9.172	0	8.904	-	6.735	1.825	0	2.08	0	1.666
TRIM49L1	12.353	+	14.078	+	14.112	1.664	0	2.06	0	1.874
ZNF658B	5.544	+	6.273	+	6.647	0.727	0	0.827	+	1.01

4.2.2 Unequal recombination model

In a recently developed model we considered unequal recombination and selection to describe the evolution of tandem gene arrays (Otto et al., 2022). We shortly summarize the main findings. Consider two chromosomes with gene arrays of size y_1 and y_2. A recombination event happens at rate r and may produce a gamete of gene array size according to the trapezodial distribution, such that

$$Prob[y|y_1, y_2] = \frac{1}{y_1 y_2} \begin{cases} 0, & y < 1 \\ y, & 1 \leq y < \min(y_1, y_2) \\ \min(y_1, y_2), & \min(y_1, y_2) \leq y < \max(y_1, y_2) \\ y_1 + y_2 - y, & \max(y_1, y_2) \leq y < y_1 + y_2 - 1 \\ 0, & y \geq y_1 + y_2 \end{cases} .$$

See Figure 4.2A for an illustration. We apply a fitness function, where each newly arising copy has a positive, yet decreasing benefit s_x. This is motivated by assuming a beneficial effect, yet with diminishing returns, either of increased gene dosage or of increased allelic diversity within an individual (Otto et al., 2022). At the same time, we assume additional copies to be selected against with an increasing selective disadvantage s_y. This is motivated by assuming an increasing cost of replication, of gene processing and of maintaining genome integrity. Both effects are cast in a double-epistatic fitness function with two selection coefficients (s_x, s_y), governed by a single epistasis parameter (ε). To avoid the trivial long-term evolution equilibrium of one copy, we assume $s_x > s_y$. Furthermore, we assume $\varepsilon = 0.05$ to be constant in the following.

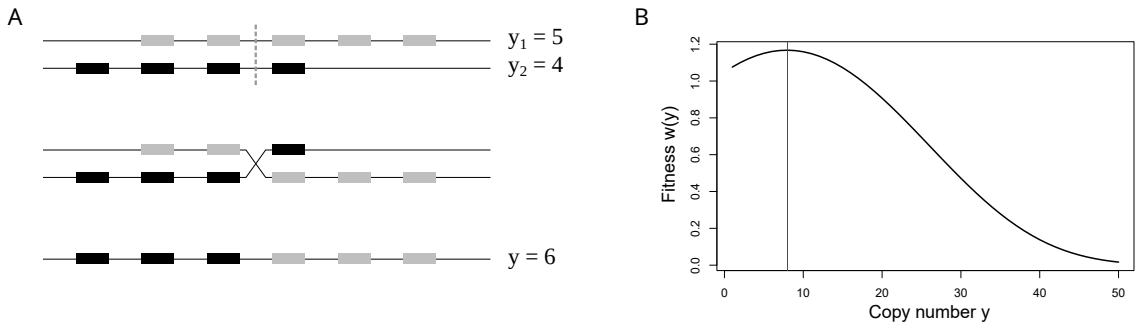


Figure 4.2: A. Sketch of the unequal recombination process. Starting with two chromosomes with $y_1 = 5$ and $y_2 = 4$ gene copies, two breaking points are chosen. One of the recombinants is then propagated. Its copy number $y = 6$ is Trapezodial as shown in (Otto et al., 2022). B. Example of the fitness function $\omega(y)$ equation (4.1) with $\varepsilon = 0.05, s_x = 0.05, s_y = 0.0025$, which leads to an optimal copy number $y_{\text{opt}} \approx 8$ copies.

Summarizing, fitness of a diploid individual with total gene copy number y is given by

$$\omega(y) = \exp\left\{\frac{1}{\varepsilon}\left((s_x + s_y) - s_x \cdot e^{-\varepsilon y} - s_y \cdot e^{\varepsilon(y-2)}\right)\right\} \tag{4.1}$$

This leads to an optimal copy number y_{opt} of

$$y_{\text{opt}} = 1 + \frac{\log(s_x/s_y)}{2\varepsilon}, \tag{4.2}$$

which is determined by the ratio s_x/s_y when ε is kept fixed. See Figure 4.2B for an example. The population is then simulated according to a Wright-Fisher model with non-overlapping generations and with selection and recombination described as above. It was shown, that in the deterministic model the equilibrium copy number distribution is centered around y_{opt} and is well approximated by a Gamma distribution (Otto et al., 2022). Furthermore, it holds that the coefficient of variation $\mathcal{C}_V = \sigma/\bar{y}$ is correlated to the logarithm of the recombination - selection ratio $\log(r/s_x)$. With strong selection and low recombination the distribution is tightly distributed around the optimal value, whereas higher r and lower s_x lead to a widespread distribution. Therefore, we introduce two new parameters:

- $q_S = s_x/s_y$, the *'selection ratio'*, which determines the optimal copy number, such that for $\varepsilon = 0.05$ we find

$$y_{\text{opt}} = y_{\text{opt}}(q_S) = 1 + 10 \cdot \log(q_S)$$

- $q_R = r/s_x$, the *'recombination/selection ratio'*, which measures the randomness produced by the unequal recombination process versus the selective pressure of the fitness function and therefore determines the coefficient of variation $\mathcal{C}_V = \sigma/\bar{y}$ of the equilibrium distribution.

4.2.3 Regression

To analyze the equilibrium distribution of the unequal recombination process under drift, we simulated the population evolution under different parameter settings. Population size is kept at $N = 5,000$ and assumed to be at an initial state of 5 copies on each chromosome. The different input parameters are given in Table 4.2.

Together, they define 324 triples r, s_x, s_y. Additionally, we generated 160 random pairs such that q_R is between 0.01 and 5 and y_{opt} is between 4 and 60 and combined them with

Table 4.2: Parameters for regression simulations.

4 recombination rates r	0.1%, 0.2%, 0.5% and 1%
9 recombination ratios $q_R = r/s_x$	0.01, 0.02, 0.05, 0.1, 0.5, 1.0, 2.0, 5.0
9 optimal copy number values y_{opt}	10, 15, 20, 25, 30, 35, 40, 45, 50

Table 4.3: Estimates of selection coefficients s_x, s_y under four recombination rates $r = 0.001, ..., 0.01$ based on regression equation (4.3). The displayed gene families are the ones of Figure 4.1 for all three populations. Values in brackets are out of the range $0.001 < s_x < 0.1$ in YRI and hence not used in simulations.

					$r = 0.001$		$r = 0.002$		$r = 0.005$		$r = 0.01$	
Gene	Pop	Mean	SD	y_{opt}	s_x	s_y	s_x	s_y	s_x	s_y	s_x	s_y
AMY1A	CEU	9.1511	2.7133	9.2708	0.0012	0.0005	0.0025	0.0011	0.0062	0.0027	0.0125	0.0055
	CHB	11.128	3.3503	11.282	0.0011	0.0004	0.0022	0.0008	0.0054	0.0019	0.0109	0.0039
	YRI	8.6279	1.9074	8.7386	0.0048	0.0022	0.0097	0.0045	0.0242	0.0111	0.0483	0.0223
PGA3	CEU	6.1808	1.5646	6.2491	0.0029	0.0017	0.0058	0.0035	0.0146	0.0086	(0.0292)	(0.0173)
	CHB	8.4731	1.3526	8.5811	0.0144	0.0068	0.0289	0.0135	0.0722	0.0338	(0.1445)	(0.0677)
	YRI	7.0444	1.2053	7.1277	0.0122	0.0066	0.0245	0.0133	0.0611	0.0331	(0.1223)	(0.0663)
SULT1A3	CEU	7.4058	1.0872	7.4953	0.0186	0.0097	0.0373	0.0195	0.0932	0.0487	(0.1865)	(0.0974)
	CHB	7.0165	0.9041	7.0993	0.0259	0.0141	0.0518	0.0282	0.1295	0.0704	(0.2591)	(0.1408)
	YRI	7.6269	1.1971	7.7202	0.0155	0.0079	0.0311	0.0158	0.0774	0.0395	(0.1548)	(0.0791)
DEFA1	CEU	7.8911	1.6708	7.9889	(0.0058)	(0.0029)	(0.0116)	(0.0058)	0.0291	0.0145	0.0581	0.0289
	CHB	7.0561	1.6041	7.1396	(0.0045)	(0.0024)	(0.0091)	(0.0049)	0.0225	0.0122	0.0451	0.0244
	YRI	7.4421	2.6428	7.5321	(0.0005)	(0.0002)	(0.0009)	(0.0005)	0.0023	0.0012	0.0046	0.0024

the four recombination rates, leading to a total parameter set of 964 combinations, where we disregarded those triples with selective strengths $s_x > 0.1$ to keep a realistic parameter range.

For each of this parameter combinations, we evolve the population under the given selection scheme for 5 million generations. The first 200,000 generations were discarded as burn-in and the population statistics (mean copy number $\bar{y}$ and standard deviation σ) are recorded every 20,000 generations.

In total, this results in $\approx 185,000$ data points, which we used to determine the relationship of input parameters (r, s_x, s_y) and output population statistics $(\bar{y}, \sigma)$.

As indicated in (Otto et al., 2022) we suggest a mean copy number $\bar{y}$ close to its optimal value y_{opt} and a correlation of $\mathcal{C}_V$ to $\log(q_R)$. Indeed, with r^2-values of 0.9842 and 0.9088 we find

$$\begin{aligned} \bar{y} &= 0.0379 + 0.983 \cdot y_{\text{opt}} \\ \mathcal{C}_V = \frac{\sigma}{\bar{y}} &= 0.323 + 0.0566 \cdot \log(q_R) - 0.00152 \cdot y_{\text{opt}} - 0.000036 \cdot \log(q_R) \cdot y_{\text{opt}} \end{aligned} \tag{4.3}$$

We calculated the q_S and q_R ratios based on $\bar{y}$ and $\mathcal{C}_V$ from gene copy numbers (see Tab 4.1) using the regression formula (4.3) with four recombination rates $r = 0.001$, 0.002, 0.005 and 0.01. Results for the four candidate genes shown in Figure 4.1 are given in Table 4.3.

4.2.4 Demography simulations

To determine whether significant changes of mean and variance of the copy number distribution (Table 4.1) can be explained by demographic history of human populations, we examined in total 6 different scenarios (enumerated as I - VI), see Figure 4.3.

Simulation of the bottleneck model First, we ran a simple bottleneck model of three different population reductions. Each is divided into three phases: (1) Burn-in phase. For each gene we used the estimated (r, s_x, s_y)-triple based on the dataset from YRI. These parameters were chosen as input to produce an equilibrium population of $N = 10,000$ by a burn-in process of 200,000 generations. 'Independent' equilibrium populations are produced by recording the population state every 20,000 generations. (2) Bottleneck. From equilibrium we reduced the population size to $N = 100,500$ or $1,000$, denoted scenario I, II and III, and kept it such for 5,000 generations. (3) Recovery phase. At the end of the bottleneck, the population is reset to $N = 10,000$ and the copy number distribution is recorded every 50 generations until generation 1000 after the bottleneck. We ran the bottleneck simulations I - III on all gene families given in Tab 4.1, with recombination rates $r = 0.001$, 0.002, 0.005 and 0.01, and discarded parameter combinations with s_x outside the interval [0.001, 0.1] in YRI. This gives a final total of 42 gene families and 95 gene-r combinations. For each gene, recombination rate and bottleneck population size combinations, 10,000 replicates are produced (from 100 'independent' starting equilibria).

We then traced mean and $\mathcal{C}_V$ along the recovery phase and compared these with the empirical data from CHB and CEU populations.

Simulation of the human population history A more realistic population history of human is given by the *Genetic Algorithm for Demographic Model Analysis* (GADMA) (Noskova et al., 2020), which also includes migration between subpopulations. We ran simulations on four candidate genes (`AMY1A, PGA3, SULT1A3, DEFA1`) with the following modification of the GADMA-demography: As ancestral population ($N = 9,900$ in GADMA), we used the equilibrium populations ($N = 10,000$) from the previous section. Therefore, we started the simulation 5992 generations before present, roughly corresponding to the onset of the 'out-of-Africa' expansion, when the Eurasian population split from the ancestral population and experienced a sharp bottleneck. To reduce computation time, we did not simulate the continued evolution of the African (YRI) population, since we assumed it to be in equilibrium; for migration from YRI to Eurasian populations, we drew samples from the ancestral population. At 896 generations before present, CEU and CHB split from each other and started to evolve including reciprocal migration and exponentially increasing population size. In the following, we refer to this simulation as scenario IV. At 'present', copy number distributions (mean and variance) were recorded. For each gene and recombination rate combination, 10,000 replicates were produced.

We also ran the same population model with a change of the selection parameter either at 500 generations or 896 generations before present (the latter being the CEU/CHB split time). The new selection parameters (s_x and s_y) are different for CEU and CHB populations, and are estimated from present CEU/CHB distributions (see Table 4.3). These simulations are hereafter called scenario V (selection change 500 generations before present) and VI (896 gpb).

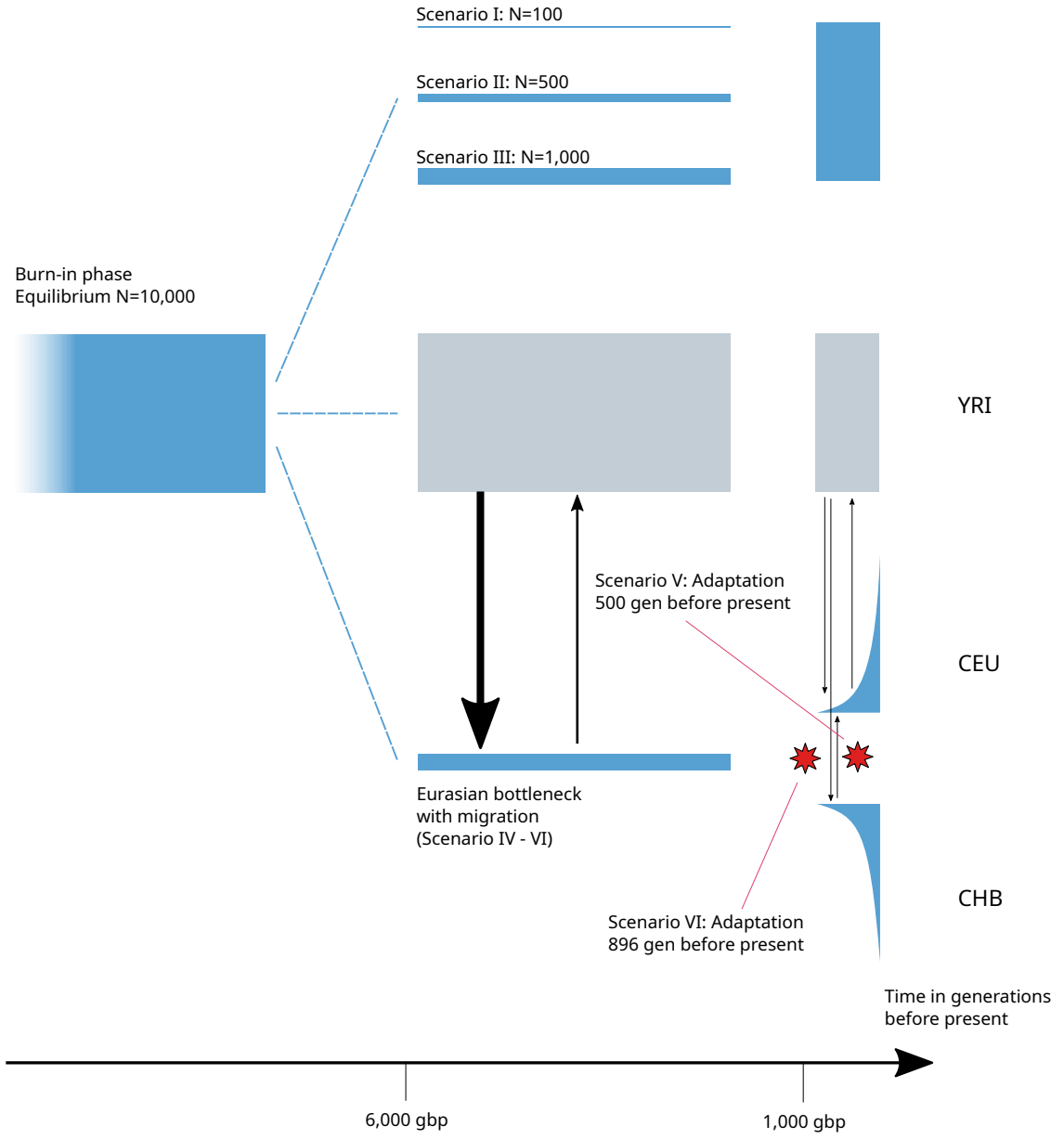


Figure 4.3: Illustration of simulated demography scenarios. The height of the blue boxes denote the population size. Scenario I-III cover a simple bottleneck lasting 5,000 generations, whereas scenario IV-VI are modifications of the GADMA-model with migration between subpopulations and change of selection parameters (indicated by red stars) in scenario V and VI. The gray color in the YRI population indicates, that it is at equilibrium.

4.3 Results and Discussion

In this study, we conducted an analysis of multicopy gene family evolution using a model that incorporates unequal recombination and selection. Our investigation aimed to examine the copy number changes observed in subpopulations of Europe, Asia, and Africa and to determine whether these changes could be attributed solely to demographic factors or if additionally selective pressures played a significant role. Our findings reveal that the observed copy number variations in several genes cannot adequately be explained by demographic processes alone, suggesting the involvement of selection.

Based on the data of (Brahmachary et al., 2014), we chose 42 gene families of intermediate copy number that show significant differences in their distribution among different populations (Table 4.1). This dataset relies on phase I of the 1000 Genomes project and – despite of been published almost a decade ago – turned out to be most suitable for our analysis. More recent data from the human pangenome project still lack phasing and a sufficiently large sample size across different populations (Liao et al., 2023).

Using equation (4.3), we ran the simple bottleneck scenarios I-III of in total 95 parameter combinations (r, s_x, s_y), to see whether population size changes can explain the differences in copy number distribution. Figure 4.4 shows mean gene copy number of 10,000 simulated bottleneck populations over time for one chosen gene (here: `PGA3`). Gray boxes indicate the centerd 50% quantile, white the 95% and the whiskers 99% quantiles. With strong bottleneck (reduction to $N = 100$ for 5,000 generations) and under low recombination and hence weak selection ($r = 0.001$, and $q_R = r/s_x$, $q_S = s_x/s_y$ constant) we find the widest variation among the 10,000 replicates. Higher r and stronger selection shift the mean value back to the one of YRI population, which was the basis of the parameter estimates. In this particular example, we find compared to YRI (black horizontal line) a higher mean copy number of `PGA3` in CHB (red line) and a lower mean value in CEU (blue line). It is the only gene in our chosen set, that shows opposite direction of significant mean copy number change. Only under strong bottleneck and low recombination these changes can be explained with same selection parameters as in the ancestral population and lay within the 99% quantile.

An overview of all 95 bottleneck results is given in Table 4.4. We consider scenario I – the strongest bottleneck – and the time point after 1,000 generations of recovery (i.e. first row of Figure 4.4, last boxplot of each column). If the mean or resp. $\mathcal{C}_V$ lays within the 95% quantile, we indicate non-significant differences with a 0. Significant changes are marked with a single $*$ ($\alpha = 5\%$) or double asterisk $**$ ($\alpha = 1\%$). As an example, we find for `PGA3` a mean value which is significantly smaller in CEU than in YRI (marked with –). With $r = 0.001$, this might be explained by a bottleneck (denoted by 0), whereas for $r = 0.002$ and $r = 0.005$ we find a significant difference ($**$) and the bottleneck explanation to be highly unlikely. Higher recombination $r = 0.01$ led to s_x values greater than 0.1 in CHB and YRI (see Tab 4.3) and hence was omitted.

From the candidates with a significant difference in mean or variance we selected the three genes coding for digestive enzymes, `AMY1A`, `SULT1A3`, `PGA3`, and the defense gene `DEFA1` for

Table 4.4: Results of bottleneck simulations. We ran simulations of scenario I (the strongest bottleneck with a reduction to $N = 100$) with parameters (r, s_x, s_y) estimated from the YRI-data and tested, whether after 1000 generations of recovery the mean and standard deviation σ of the CEU and CHB-data can be explained by a bottleneck. Blank space indicates that this parameter combination led to an s_x value out of the range of 0.001 and 0.1, and hence no simulation was run. The columns with 0,+ and – indicate whether there is a significant difference to the empirical data set (see Tab 4.1). In the r1 - r10 columns, a 0 indicates that the data can be explained by a bottleneck. * and ** show significant differences (5% and 1%) of the simulated and empirical data.

	CEU mean					CEU sd					CHB mean					CHB sd				
Gene		r1	r2	r5	r10		r1	r2	r5	r10		r1	r2	r5	r10		r1	r2	r5	r10
AMY1A	0	0	0	0	*	+	0	*	*	**	+	*	**	**	**	+	*	**	**	**
ANKRD20A3	+	**	**	.	.	0	0	0	.	.	+	**	**	.	.	0	0	0	.	.
BOLA2B	-	0	**	.	.	0	*	**	.	.	-	0	**	.	.	0	*	**	.	.
CBWD3	0	0	.	.	.	0	0	.	.	.	+	0	.	.	.	+	*	.	.	.
CDC37P1	+	**	**	**	**	+	*	**	**	**	0	0	0	*	**	+	*	**	**	**
CLEC18A	+	0	0	*	.	0	0	*	*	.	0	0	0	0	.	0	0	0	0	.
CSH	+	0	.	.	.	0	**	.	.	.	+	**	.	.	.	0	**	0	0	0
DEFA1	0	.	.	0	0	-	.	.	0	**	0	.	.	0	0	-	.	.	0	**
DEFB130	+	0	*	.	.	0	**	**	.	.	0	0	0	.	.	0	**	**	.	.
FAM72A	+	0	.	.	.	+	0	.	.	.	+	0	.	.	.	0	**	.	.	.
FAM75A1	0	0	0	.	.	0	0	0	.	.	+	*	**	.	.	+	0	**	.	.
FAM75A5	0	0	0	.	.	0	0	0	.	.	+	0	**	.	.	+	0	**	.	.
FCGBP	+	0	0	0	0	-	*	**	**	**	+	0	0	0	0	0	0	0	**	**
FOXD4L2	+	0	.	.	.	0	0	.	.	.	+	**	.	.	.	+	*	.	.	.
GOLGA6L9	0	0	0	.	.	0	0	0	.	.	+	0	*	.	.	0	0	0	.	.
GOLGA8G	+	*	**	.	.	0	0	0	.	.	+	0	0	.	.	0	0	0	.	.
GSUBP1	+	**	**	**	.	0	0	0	0	.	+	0	0	**	.	0	0	0	0	.
HIST2	+	0	.	.	.	0	0	.	.	.	+	*	.	.	.	0	0	.	.	.
LIMS3	-	*	.	.	.	0	**	.	.	.	-	0	.	.	.	0	**	.	.	.
LOC23117	0	0	0	.	.	0	*	*	.	.	-	*	**	.	.	0	*	**	.	.
LOC653606	0	0	.	.	.	0	0	.	.	.	-	**	.	.	.	+	0	.	.	.
MUC12	+	*	**	**	**	0	0	*	**	**	0	0	0	0	0	-	0	**	**	**
NBPF11	-	*	**	**	.	-	*	**	**	.	0	0	**	**	.	0	*	*	*	.
NBPF16	0	0	0	0	.	0	0	0	0	.	+	0	0	0	.	0	0	0	0	.
NPIP	-	**	**	.	.	0	*	*	.	.	-	**	**	.	.	0	*	*	.	.
PGA3	-	0	**	**	.	+	0	0	0	.	+	*	**	**	.	0	0	0	0	.
PPIAP21	+	**	**	.	.	0	0	0	.	.	+	**	**	.	.	0	0	0	.	.
PRAMEF14	+	*	**	.	.	+	**	**	.	.	+	**	**	.	.	+	0	**	.	.
PRAMEF20	0	0	.	.	.	0	0	.	.	.	+	0	.	.	.	+	0	.	.	.
PRAMEF5	-	**	**	.	.	+	*	*	.	.	-	**	**	.	.	+	*	**	.	.
PRAMEF8	0	0	0	.	.	+	0	**	.	.	0	0	0	.	.	0	0	0	.	.
PRR11	+	**	**	.	.	0	0	*	.	.	+	**	**	.	.	0	**	**	.	.
PRR20A	-	.	.	0	**	-	.	.	0	*	-	.	.	0	**	0	.	.	0	0
PSG3	+	0	*	.	.	0	0	0	.	.	0	0	0	.	.	+	0	*	.	.
RGPD1	0	0	.	.	.	+	0	.	.	.	0	0	.	.	.	+	0	.	.	.
SPYDE3	-	**	**	.	.	-	0	0	.	.	-	**	**	.	.	0	0	0	.	.
SULT1A3	0	0	0	0	.	0	0	*	*	.	-	0	**	**	.	0	**	**	**	.
TBC1D3	-	**	**	**	**	0	0	0	0	0	-	**	**	**	**	+	0	*	**	**
TCEB3C	-	0	**	**	**	0	0	0	0	0	-	0	**	**	**	0	0	0	0	0
TP53TG3	0	0	0	0	0	0	0	0	0	0	-	0	**	**	**	0	0	0	*	*
TRIM49L1	+	**	**	.	.	0	0	0	.	.	+	**	**	.	.	0	0	0	.	.
ZNF658B	+	0	**	.	.	0	0	0	.	.	+	**	**	.	.	+	0	0	.	.

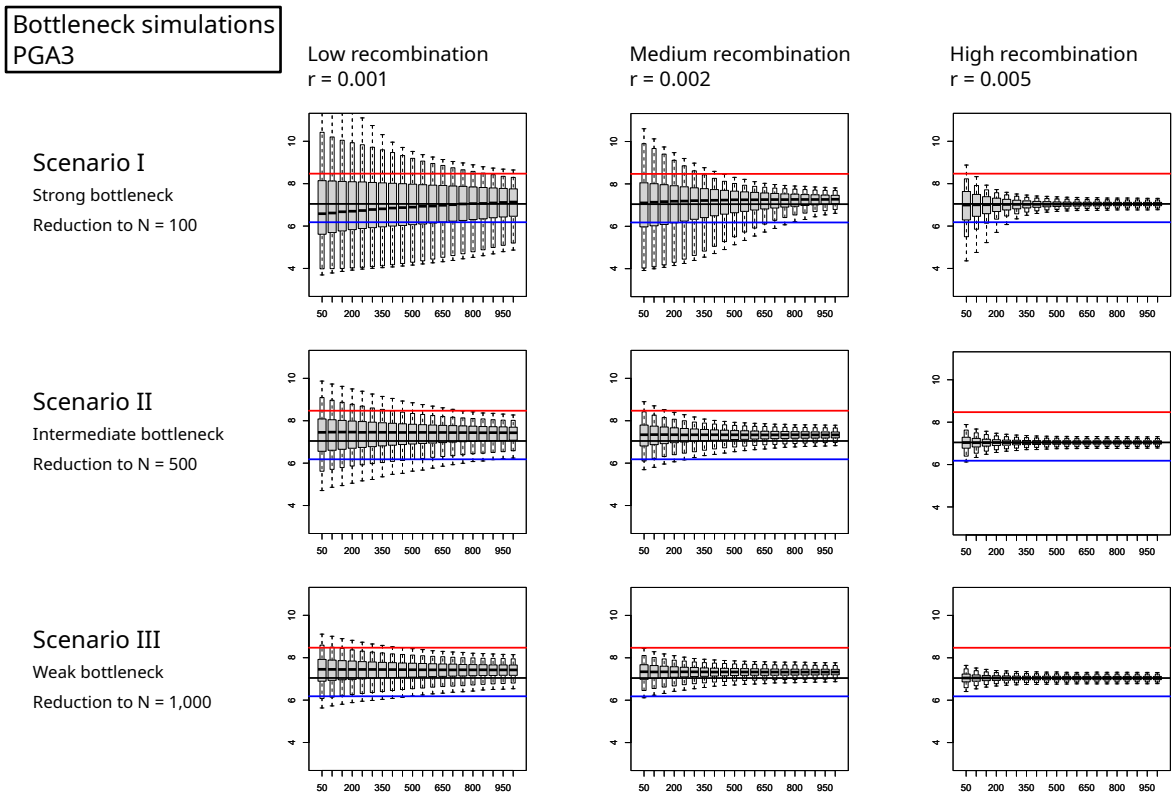


Figure 4.4: Mean copy number over time. After population reduction to $N = 100, 500, 1000$ (top to bottom) we traced the mean-value of 10,000 replicates over time (x-axis in generations). The input parameters s_x, s_y were estimated for $r = 0.001, 0.002, 0.005$ (left to right) from the YRI-data set for the candidate gene `PGA3` (see Tab 4.3) and kept constant over time, to see the effect of the bottleneck and recovery. Whiskers mark the 99% quantile, the white box the 95% quantile. Horizonal lines mark the values from the original data set of (Brahmachary et al., 2014) (black: YRI, red: CHB, blue: CEU).

a more detailed analysis and tested the GADMA demography without and with selection change according to the estimates from regression (scenario IV-VI).

Figure 4.5 shows mean copy number and coefficient of variation $\mathcal{C}_V$ at present, simulated according to scenarios I,IV,V and VI for 10,000 replicates each. As in the simplified bottleneck scenario, scenario IV with subpopulation migration returns to values of basis data set YRI. Hence, under high recombination and strong selection the mean and standard deviation values of CHB and CEU show significant differences to the simulations.

However, with a change of selection the data of scenario V and VI show a different pattern. With new s_x and s_y estimated for subpopulations according to equation (4.3), the mean and $\mathcal{C}_V$ are shifted and the empirical data lies often within the 95%- and 99%-quantiles of the simulated data distributions. Deviations are a result of the sensitivity of the logarithmic regression. We observe no strong difference between V and VI, suggesting that 500 generations represent a sufficiently large time span to reach a new equilibrium.

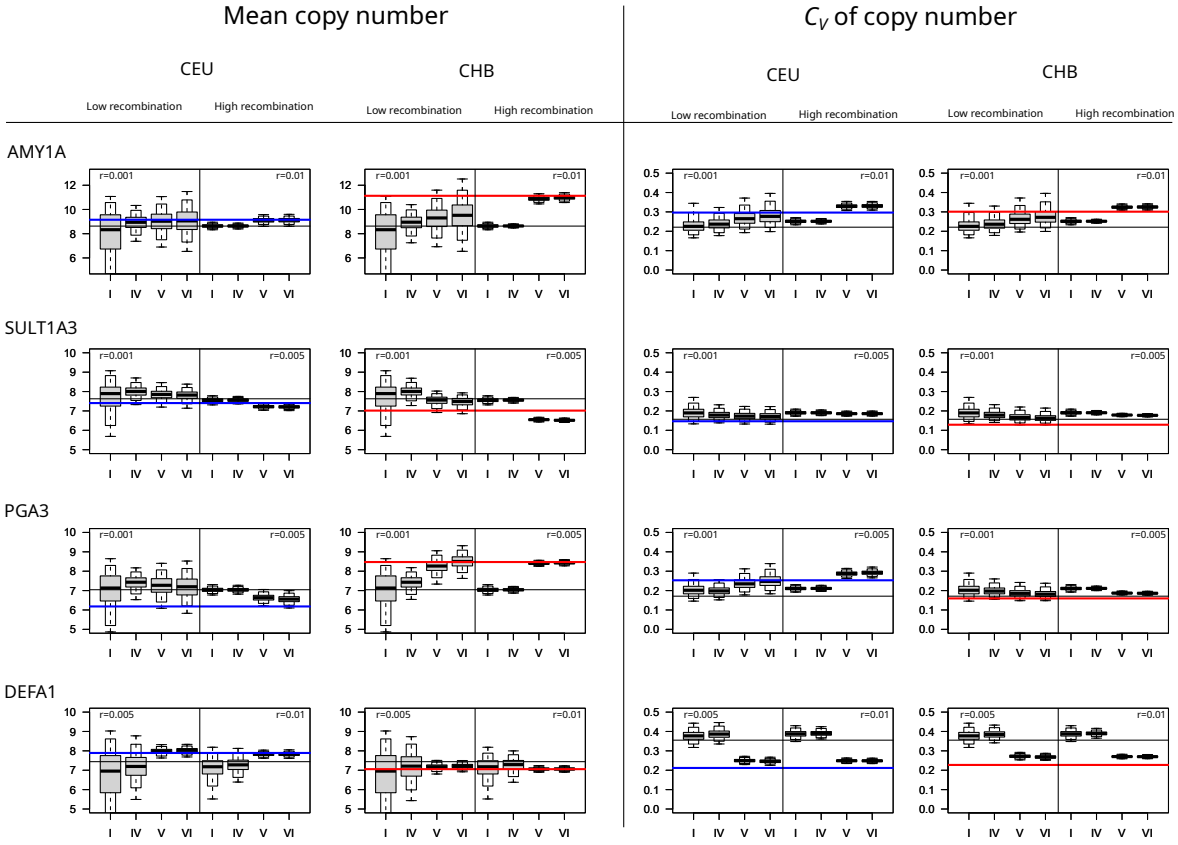


Figure 4.5: Comparison of mean copy number $\bar{y}$, and coefficient of variation $\mathcal{C}_V = \sigma/\bar{y}$ in four scenarios I,IV,V,VI for candidate genes `AMY1A`, `SULT1A3`, `PGA3`, `DEFA1`. In scenario V and VI the initial selection coefficients of YRI (see Table 4.3) were changed 500 resp. 986 generations before present. Simulation results are shown for lowest and highest recombination rate. Black lines refer to mean and $\mathcal{C}_V$ of the experimental data in YRI, blue to CEU and red to CHB.

Hence, the differences of the four candidate genes may be explained by adaptive processes and different selective pressures.

The `AMY1A` gene, which encodes amylase, an enzyme that breaks down starch, has strongly increased mean and σ in the Asian population, likely linked to adaptations to high grain intake. In the European population, while the variation is increased, the change in mean copy number is small. These findings are in agreement with results of several studies that indicate that individuals from populations with high-starch diets have, on average, more gene copies than those with traditionally low-starch diets (Perry et al., 2007; Pajic et al., 2019; Atkinson et al., 2018). Under our model selection strength is relaxed in CEU and CHB, such that a higher copy number is not selected against and a more widespread distribution of CNV can evolve. A recent study (Inchley et al., 2016) has suggested a more complicated model of Amylase evolution, involving two steps: an expansion from one to several copies after the human-Neanderthal split, but before separation of modern human populations, and a subsequent shift of the optimal gene copy number, independently in different populations.

This study also suggests that increase of AMY1 copy number occurred in South America even more dramatically than in East Asia, a hypothesis which should be tested in the framework of our model as soon as suitable data become available.

SULT1A3 is a gene in the SULT (sulfotransferases) family, which catalyze sulfation of a variety of substrates, especially catecholamines including dopamine and epinephrine (Brix et al., 1999; Dajani et al., 1999). Polymorphisms in SULT1A3 and SULT1A4 have been shown to affect metabolism of therapeutic drugs (eg., (Hui and Liu, 2015; Bairam et al., 2019)), and these genes have therefore been studied extensively in the framework of medico- and pharmacogenetics (Thomae et al., 2004; Hildebrandt et al., 2004). In the analyzed dataset, it has reduced mean copy number in Asia but not in Europe. There have been a few studies on copy numbers of SULT1A3/4 genes. Hildebrandt et al. (2004) first noted possible duplication of SULT1A3 and identified a duplicated copy in all four different human populations. More recently, a study of 172 human individuals discovered variable SULT1A3/4 copy numbers from 1 to 10, and associated its copy number with risk and onset of neurodegenerative disease (Butcher et al., 2017). Note that SULT1A3 and SULT1A4 are closely related paralogs that are often difficult to distinguish, and studies on copy number usually put them together.

PGA3 (Pepsinogen, precursor for pepsin, an enzyme that breaks down protein to smaller peptides) is associated with prostate-specific antigen production. It is the only gene in our list to have opposite changes in two derived populations: its mean copy number increases in Asia and decreases in Europe. As Asian and European humans share most of the same bottleneck period, the diverging copy number distribution is highly unlikely to be a demographic effect, and complex selection patterns are needed to explain the data. The copy number variation on the Pepsinogen (PGA) locus was originally discoverd with electrophoresis and three individual genes (named PGA 3, 4, 5) were initially found (Taggart et al., 1985). Pepsinogen genes have been shown to duplicate and become lost recurrently in vertebrates (Castro et al., 2014). The pepsinogen genes were also shown to have variable expression level in tumor cells, particularly a reduction of PGA expression in esphagael, stomach and thyroid cancers (Shen et al., 2020). This could be an additional source of selective pressure besides protein metabolism. While the simplest explanation is that dietary differences between Asian and European populations during the spread of of agriculture (in the last 5000-10000 years) is the driver of PGA copy number changes, alternative hypotheses involving tumor suppression or interaction with other enzymes must be considered.

Finally, we analyzed the immune gene Alpha-defensin DEFA1. It codes for defensins, proteins that are involved in innate (non-learned) immunity, specifically in antimicrobial defense against a broad spectrum of microorganisms, including bacteria, fungi, and viruses. DEFA1 shows a decrease in variance in both Asia and Europe, indicating stronger selective pressures. More precisely, when considering the distribution in Figure 4.1, one observes four individuals in YRI population with high copy number which indicates a relaxed selective pressure in Africa. Alpha-defensins are expressed in neutrophil cells and intestinal epithelial cells, acting as microbiocidal agents (Ganz et al., 1985; Ayabe et al., 2000; Nassar et al., 2007). The genes DEFA1 and DEFA3 code for some of the Alpha-defensins (HNP1/2/3), and appear to be "in-

terchangeable variant cassettes" within a tandem array of 19kb (Aldred et al., 2005). Copy number variation of `DEFA1` is present in all apes including gibbon, but the version identified as `DEFA3` is human-specific; the copy number is also demonstrated to affect expression level (Aldred et al., 2005). Low copy number of `DEFA1/3` is shown to be associated with hospital-acquired infection (Zhao et al., 2018) as well as kidney diseases (Ai et al., 2016). On the other hand and counterintuitively, a high copy number of `DEFA1/3` may lead to more severe cases of sepsis (Chen et al., 2010, 2019) and is associated with Crohn's Disease (Jespersgaard et al., 2011), and thus selected against. The trade-off between infective and autoimmune diseases could lead to selection towards an intermediate copy number of Alpha-defensins. Therefore, our results suggest a possibility that the out-of-Africa expansion is accompanied by such a change in environmental pathogen diversity that a delicately tuned dosage of defensin is required. This can be corroborated by the fact that YRI has a few individuals with very high (outliers) copy numbers of `DEFA1`, which can not be found in CHB or CEU.

In conclusion, while both demographic effects and shifts in selection schemes can result in changes in copy number distributions, in some of our candidate genes the former is not sufficient to explain the observation. Adaptive processes can induce new relationships between copy number and fitness, and impact the resulting copy number distribution. Importantly, changes in the strength, or direction of selection may become manifest not only in mean copy number, but also in the variance or compound statistics, such as the coefficient of variation.

5 The structured coalescent in the context of gene copy number variation

Authors: Moritz Otto and Thomas Wiehe
Status: Published in Theoretical Population Biology, August 2023

DOI: https://doi.org/10.1016/j.tpb.2023.08.001

Abstract

The *Structured Coalescent* was introduced to describe the coalescent process in spatially subdivided populations with migration. Here, we re-interpret migration routes of individuals in the original model as "migration routes" of single genes in tandemly arranged gene arrays. A gene copy may change its position within the array via unequal recombination. Hence, in a coalescent framework, two copies sampled from two chromosomes may coalesce only if they are at exactly homologous positions. Otherwise, one or multiple recombination events have to occur before they can coalesce, thereby increasing mean coalescence time and expected genetic diversity among the copies in a gene array.

We explicitly calculate the transition probabilities on these routes backward in time. We simulate the structured coalescent with migration and coalescence rates informed by the unequal recombination process of gene copies. With this novel interpretation of population structure models we determine coalescence times and expected genetic diversity in samples of orthologous and paralogous copies from a gene family. As a case study, we discuss the site frequency spectrum of a small gene family in the two scenarios of high and of no gene copy number variation among individuals. These examples underline the significance of our model, since standard test-statistics may lead to misinterpretations when analyzing sequence data of multi-copy genes due to their different expected genetic diversity.

5.1 Introduction

Duplicated genes can make up a large portion of the entire gene complement of many eukaryotes. For instance, in human around 40% of the genes are considered to be duplicates (Zhang, 2003), in plants on average about 65% (Panchy et al., 2016). Among the largest known gene families are those of the NLR immune receptors in some animals (about 400 members in *D. rerio* (Howe et al., 2016)) and the evolutionarily related NBS-LRRs in some plants (Jones et al., 2016). There are several mechanisms which generate new duplicates (Panchy et al., 2016; Ohno, 1970; Magadum et al., 2013). One mechanism is whole genome duplication, another one is tandem duplication. Only the latter is expected to affect a limited number of genes and to potentially generate large numbers of copies which are clustered on a chromosome. With many species now being re-sequenced at (sub-) population level, analysis of gene copy number variation (gCNV) is gaining growing attention. As repeatedly reported, gCNV of signal receptors or of immunity genes is involved in adaptive responses to both biotic (e.g., pathogens) and abiotic (e.g., heat, drought) stresses (Wan et al., 2021; Kondrashov, 2012; Qian and Zhang, 2014). Thus, population genetic insights should contribute to understanding the basis of such adaptive mechanisms.

One of the persisting problems, even with modern sequencing technology, is to uniquely identify the members of a large gene family and to correctly map them to a reference genome. This implies difficulties not only in correctly counting copy number for a given individual, hence in the construction of presence/absence matrices, but also in computing meaningful measures of genetic diversity within copies. One way to overcome this problem with a bioinformatic approach is to abandon the concept of a reference genome altogether, replacing it by a pan-genome graph which integrates – in particular – the structural variation present in a species (Golicz et al., 2020; Hübner, 2022). Still, this would not solve the population genetic problem of measuring diversity within genes, as long as these measures depend on a clear distinction between paralogs (duplicates within the same genome) and orthologs (related genes in different species / individuals).

However, acknowledging this difficulty, one may try to overcome it by integrating both orthologs and paralogs in a joint framework, as suggested here. Focusing on the evolution of tandemly arrayed genes, we combine a recent model of gCNV driven by unequal recombination (Otto et al., 2022) with the classical ideas of the structured coalescent (Takahata, 1988; Wakeley, 2001). This describes genetic distance of individuals in spatially separated islands and under different regimes of migration. One key insight is that any two lineages may only coalesce if drawn from the same island. Lineages (i.e. individuals) from different islands have to migrate before they can coalesce. A large volume of theoretical literature is dedicated to studying coalescent times and their limiting properties in simple symmetric and asymmetric island layouts (Notohara, 1990; Nordborg, 1997; Austerlitz et al., 1997; Wilkinson-Herbots, 1998).

Here, we re-visit results of this ground-breaking work and put the structured coalescent in a new context: considering a single gene copy, which is located in a tandem gene array (see

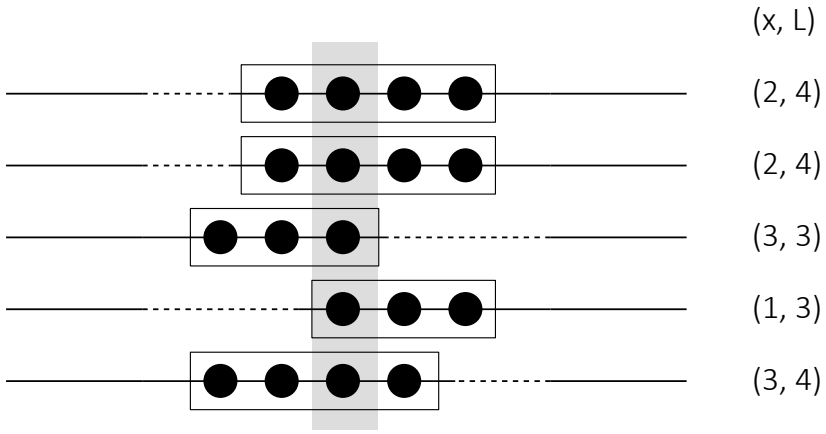


Figure 5.1: Consider an alignment of five chromosomes containing arrays of different numbers of gene copies. What is the expected pairwise genetic diversity, if the relative positions of the genes with respect to their arrays, shown in parentheses, are taken into account?

Figure 5.1), it may change its position in the array by unequal recombination (see Figure 5.2). Hence, in a coalescent framework, two copies of different chromosomes can only coalesce, if they are both in the same array position. We may call such copies orthologous. Otherwise, they are 'paralogous' and one or more recombination events have to bring them to the same position before they can coalesce. Thereby coalescence time is increased. Taking into account this possibility, we reconsider the question of genetic diversity: depending on the recombination rate, genetic diversity of paralogs can be substantially increased compared to orthologs. We explicitely calculate the transition probability of a single gene copy along its gene array backward in time. We simulate the structured coalescent using `msprime` (Baumdicker et al., 2022) with migration and coalescence rates according to the unequal recombination process of gene copies.

Furthermore, we consider samples larger than two, and explore how the site frequency spectrum and the common summary statistics, such as Tajima's D (Tajima, 1989), are affected when orthologs and paralogs are analyzed jointly. We close with an application, inspired by the HMA4 gene array in *Arabidopsis halleri*.

5.2 Methods and Model

Transition probabilities Consider a diploid population of size N. Let each individual carry a varying number of gene copies of a particular gene and assume them to be tandemly arranged. For a given chromosome denote the length of its gene array (i.e. the number of gene copies) by ℓ. Under neutrality and unequal recombination, as described in (Otto et al., 2022), the offspring at generation $t+1$ is generated from the population at generation t as follows:

For each i in $1, ..., 2N$:

- Choose two arrays of sizes $\ell_1, \ell_2 \geq 1$ from the current population
- Decide whether a recombination event happens. If so, choose two break points $B_1 \sim U(1, ..., \ell_1)$ and $B_2 \sim U(1, ..., \ell_2)$
- Splice head and tail of the split arrays together and propagate one of the resulting arrays to the next generation

For $N = \infty$ and under neutrality array size ℓ is discrete-Gamma distributed at equilibrium (Otto et al., 2022). It has shape $\kappa = 2$ and expectation which is identical to the initial mean array size L since the recombination process is symmetric.

Finite population sizes introduce an additional layer of stochasticity and an analytic representation of the compound process and its stationary density is not known. However, simulations suggest that the stochastic effect of drift is small compared to the effect of unequal recombination already with moderately large population size. In the following, we assume a sufficiently large population, which is in equilibrium, such that gene array lengths are sampled from a discrete $\mathcal{G}(2, L)$ distribution, i.e.

$$\ell \sim \mathcal{G}(2, L), \qquad p^{\mathcal{G}(2,L)}(k) := \text{Prob}[\ell = k] = \frac{(e^{2/L} - 1)^2}{e^{2/L}} \cdot k \cdot e^{-2/L \cdot k}. \tag{5.1}$$

Now focus on one particular copy in an array and trace its position backward in time. Note that from now on we analyze the haploid genealogy. Encode the copy's current position by (x, y), where x designates its position from the head of the gene array and y from the tail such that the total length of the gene array is $\ell = x + y - 1$.

Change of position may occur by unequal recombination. To trace such events, decide whether the array in the current generation was produced by recombination of head and tail from different arrays in the previous generation. If so, draw a random variable $H \sim Ber(\frac{x}{x+y-1})$ to decide whether the particular copy at position (x, y) was propagated by the head or tail part. Choose a break point $B_1 \sim U(1, ..., \ell_1 = x + y - 1)$, draw a second gene array of size $\ell_2 \sim \mathcal{G}(2, L)$ from the population and a break point $B_2 \sim U(1, ..., \ell_2)$. Finally, fuse the two parts together (see Figure 5.2).

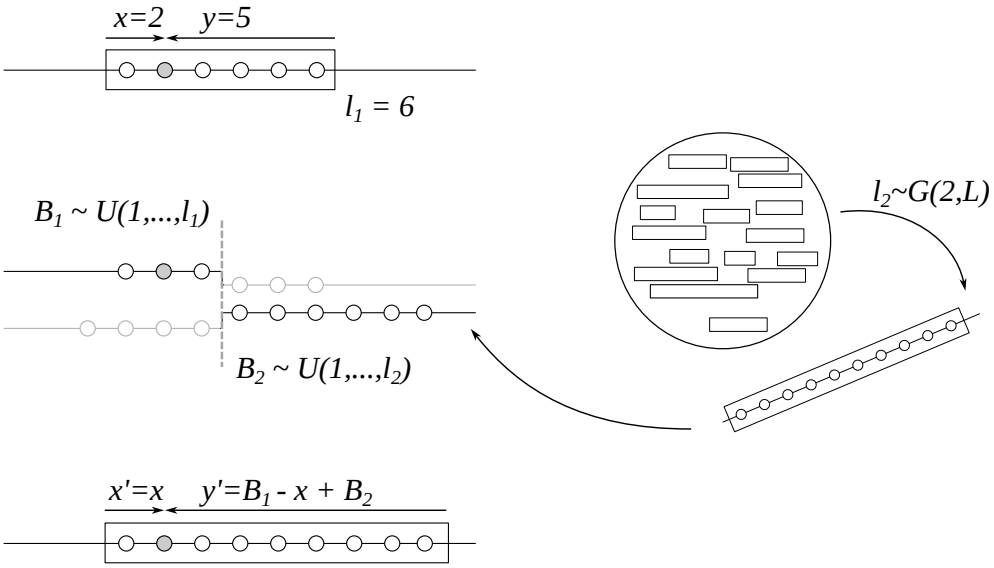


Figure 5.2: Illustration of the position change process backward in time of a gene copy. Given a copy (grey circle) at position (x, y) in a gene array, it may have been originated from two recombinants. The breaking point, that seperates the head and tail is marked as B_1. The part on which the copy is located is maintained (here: head) and fused with the counterpart (tail) of a gene array sampled from the equilibrium distribution $\mathcal{G}(2, L)$. Hence, the position of the copy changes to (x', y'), where x' remains constant and y' is the concatenation of the remaining part of the head $(B_1 - x)$ and the tail from the chosen recombinant (B_2).

Note that either the head or the tail is maintained, and hence x or y remains constant. In terms, we find:

$$
\begin{aligned}
&H \sim \mathrm{Ber}(\frac{x}{x+y-1}), B_1 \sim U(1,...,\ell_1 = x+y-1), \ell_2 \sim \mathcal{G}(2,L), B_2 \sim U(1,...,\ell_2) \\
&\text{if } H = 0 \\
&\quad | \quad x' = x \text{ and } y' = |B_1 - x| + B_2 \\
&\text{if } H = 1 \\
&\quad | \quad x' = |B_1 - x| + B_2 \text{ and } y' = y
\end{aligned}
\tag{5.2}
$$

Proposition 1. *Using the notation $X = |B_1 - x| + B_2$, we find*

$$
P[X = k|x, y, L] = \frac{1}{Z} \cdot \left(2e^{-2/L \cdot k} - e^{-2/L \cdot c_1} - e^{-2/L \cdot c_2}\right), \tag{5.3}
$$

where

$$
Z = (x+y) \cdot \frac{1 - e^{2/L}}{e^{2/L} - 1}, c_1 = \max(0, k-x), c_2 = \max(0, k-y)
$$

Proof. This is a straightforward computation after finding that B_2 is geometrically distributed and $|B_1 - x|$ is the sum of two uniform distributions. □

Therefore, we can interprete the position change of a gene copy in its array as a Markov process on an $\mathbb{N}\times\mathbb{N}$-lattice with either horizontal or vertical moves and transition probabilities given by

$$\begin{aligned} P^L[(x,y)\to(x,y)] &= 1 - r_{(x,y)} \\ P^L[(x,y)\to(x,k)] &= r_{(x,y)}\cdot\frac{y}{x+y}P[X=k|x,y,L] \\ P^L[(x,y)\to(k,y)] &= r_{(x,y)}\cdot\frac{x}{x+y}P[X=k|x,y,L], \end{aligned} \tag{5.4}$$

where $r_{(x,y)}$ denotes the (unequal) recombination rate. We distinguish the two cases, where $r_{(x,y)} = r$ constant and does not depend on its current state and $r_{(x,y)} = r_0\cdot\ell$, where the recombination rate scales linearly with the size of the gene array. This is motivated by the fact that under a constant rate per nucleotide the number of possible recombination points increases with additional gene copies.

Stationary distribution The above described Markov process is irreducible and aperiodic and therefore provides a stationary distribution π on $\mathbb{N}\times\mathbb{N}$ that satisfies $\pi P = \pi$, i.e. for all $x, y \in \mathbb{N}$:

$$\begin{aligned} \pi(x,y) &= \sum_{(a,b)}\pi(a,b)P[(a,b)\to(x,y)] \\ &= \sum_k \frac{k}{x+k}P^L[X=y|x,k] + \frac{k}{k+y}P^L[X=x|k,y] \\ \text{and} \quad & \sum_{(x,y)}\pi(x,y) = 1. \end{aligned} \tag{5.5}$$

Note, that the process and therefore the stationary distribution only depend on the initial mean gene array length L of the population, from which the recombinant was generated. To determine π, we reduce the system to a finite state space from $\mathbb{N}\times\mathbb{N}$ to $n_0\times n_0$, i.e. neglecting the possibility for a gene copy to reside in an array of size $L > n_0$. Let n_0 be the 99.9% quantile of the discrete Gamma distribution (5.1), i.e.

$$n_0 = \min\left\{ n \quad | \quad \sum_{k=1}^{n} p^{\mathcal{G}(2,L)}(k) > 0.999 \right\}. \tag{5.6}$$

With this reduction to finiteness we find a numerical solution of π. Given π we can determine the *mixing time* of the Markov process, i.e. the time it takes for the process to converge to its stationary distribution. Given a transition matrix P on the state space $\mathcal{S}$ the distance $d(t)$ at time t to its stationary distribution is defined as the total variation $\|\cdot\|_{TV}$, i.e.

$$\begin{aligned} d(t) &= \max_{s_0\in\mathcal{S}} \|P^t_{s_0} - \pi\|_{TV} \\ &= \max_{s_0\in\mathcal{S}} \sum_{s_1} \frac{1}{2}|P^t_{s_0}(s_1) - \pi(s_1)| \end{aligned}$$

Consistent with literature (see (Levin and Peres, 2017)), the mixing time t_{mix} is defined as the time, at which $d(t)$ is less than $\frac{1}{4}$, i.e.

$$t_{mix} = \min\left\{t \quad | \quad d(t) < \frac{1}{4}\right\}, \tag{5.7}$$

where $\frac{1}{4}$ is chosen as threshold in order to satisfy the equation $d(s \cdot t_{mix}) \leq 2^{-s}$. Since π is numerically determined, one can determine t_{mix} numerically as well.

Furthermore, we aim to determine the probability that two particles that independently move on the $\mathbb{N} \times \mathbb{N}$-lattice meet at the same position. Therefore, we define the distance

$$D_{\left(\binom{x_1}{y_1},\binom{x_2}{y_2}\right)} = \max\left\{|(x_1 - x_2|, |y_1 - y_2|\right\}. \tag{5.8}$$

Starting with any tuple of initial positions we trace their relative distance over time t. Since the particle trajectories do not affect each other, the process $D(t)$ also is a Markov process with a unique stationary distribution. We are interested in determining the probability $\text{Prob}(D_\infty = 0)$, representing the event that two particles are eventually at the same position.

Structured Coalescent Recall the definition of the structured coalescent, with notations as in Wilkinson-Herbots (1998). Consider a haploid population of size N divided into a finite or infinte number of subpopulations which are all large and panmictic. Denote the label set of these islands by $\mathcal{S}$ and the population size in island $i \in \mathcal{S}$ as N_i, such that $\sum_{i\in\mathcal{S}} N_i = N$. The individuals reproduce and migrate independently in non-overlapping generations under the assumption of constant population size and hence migration equilibrium. At a particular generation draw a sample of k individuals and trace their ancestry. Let $\alpha_i(t)$ be the number of distinct ancestors of this sample which are located in subpopulation $i \in \mathcal{S}$ at t generations ago. Define ε^i to be the vector with 1 at position i and zeros otherwise, i.e. $(\varepsilon^i)_j = \delta_{i,j}$. Then, $\alpha(t)$ may change to one of the following:

- $\alpha(t+1) = \alpha(t) - \varepsilon^i$, if a coalescent event happens in subpopulation i, or
- $\alpha(t+1) = \alpha(t) - \varepsilon^i + \varepsilon^j$, if an individual migrates from i to j.

Furthermore, define M_{ij} to be the migration rate from island i to j and $1/c_i$ to be the coalescence rate in subpopulation i, which depends on the subpopulation size N_i. Under reasonable assumptions about reproduction and migration (see (Wilkinson-Herbots, 1998)), the ancestral process $\alpha(t), t \geq 0$ is well approximated by the continuous time Markov process defined by the Q matrix

$$\mathcal{Q}_{\alpha\to\beta} = \begin{cases} -\sum\limits_{i\in\mathcal{S}} \alpha_i \frac{M_i}{2} + \frac{1}{c_i}\binom{\alpha_i}{2} & \text{if } \beta = \alpha \\ \alpha_i \frac{M_{ij}}{2} & \text{if } \beta = \alpha - \varepsilon^i + \varepsilon^j \\ \frac{1}{c_i}\binom{\alpha_i}{2} & \text{if } \beta = \alpha - \varepsilon^i \\ 0 & \text{else} \end{cases}, \tag{5.9}$$

where $M_i = \sum_j M_{ij}$. Note, that this process assumes constant coalescence rates $1/c_i$, i.e. a population which is in migration equilibrium. With sufficiently large population size, this is a reasonable restriction.

From now on let $k = 2$, i.e. focus on the coalescent process of two lineages. Denote by T_{ij} the coalescence time of two lineages that are at time $t = 0$ located at islands $i, j \in \mathcal{S}$. Then, as shown in Wilkinson-Herbots (1998), the Laplace transformation of the coalescence time distribution, i.e.

$$\varphi_{ij}(s) = E[e^{-s \cdot T_{ij}}],$$

for $s \geq 0$ can be determined by solving the following linear equation system:

$$\begin{cases} \left(\frac{1}{c_i} + M_i + s\right)\varphi_{ii}(s) - \sum_{k \neq i} M_{ik}\varphi_{ik}(s) = \frac{1}{c_i} & \text{, for } i \in \mathcal{S} \\ \left(\frac{M_i}{2} + \frac{M_j}{2} + s\right)\varphi_{ij}(s) - \sum_{k \neq i} \frac{M_{ik}}{2}\varphi_{jk}(s) - \sum_{k \neq j} \frac{M_{jk}}{2}\varphi_{ik}(s) = 0 & \text{, for } i \neq j \in \mathcal{S} \end{cases} \tag{5.10}$$

In terms of the unequal recombination process (see equation (5.4)), we find

- state space $\mathcal{S} = \mathbb{N} \times \mathbb{N}$, which encodes the position (x, y) in gene array
- migration rates $M_{(x,y),(\cdot,\cdot)} = P^L[(x, y) \to (\cdot, \cdot)]$, which are the position changes in the gene array
- coalescence rates $\frac{1}{c_{(x,y)}} = \frac{1}{N_{(x,y)}}$, where $N_{(x,y)}$ is the equilibrium number of individuals with gene array size $\ell = x + y - 1$, i.e.

$$N_{(x,y)} = N \cdot \frac{(e^{2/L} - 1)^2}{e^{2/L}}(x + y - 1)e^{-2/L \cdot (x+y-1)} \tag{5.11}$$

  Hence, the sum of all subpopulation islands equals $\sum N_{x,y} = L \cdot N$, the total number of gene copies in the population.

An illustration of the process is shown in Figure 5.3. Note, that gene copies are fully described by their position within the gene array. Whether two copies are located on the same or different haplotypes does not affect their ability to coalesce. Still, when tracing the trajectory of a small sample of gene copies, we assume them to change their positions independently. In other words, they move as particles in the Markov process of the structured coalescent in state space $\mathcal{S}$.

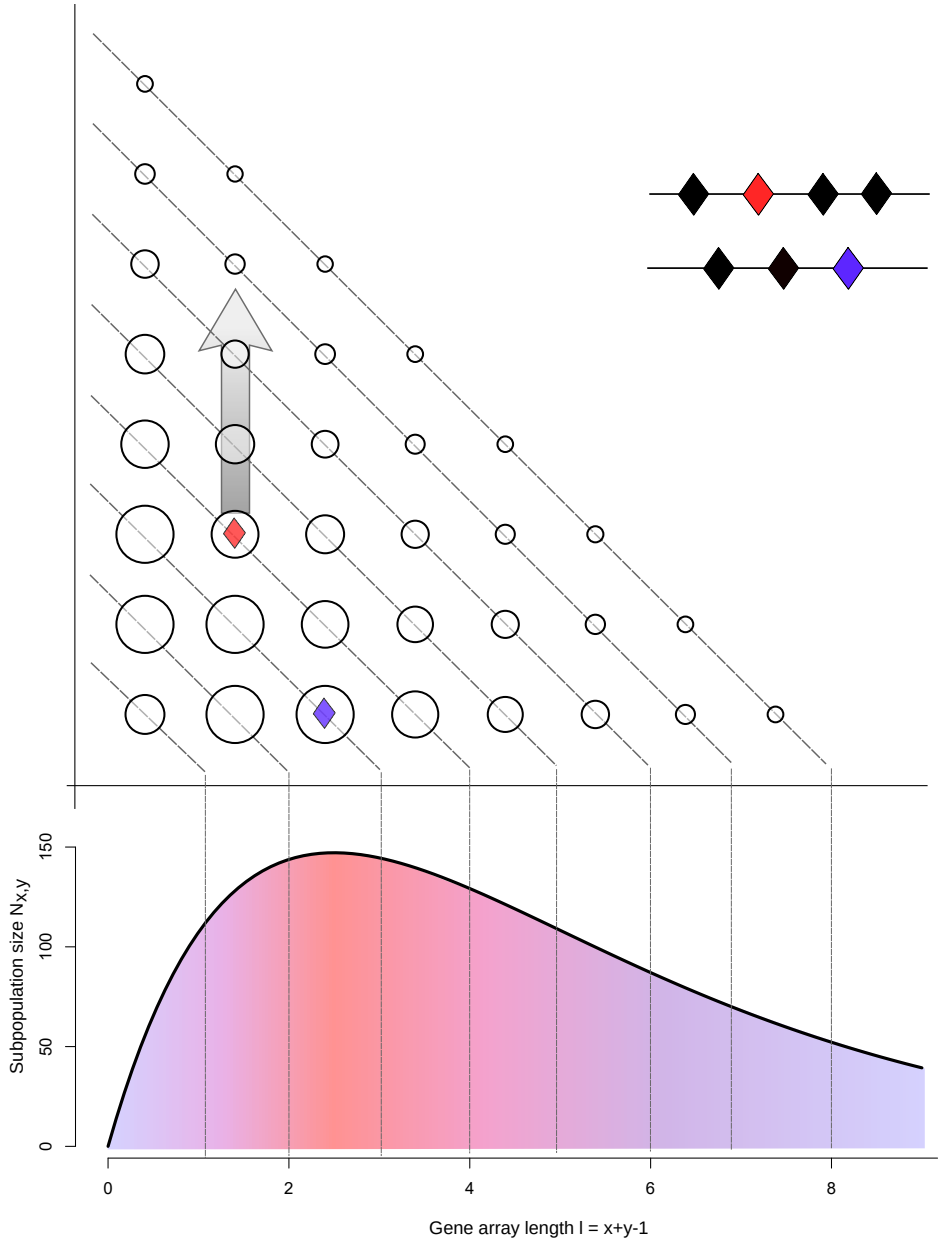


Figure 5.3: Example of $k = 2$ gene copies in the $\mathbb{N} \times \mathbb{N}$-lattice. The coalescence rates are defined as $1/N_{x,y}$, which is the subpopulation size in the equilibrium population with $\ell = x + y - 1$ copies (bottom figure). The genes may change their position either horizontally or vertically according to unequal recombination (see black arrow).

msprime To obtain numerical results we resorted to simulations with `msprime` (Baumdicker et al., 2022) of the structured coalescent with underlying migration and coalescence rates given by the unequal recombination process.

Although some simpler migration dynamics may be tackled in terms of the eigenvectors of the transition matrix, the dynamics defined in equation (5.4) on a state space $\mathcal{S} = \mathbb{N} \times \mathbb{N}$ is not easily decomposed in its eigenvalues. Also, a numerical solution of equation (5.10) is not feasible due to the infinite state space. Even when truncating to $n_0 \times n_0$, as in (5.6), the number of unknown parameters is of order $(n_0)^4$. As an example, for $L = 5$ we find $n_0 = 23$, which results in $23^4 \approx 280,000$ possible pairs of starting points and hence unknown variables $\varphi_{ij}(s)$ for any given $s \geq 0$.

We ran `msprime` with a `python` wrapper-script that takes as input the finite state space, a migration matrix and subpopulation sizes and simulates for a given set of starting point tuples $10,000$ coalescence events. To test the validity of the procedure, we ran the code on two examples, for which a numerical solution was known and compared expected and simulated outcomes. The details of this validation procedure on a symmetric and a continental island model are given in the Appendix. Also, we uploaded a jupyter-notebook on github, as reference for the details of the simulations[1]

Finally, we analyzed the unequal recombination structured coalescent, see equation (5.4). The parameter range was chosen as $L = 5, 10$, $N = 1000, 5000, 10000$ and $r = 1\%, 5\%$ constant and $r_{(x,y)} = r_0 \cdot \ell$, where $r_0 = 1\%$ and $\ell = x + y - 1$. We used a 5×5 sample grid spanning $n_0 \times n_0$, i.e. $\{1, 5, 9, 13, 17\}^2$ (case $L = 5$) and $\{1, 8, 16, 23, 31\}^2$ (case $L = 10$), leading to 625 starting pairs $\left(\binom{x_i}{y_i}, \binom{x_j}{y_j}\right) \in \mathcal{S}$.

Site Frequency Spectrum We explored in detail the following two extreme cases. In scenario 1 consider five chromosomes, all with three gene copies, leading to a total copy sample of $k = 15$. We assume them to be drawn from a population of size $N = 80,000$ and mean gene array length $L = 3$. These parameters are motivated by an example from the plant *Arabidopsis halleri* in which we analyzed the three copies of the heavy metal ATPase4 (HMA4) gene. They encode a Zinc and Cadmium pump which facilitates root-to-shoot transport of these metals (Hanikenne et al., 2008; Roux et al., 2011; Briskine et al., 2016).

In scenario 2 we also consider $k = 15$ gene copies from five chromosomes, but all with different gene array sizes ranging from 1 to 5. Hence, in the sense of the structured coalescent, we start in scenario 1 with $k = 15$ particles equally distributed on 3 islands. Those that are located on the same position may coalesce without a recombination event. In scenario 2 the particles are all placed on different islands and hence need to migrate before coalescence. Superimposed on this process we consider mutation events under an infinite sites model occuring with rate μ per generation. Hence, each simulation run leads to a $0-1$-SNP-matrix of dimension $k = 15 \times m$, where m indicates the total number of mutations. If gene copy i is affected by mutation j, the entry of the matrix is 1 (and 0 otherwise). We indicate the frequency spectrum by $(\xi)_{i=1,\ldots,14}$, where ξ_1 denotes the relative number of singletons,

[1] https://github.com/Moritz-Otto/motto-structured_coalescent

ξ_2 the doubletons and so on. An illustration of the setup is shown in Figure 5.4. For each scenario, we chose two mutation rates $2N\mu = \theta = 1$ and 5 and three recombination rates $r = 0.01/N, 1/N, 100/N$ and built the average frequency spectrum out of $10,000$ SNP-matrices for each parameter combination. Furthermore, we estimated the mutation rate based on the mean pairwise differences θ_π and Watterson's estimator θ_W. From this, one can calculate Tajima's D (Tajima, 1989) and the ratio $\theta\pi/\theta_W$ to detect deviations from neutrality.

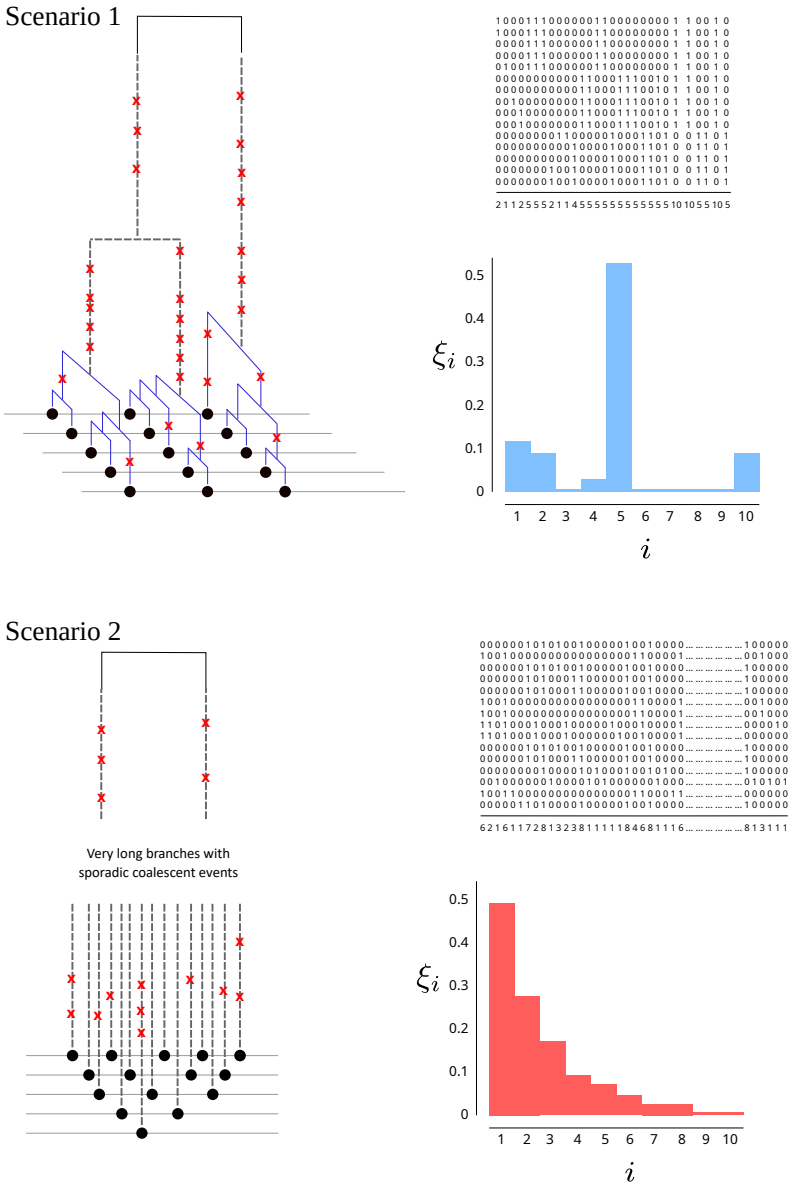


Figure 5.4: Illustration of scenario 1 (top) and 2 (bottom) of frequency spectrum simulations. In scenario 1 one considers $k = 15$ copies equally distributed on 5 chromosomes and no copy number variation. Hence, the copies located at the same position may coalesce quite fast (blue coalescence lines). Then, located at different positions, it may take same longer time to coalesce (indicated by dotted line). During the process, mutations may occur (red crosses), resulting in a 0-1-SNP-matrix (right, top). In this example, it is a 15×29 matrix, where its column sum counts the occurences of a SNP in the sample. The frequency of these is denoted by ξ_i, i.e. the number of singletons, doubletons etc. Due to the long coalescence time after the first coalescence events, we expect a high ξ_5. In contrast, in scenario 2 (bottom) we expect a frequency spectrum closer to the one expected under neutrality with a high number of singletons and a higher total number of mutations due to the elongated branches.

5.3 Results

Stationary distribution For given mean gene copy number L reduce the state space $\mathcal{S}$ to $n_0 \times n_0$ as in equation (5.6) and solve the linear equation system $\pi P^L = \pi$ with $\sum \pi = 1$. An example for $L = 20$ is shown in Figure 5.5**A**.

The stationary distribution fits well a bi-variate Gamma distribution. Its maximum is located at $(L/2, L/2)$, which is expected. Due to the recombination with gene arrays sampled from a population with $\mathcal{G}(2, L)$ distribution one expects the gene copy to be located in the center of a gene array of size L. We observe a decline from the diagonal of the distribution (i.e. the gene array size) and to the axes (the position within the array).

The stationary distribution is uniquely defined by L. Hence, we numerically calculated t_{mix} as in equation (5.7) for $L = 5, 10, 15, 20, 30$ (see Figure 5.5**B**). We find a strong correlation ($\rho^2 = 0.998$) and an almost perfect linear fit

$$t_{mix}(L) \approx 0.716 \cdot L + 3.54 \tag{5.12}$$

Therefore, when starting with a random distribution of gene copies that evolve according to P^L, the difference to the stationary distribution π after t generations can be bounded by

$$d(t) \leq \exp\left\{-\frac{\ln(2) \cdot t}{0.716 \cdot L + 3.54}\right\}.$$

Hitting probability Consider two particles moving independently on the $\mathbb{N} \times \mathbb{N}$-lattice with the dynamics defined in equation (5.4) with given L. We want to study their hitting time and therefore trace their relative distance to each other with D defined as in equation (5.8). The stationary distribution of this Markov process on $\mathbb{N}$ is given by the discrete Gamma distribution (5.1), see Figure 5.5**C**. We formulate as a conjecture

Conjecture 1. *Consider two particles moving independently on the $\mathbb{N} \times \mathbb{N}$-lattice with the dynamics defined in equation (5.4), given L fixed. Trace their relative distance D defined in equation (5.8). Over time, the distance is a Markov process on $\mathbb{N}$ with stationary distribution given by the discrete Gamma distribution equation (5.1), i.e.*

$$\begin{aligned} P[D_\infty = d] &= \sum_{\binom{x_1}{y_1},\binom{x_2}{y_2}} \pi(x_1, y_1)\pi(x_2, y_2) \cdot 1_{\left\{D\left(\binom{x_1}{y_1},\binom{x_2}{y_2}\right)=d\right\}} \\ &= p^{\mathcal{G}(2,L)}(d). \end{aligned} \tag{5.13}$$

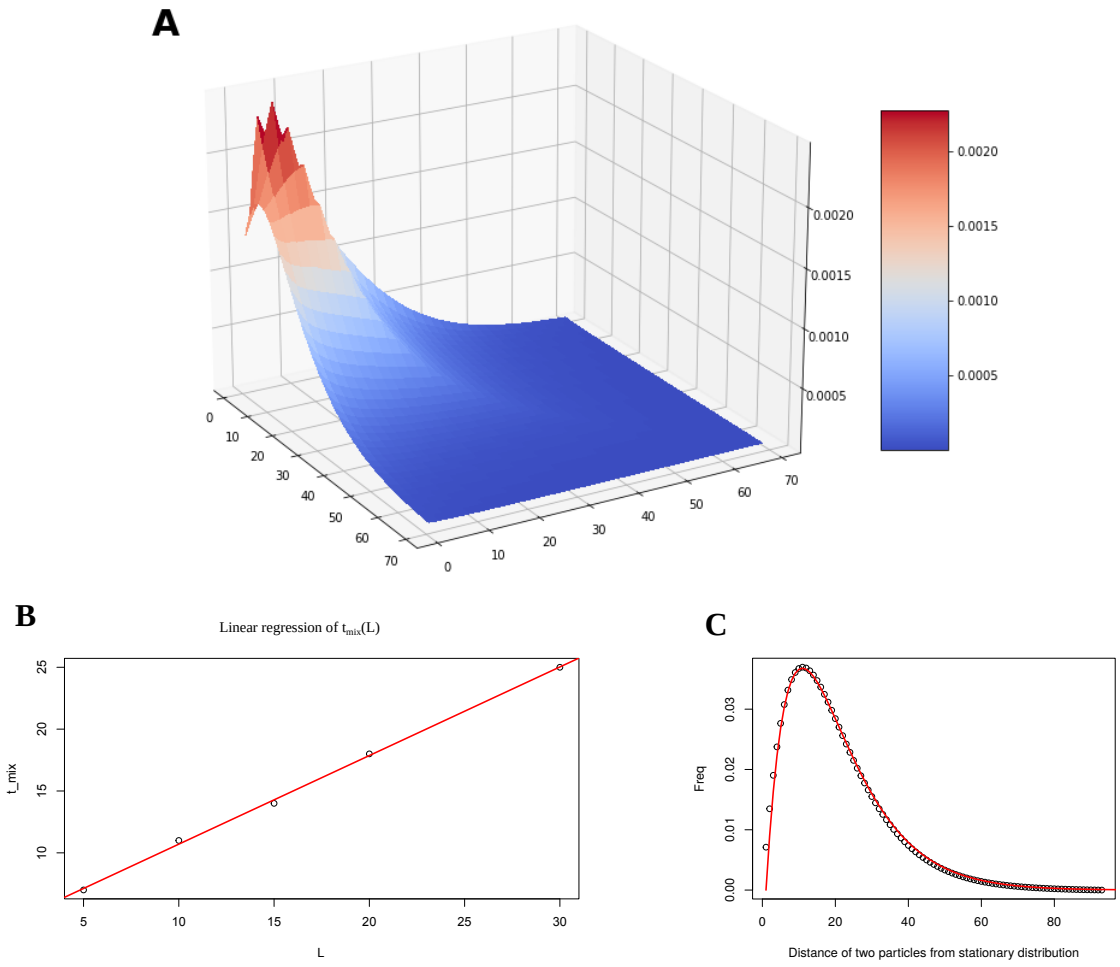


Figure 5.5: **A** Numerical solution of the stationary distribution π of the unequal recombination process, i.e. $\pi P^L = \pi$ with $L = 20$. **B** Linear regression of $t_{mix}(L)$. **C** Evidence for Conjecture 1, that the relative distance Markov process of two copies has stationary distribution $\mathcal{G}(2, L)$. Example shows $L = 20$. The small difference may occur by rounding errors and the truncation of the process to $n_0 \times n_0$.

Unequal Recombination We ran `msprime` on the complex model of gene copy trajectories with migration dynamics given by the unequal recombination process defined in equation (5.4). Figure 5.6 shows the mean coalescence time of 10,000 coalescence events of any pair from the 5×5 sample grid.

As in the simpler scenarios of symmetric or continental island models (see Appendix), the distribution has two modes, indicating whether gene copies are at the same or different positions at time $t = 0$. We find short coalescence times for those that may coalesce without a recombination event and long times for those at different positions.

Since the state space $\mathcal{S}$ is smaller for $L = 5$ than for $L = 10$, we observe shorter coalescence times for the recombination process with $L = 5$ than with $L = 10$. With a large number of subpopulations and small population size we reach a maximum mean coalescence rate for

$L = 10$, $r = 1\%$ and $N = 1,000$ at circa $42N$ (Figure 5.6**B**, first column in first row). Hence, we expect a high number of nucleotide differences, if we analyze two copies sampled from different gene array positions, if the population is small, has a large mean gene array length and a low recombination rate. Vice versa, we find for $L = 5$, $N = 10,000$ and $r = 5\%$ shorter coalescence times. Still, we expect an increased coalescence time of a factor of 5 for two copies at different positions. With smaller population structure, i.e. less states, larger population size N and higher migration resp. recombination rate r one expects the effect of spatial isolation to vanish and to approach the case of panmixia.

Coalescence times are substantially reduced when the recombination rate depends on the gene array size (Figure 5.6, third column). With higher recombination (resp. migration) rate the process gains mobility and with decreasing coalescence times the effects of subpopulation structure vanish.

Site frequency spectrum In the simple scenario of a haploid population of size N evolving under neutrality the expected site frequency spectrum is $\xi_i = \theta\frac{1}{i}$, where $\theta = 2N\mu$ is the population scaled mutation rate. In Figure 5.7 we show the scaled mean frequency spectrum of a sample of $k = 15$ gene copies under different recombination and mutation rates. As expected (see Figure 5.4) we find an increased value of ξ_5 and ξ_{10} in scenario 1, if recombination rate is low. Indeed, it corresponds to the expected frequency spectrum for $k = 3$. More precisely, since we observe an almost instantanuous coalescence for copies located at the same position, the number of distinct lineages reduces to 3. Hence, mutations that occur on those branches affect either 5 copies (equivalent to singletons in $k = 3$) or 10 (doubletons). The effect vanishes with higher recombination and the spectrum resembles the expected neutral frequency spectrum. In scenario 2 we do not observe this effect, since the copies are already located at different positions. Hence, the shape of the frequency spectrum does not change. As seen in Figure 5.6, the coalescence times increase with the effect of isolation. Therefore, the height of the coalescence tree increases and one expects a higher number of mutations. Also increasing the mutation rate increases the expected number of mutations. In Figure 5.7 we see that increasing θ by a factor of 5 also increases the number of mutations by the same factor, as expected. With increased recombination rate, the coalescence times reduce and therefore also the number of mutations decrease. When choosing $L = 3$, $N = 80,000$ and $r = 0.00125$ we can not distinguish scenario 1 and 2 by their frequency spectrum. We also calculated the θ_π/θ_W ratio. As expected, for all parameter settings in scenario 2 and high recombination in scenario 1 we observe a ratio close to 1. If recombination is low in scenario 1 we observe a higher θ_π, since the frequency spectrum shows a lack of rare alleles, i.e. almost no singletons. For $rN = 0.01$ the ratio is $\theta_\pi/\theta_W \approx 1.5$ and for $rN = 1$ it is ≈ 1.3.

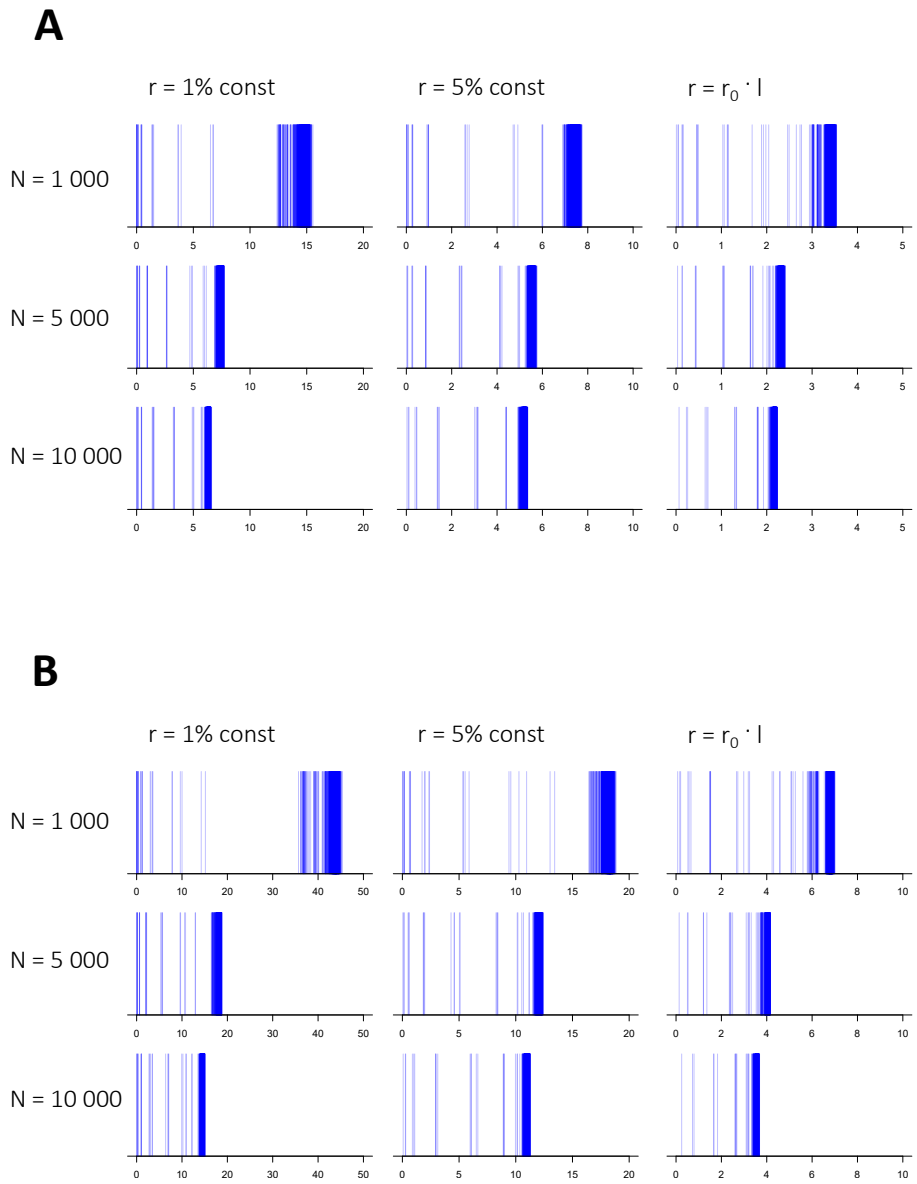
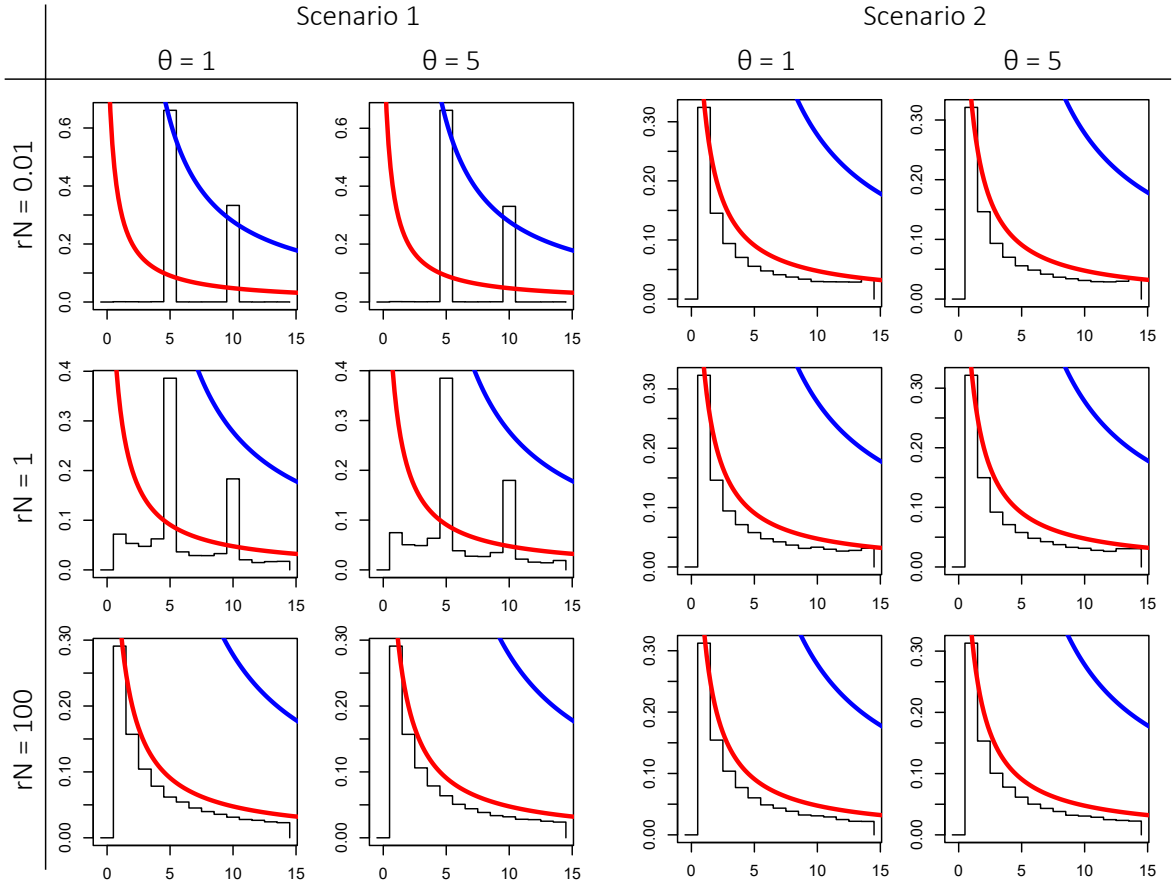


Figure 5.6: Mean coalescence times of the unequal recombination structured coalescent for mean gene array size (**A**) $L = 5$ and (**B**) $L = 10$. Every blue line marks one mean coalescence time of the 625 pairs in the spanning 5×5 grid. The mean was calculated from 10,000 coalescence events, simulated with `msprime`.



	Scenario 1				Scenario 2			
	$\theta = 1$		$\theta = 5$		$\theta = 1$		$\theta = 5$	
	m	θ_π/θ_W	m	θ_π/θ_W	m	θ_π/θ_W	m	θ_π/θ_W
$rN = 0.01$	4 216	1.54	21 112	1.54	9 198	0.95	46 192	0.95
$rN = 1$	55	1.38	283	1.38	105	0.97	536	0.97
$rN = 100$	12	0.99	59	0.99	12	0.99	60	0.99

Figure 5.7: Frequency spectrum $(\xi_i)_{i=1,\dots,15}$ of the structured coalescent process with $k = 15$ gene copies. Red line marks neutral frequency spectrum $1/i$, blue line marks scaled neutral spectrum $5/i$. Table below shows the (rounded) mean number of mutations m as well as ratio of mutation rate estimates θ_π and θ_W.

5.4 Discussion

In this study we have introduced a novel interpretation of population substructure models, in which we do not trace the migration routes of individuals but of gene copies in gene arrays. In its principle, it can be applied to any genomic components that change their position along the DNA. The challenge occurs in modelling the state space and transition rates. Here, we used the unequal recombination model of (Otto et al., 2022) and calculated explicitly the transition probabilities with which a particular gene copy changes its position.

To do so, we assumed a sufficiently large population at migration equilibrium (or, respectively gene copies in recombination equilibirum) with the structured coalescent as underlying genealogical model. We described single gene copies as particles in a two-dimensional state space with independent trajectories. However, when tracing multiple copies located on the same haplotype, one has to keep in mind that one recombination event can induce a position change of several gene copies located on this haplotype. A coalescent event can happen only if copies on different haplotypes share their position.

This Markov process is interesting in itself: it has a two-dimensional state space and we conjecture that its stationary distribution has one-dimensional marginals which are Gamma distributed. We numerically calculated the 2-D stationary distribution, the mixing time and the distance distribution of two copies.

We showed that the unequal recombination process, considered here, and the structured coalescent agree with respect to the analytic results derived by Wilkinson-Herbots (1998). However, in her analysis it became already clear that even with a simple island structure the solution of equation (5.10) can be quite challenging. Even numerical results are difficult to compute if the migration graph contains multiple connected islands. In our model this implies that either the entire array size must be limited to a small value or that the width of a positional shift, in other words the number of accessible islands, must be capped at a moderate value. This means that larger changes in copy number could only be accomplished by multiple unequal recombination events. However, to not limit the applicability of our model to only very small gene families, we favour the second solution. The incurred error should be negligible, since the probability of a large shift is anyway very small (see Figure 5.8B).

Coalescence time of two genes can be substantially longer than $2N$ in our model. Therefore, pairwise genetic diversity may also be much higher than $2N\mu$. But what about the coalescent tree of larger samples, which may contain a mixture of orthologous and paralogous gene copies collected from several individuals? This question is of practical interest, because orthology and paralogy of gene copies is often not easily distinguishable in experimental data, especially when gene families are large (de Weyer et al., 2019). Nevertheless, researchers sometimes apply to such data the usual population genetic statistics and tests of neutrality, Tajima's D for instance, without critically examining the validity of such an approach.

We investigated this problem for a moderate sample size of $k = 15$, taken from $n = 5$ chromosomes, and considered the two extreme cases: without gCNV, i.e. all chromosomes carry the same number of gene copies (here: 3) and with high gCNV: the five chromosomes

carry 1, 2, 3, 4 and 5 copies. In both cases mean copy number is 3 (see Figure 5.7). The two cases exemplify the effect on the frequency spectrum: without gCNV, as expected, there is clear clustering of all orthologous copies. All coalescent events have occurred long before only two unequal recombination events unite the remaining three paralogous copies. This leads to a spectrum with several clearly distinguishable modes. In contrast, high gCNV requires many more unequal recombination events. Depending on the rate, the coalescent and recombination events may be interspersed on an more or less elongated tree. Regarding its shape, the spectrum tends to resemble the $1/x$-spectrum of a standard Kingman-coalescent. However, due to the elongated tree size, the number of segregating sites is expected to be much higher than for a sample of single copy genes.

Under low recombination rates and high gCNV (see Figure 5.4C), the frequency spectrum can severely differ from the one expected under neutrality and increase the time to the most recent common ancestor up to a factor of ≈ 42. Hence, when analyzing data from large gene families by pooling all sequences, population statistics based on the standard frequency spectrum should be used and interpreted with caution. Inferences based on summary statistics derived from the frequency spectrum may be heavily biased. As shown, the θ_π/θ_W ratio can reach values of 1.5 if recombination is low. This leads to a positive Tajima's D, which might be misinterpreted as balancing selection or a sudden population contraction.

There are several routes of investigation to be pursued further. Here, we assumed a time independent distribution of gene families under neutrality. Since multicopy gene families are often involved in adaptive processes, one may enlarge the model with positive selection for some mean copy number, as in Otto et al. (2022). Additionally to the size of gene arrays, sequence similarity may also affect the rate and break point choice of recombination. This, together with the question how unequal recombination affects coalescent tree topology and its statistics is subject of current further investigations. Another complication in the analysis and interpretation of experimental data is the possibility of ectopic gene conversion between copies to obscure genealogical signals. Finally, inter-chromosomal recombination may be another source of generating gCNV. Large gene families, such as the NLR receptors, may indeed be subject to a combination of all these mechanisms.

In any case, one should know which patterns of genetic diversity to expect, before drawing conclusions about the adaptive role of a particular copy or a gene cluster.

Appendix

Validation of algorithm The first example was a symmetric island model with $n = 5$ islands with equal subpopulation sizes, i.e. $N_i = N/5$ for $i = 1, ..., 5$ and where all islands are connected with each other (see Figure 5.8). With `msprime` we simulated $10,000$ coalescent events for each of all 25 pairs of starting points, with population sizes $N = 1000$, 5000, 10000 and migration rates $m = 0.01$, 0.001. To compare the simulation results with the result given by Wilkinson-Herbots (1998), we calculated the Laplace transformation of the empiricial distribution, i.e.

$$\phi_{ij}(s) = \frac{1}{10000} \sum_{k=1}^{10000} e^{-s \cdot T_k},$$

for 30 discrete points $s \in (1e-8, 1e-2)$, where 10 points span the interval of $(1e-8, 1e-7)$ and 20 the interval of $(1e-7, 1e-2)$. The fine grid near 0 was used to derive the mean and variance of the distribution, since the n-th moment of a random variable X is given by the n-th derivative of the Laplace transformation evaluated at 0, i.e.

$$\mathbb{E}\left[X^n\right] = (-1)^n \frac{d^n}{ds^n} \mathbb{E}\left[e^{-sX}\right](0).$$

We numerically solved the linear equation system equation (5.10) with the same parameters, also using an in-house developed `python`-script.

In the second example we defined a continental island model, that mimics the complex unequal recombination process defined in equation (5.4). Consider the 5×5 lattice $\{-2, -1, ..., 2\}^2$. The population size increases towards the center, the *'continent'*, such that in each discrete step the population size doubles, i.e.

$$(N_{ij}) = \frac{N}{100} \begin{bmatrix} 1 & 2 & 4 & 2 & 1 \\ 2 & 4 & 8 & 4 & 2 \\ 4 & 8 & 16 & 8 & 4 \\ 2 & 4 & 8 & 4 & 2 \\ 1 & 2 & 4 & 2 & 1 \end{bmatrix}.$$

The migration dynamics is similiar to equation (5.4) such that an individual can jump from one to another island, but only either horizontally or vertically. Furthermore, jumping probabilities are biased towards the center (see Figure 5.8), with

$$\begin{aligned} &\text{Prob}[(x, y) \to (x, y)] = 1 - m \\ &\text{Prob}[(x, y) \to (x, k)] = 0.5m \cdot [0.01, 0.2, 0.5, 0.2, 0.05]_k = \text{Prob}[(x, y) \to (y, k)]. \end{aligned}$$

We used the same parameter settings as before ($N = 1000$, 5000, 10000 and $m = 0.01$, 0.001) and calculated the numerical solution for the Laplace transformation as in equation (5.10).

We ran our `msprime`-script on the solvable symmetric and continental island models. We

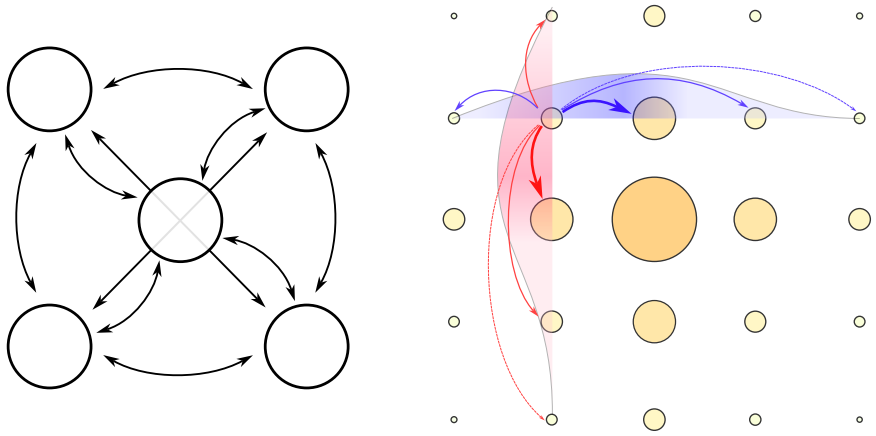


Figure 5.8: Left: Symmetric island model with $n = 5$ islands. Migration rate m is constant and all islands are connected. Right: Continental island model. Central continent marks the largest population, whereas the ones at the boundary have small population sizes. Arrows mark the transition probablitites from the island at position $(-1, 1)$. Blue arrows mark the horizontal, red arrows the vertical transition probabilities. The thickness of the arrows and the bellshape distribution indicate the transition probability.

find that $10,000$ coalescence events describe the true distribution of T_{ij} with high precision. Results are shown in Figure 5.9. Due to symmetry and connectivity of the island structure the distribution of the coalescence time reduces to two cases: individuals located on the same island at time $t = 0$ or not. With high migration and large population size these differences dissolve and the expected coalescence time gets closer to the panmictic case of $E[T_2] = N$. In contrast, in a small and strongly isolated population we find coalescence times of up to $2N$ in the symmetric island model and up to $23N$ in the continental island model. These results are in agreement with the analytic results of Wilkinson-Herbots (1998), where she concluded, that in a symmetric island model with $c_i = 1$ one has

$$\begin{aligned} \varphi_{ii}(s) &= \frac{M + (n-1)s}{M + (nM + n - 1)s + (n-1)s^2}, \quad \mathbb{E}[T_{ii}] = n, \\ &\qquad\qquad Var(T_{ii}) = n^2 + 2\frac{(n-1)^2}{M} \\ \varphi_{ij}(s) &= \frac{M}{M + (nM + n - 1)s + (n-1)s^2}, \quad \mathbb{E}[T_{ij}] = n + \frac{n-1}{M} \\ &\qquad\qquad Var(T_{ij}) = n^2 + 2\frac{(n-1)^2}{M} + \frac{(n-1)^2}{M^2}. \end{aligned} \tag{5.14}$$

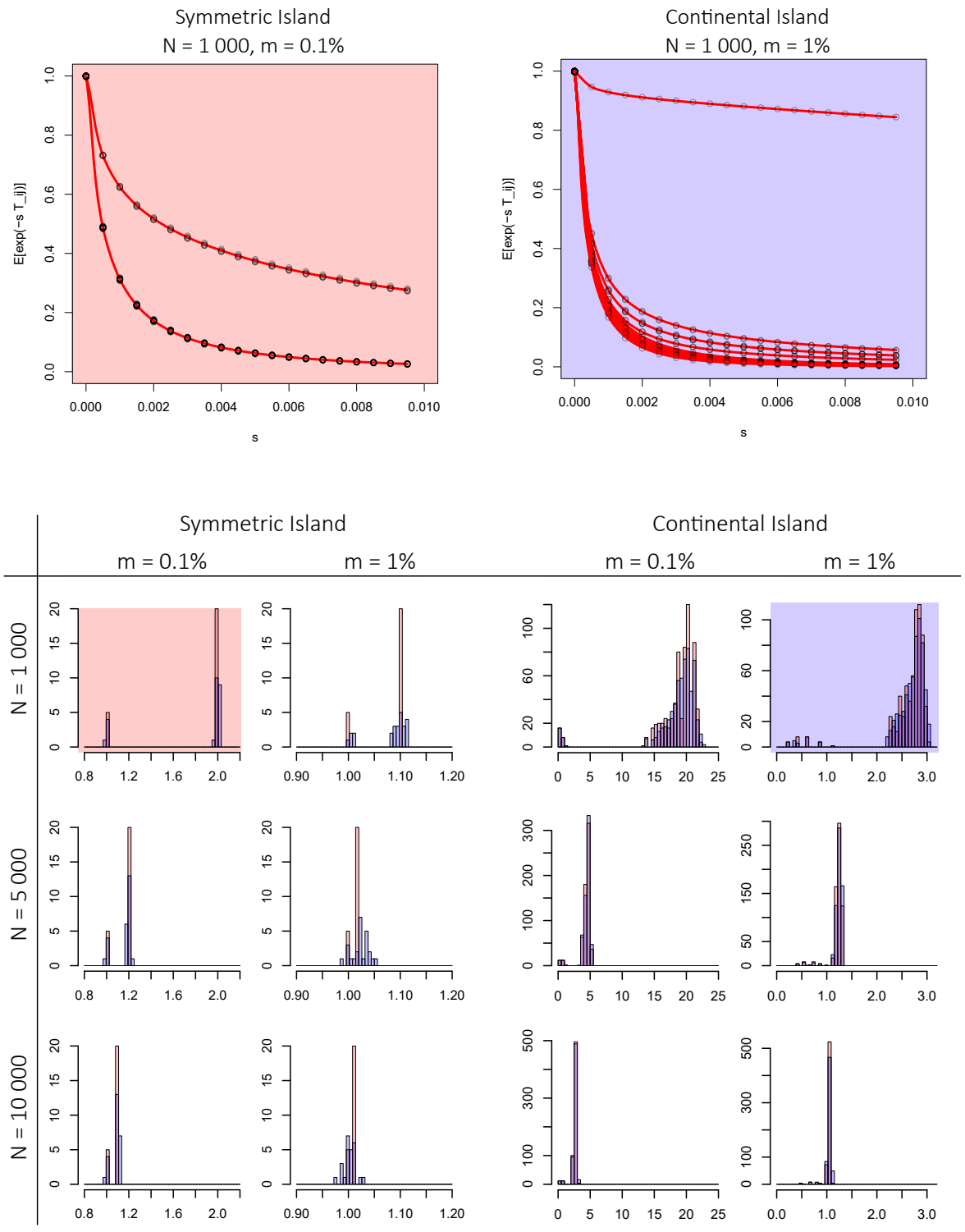


Figure 5.9: Distribution of coalescence times. Top figure: Laplace transformation of the empiricial distribution $\phi_(s) = \frac{1}{10000} \sum e^{-sT_k}$ for `msprime`-simulated data. Red line marks the numerical solution $\varphi_{ij}(s)$ of equation (5.10) for s. Bottom figure: mean values of coalescence times in symmetric island (left) and continental island (right) models. Blue bars mark the mean values calculated from `msprime`-simulated data, red bars the solution of (5.10).

6 Conclusions and outlook

6.1 Summary

In most eukaryotic genomes a large portion of genes are considered to be duplicates. While some gene families consist of an immense number, as the ≈ 400 copies of NLR immune genes in the zebrafish *Danio rerio* (Howe et al., 2016), we focused here on human genes with intermediate copy numbers ranging from 5 to 60. Together with their high mutation rate, gene copy number variation (CNV) plays a significant role in genomic variability that allows for adaptive responses to biotic and abiotic stresses (Sudmant et al., 2015a; Kondrashov, 2012). One of the molecular processes that generates CNV is that of unequal recombination.

In chapter 3 we introduced a model on the evolution of multi-copy gene families, in which the interplay of unequal recombination and selection towards an optimal copy number with high allelic diversity generates copy number variation. The selection ratio s_x/s_y, where s_x describes the benefit of allelic diversity and s_y the cost of copy number accumulation, determines the optimal copy number y_{opt} and the recombination ratio r/s_x determines the spread around y_{opt}, see Figure 3.2 and equation (3.8). We derived that under neutrality the gene copy number is Gamma-distributed and showed that even with selection, it is still well approximated by it. We analyzed data from the 1,000 Genomes Project to estimate selection and recombination parameters for three selected candidate genes with copy number distribution shown in Figure 3.3. We observed a high copy number variation and almost neutral evolution in `PRR20A`, which is likely to be a pseudogene. In contrast, `PSG3`, which is involved in pregnancy maintenance, follows a distribution close to the optimal value. We ran migration and bottleneck simulations to see the effect on the population fitness. Finally, with the implementation of a recombination rate modifier, we observe a decreasing recombination rate. On first thought, it seems counterintuitive that the variation-driving force (the recombination rate) is decreasing in a setting of diversifying selection. On second thought, recombination also generates copy numbers with low fitness and breaks those of high fitness.

In chapter 4 we equipped the model with migration and population size changes according to human demography. Starting with a data set from the 1,000 Genomes Project with 180 gene families in 165 individuals of three populations (60 African Yoruba, 60 Central Europe and 45 East Asia), we filtered those of intermediate copy number that show significant differences in either mean or variance of copy number distribution between populations, which resulted in 42 gene families. For these candidates, we estimated recombination rate and selection strength in all three populations and used bottleneck simulations with parameters of the ancestral YRI population, to test whether the differences can be explained by demography

alone under constant selective pressure. In several scenarios, we find significant differences of simulated and empirical data (see Table 4.4), leading to the rejection of this hypothesis. Considering a change of selection parameters towards the estimates of the derived Asian and European values, the simulations are often in agreement with the empirical data, see Figure 4.5. One of the chosen candidate genes that are likely to be explained with a change of selection parameters is `AMY1A`. Several studies indicate that individuals from populations with high-starch diets have, on average, more gene copies than those with traditionally low-starch diets (Perry et al., 2007; Pajic et al., 2019; Atkinson et al., 2018)

In chapter 5 we still consider gene copy number evolution under the introduced model of unequal recombination, in which we did not focus on the distribution within a population but on individiual genes. With the process of unequal recombination a gene may change its position within the gene array. Using the unequal recombination process, we derived the transition probabilities of a single gene within the array. The trajectory of a copy can be interpreted as a particle moving in a 2-dimensional lattice, where the coordinates describe the position in the gene array. Therefore, comparing the sequences of genes at different positions one expects a higher genetic diversity. This idea led to a new interpretation of the structured coalescent, which was initially introduced to describe the coalescent process in spatially subdivided populations with migration (Takahata, 1988; Nordborg, 1997; Wilkinson-Herbots, 1998). As a backward-in-time process, two particles can fuse with a defined probability, if located at the same position. The process stops, if only one particle remains, which is (in coalescence theory terms) the most recent common ancestor. As a theoretical component, we explored the stationary distribution of the Markovian jump process, its mixing time and the relative distance of two particles (see Figure 5.5). Applied to the context of gene copies we analyzed the time to the most recent common ancestor (i.e. the expected genetic diversity of two copies, see Figure 5.6) and the site frequency spectrum of gene families with and without copy number variation (see Figure 5.7). In a dynamic system with high recombination rate, the signal is close to that of a panmictic population. But if the effect of substructure is strong, we find clear deviations. This is especially important to keep in mind when analyzing sequenced data from gene families without knowing their arrangement. The position of copies in the genome can provide important extra information that affect the analysis of sequenced data and may lead to different interpretations.

6.2 Further research

The initial model presented in chapter 3 assumes a fitness function that favours allelic diversity. In theory, it is straightforward to model an infinite alleles model such that any mutation generates a new allele. However, in real data it is a challenging task to decide, how much genetic variation is needed to call a variant a new allele. Some studies consider a single nucleotide polymorphism as sufficient, whereas others require a new function of the gene. In the example of immune genes, it is reasonable to argue that a new version is considered, if it detects a different variant of a pathogen. But then again, the nomenclature is inconsistent whether to call this a new allele or a completely new gene.

As shown in the application of the structured coalescent, the position of a gene copy contains information that can be used to provide better estimates on genetic variation. However, current sequencing and assembly tools often can not offer this information. Especially when using short read sequencing techniques, it remains a challenging problem to uniquely identify the members of a large gene family and to correctly map them to a reference genome. Additionally, in diploid organisms, assigning copies to each haplotype is also a difficult task. With long read sequencing and a high quality reference pangenome that includes structural variations it may become feasible to get precise copy number counts and their sequences in the future.

The developed model considers unequal recombination as the main driver of copy number evolution. In future studies it might be reasonable to consider additional molecular processes as gene duplication or gene conversion. Currently, we investigate models on these mechanisms in different projects. A combination of the findings might result in a more realistic model of the evolution of gene copies.

From a theoretical point of view, the novel interpretation of the structured coalescent might lead to new inspirations and applications in future modelling. Now that we have built the bridge from population migration routes to molecular position changes, it is exciting to think of further applications and re-interpretations of *structure*. One example was recently developed in the context of the *seed-bank coalescent* (Blath et al., 2015, 2016). Here, the individuals may change their state from active to dormant. This shows, that the *state space* of the Markov process can appear in many different forms, not only describing the position of an individual or, as in our model, the position of a gene.

Bibliography

1000 Genomes Project Consortium et al. (2012). An integrated map of genetic variation from 1,092 human genomes. *Nature*, 491(7422):56.

1000 Genomes Project Consortium et al. (2015). A global reference for human genetic variation. *Nature*, 526(7571):68–74.

Ai, Z., Li, M., Liu, W., Foo, J.-N., Mansouri, O., Yin, P., Zhou, Q., Tang, X., Dong, X., Feng, S., Xu, R., Zhong, Z., Chen, J., Wan, J., Lou, T., Yu, J., Zhou, Q., Fan, J., Mao, H., Gale, D., Barratt, J., Armour, J. A. L., Liu, J., and Yu, X. (2016). Low α-defensin gene copy number increases the risk for IgA nephropathy and renal dysfunction. *Science Translational Medicine*, 8(345).

Aldred, P. M., Hollox, E. J., and Armour, J. A. (2005). Copy number polymorphism and expression level variation of the human α-defensin genes DEFA1 and DEFA3. *Human Molecular Genetics*, 14(14):2045–2052.

Allison, A. C. (1956). The sickle-cell and haemoglobin c genes in some african populations. *Annals of Human Genetics*, 21(1):67–89.

Atkinson, F. S., Hancock, D., Petocz, P., and Brand-Miller, J. C. (2018). The physiologic and phenotypic significance of variation in human amylase gene copy number. *The American Journal of Clinical Nutrition*, 108(4):737–748.

Austerlitz, F., Jung-Muller, B., Godelle, B., and Gouyon, P.-H. (1997). Evolution of coalescence times, genetic diversity and structure during colonization. *Theoretical Population Biology*, 51(2):148–164.

Ayabe, T., Satchell, D. P., Wilson, C. L., Parks, W. C., Selsted, M. E., and Ouellette, A. J. (2000). Secretion of microbicidal α-defensins by intestinal paneth cells in response to bacteria. *Nature Immunology*, 1(2):113–118.

Bahr, A. and Wilson, A. B. (2012). The evolution of MHC diversity: Evidence of intralocus gene conversion and recombination in a single-locus system. *Gene*, 497(1):52–57.

Bairam, A. F., Rasool, M. I., Alherz, F. A., Abunnaja, M. S., Daibani, A. A. E., Gohal, S. A., Alatwi, E. S., Kurogi, K., and Liu, M.-C. (2019). Impact of SULT1a3/SULT1a4 genetic polymorphisms on the sulfation of phenylephrine and salbutamol by human SULT1a3 allozymes. *Pharmacogenetics and Genomics*, 29(5):99–105.

Baumdicker, F., Bisschop, G., Goldstein, D., Gower, G., Ragsdale, A. P., Tsambos, G., Zhu, S., Eldon, B., Ellerman, E. C., Galloway, J. G., Gladstein, A. L., Gorjanc, G., Guo, B., Jeffery, B., Kretzschumar, W. W., Lohse, K., Matschiner, M., Nelson, D., Pope, N. S., Quinto-Cortés, C. D., Rodrigues, M. F., Saunack, K., Sellinger, T., Thornton, K., van Kemenade, H., Wohns, A. W., Wong, Y., Gravel, S., Kern, A. D., Koskela, J., Ralph, P. L., and Kelleher, J. (2022). Efficient ancestry and mutation simulation with msprime 1.0. *Genetics*, 220(3):iyab229.

Beeson, S. K., Mickelson, J. R., and McCue, M. E. (2019). Exploration of fine-scale recombination rate variation in the domestic horse. *Genome Research*, 29(10):1744–1752.

Blath, J., Casanova, A. G., Eldon, B., Kurt, N., and Wilke-Berenguer, M. (2015). Genetic variability under the seedbank coalescent. *Genetics*, 200(3):921–934.

Blath, J., Casanova, A. G., Kurt, N., and Wilke-Berenguer, M. (2016). A new coalescent for seed-bank models. *The Annals of Applied Probability*, 26(2).

Box, G. E. P. (1980). Sampling and bayes' inference in scientific modelling and robustness. *Journal of the Royal Statistical Society. Series A (General)*, 143(4):383.

Brahmachary, M., Guilmatre, A., Quilez, J., Hasson, D., Borel, C., Warburton, P., and Sharp, A. J. (2014). Digital genotyping of macrosatellites and multicopy genes reveals novel biological functions associated with copy number variation of large tandem repeats. *PLoS Genetics*, 10(6):e1004418.

Bridges, C. B. (1936). The bar "gene" a duplication. *Science*, 83(2148):210–211.

Briskine, R. V., Paape, T., Shimizu-Inatsugi, R., Nishiyama, T., Akama, S., Sese, J., and Shimizu, K. K. (2016). Genome assembly and annotation of Arabidopsis halleri, a model for heavy metal hyperaccumulation and evolutionary ecology. *Molecular Ecology Resources*, 17(5):1025–1036.

Brix, L. A., Barnett, A. C., Duggleby, R. G., Leggett, B., and McManus, M. E. (1999). Analysis of the substrate specificity of human sulfotransferases SULT1a1 and SULT1a3: site-directed mutagenesis and kinetic studies. *Biochemistry*, 38(32):10474–10479.

Butcher, N. J., , Horne, M. K., Mellick, G. D., Fowler, C. J., Masters, C. L., and Minchin, R. F. (2017). Sulfotransferase 1a3/4 copy number variation is associated with neurodegenerative disease. *The Pharmacogenomics Journal*, 18(2):209–214.

Carpenter, D., Dhar, S., Mitchell, L. M., Fu, B., Tyson, J., Shwan, N. A. A., Yang, F., Thomas, M. G., and Armour, J. A. L. (2015). Obesity, starch digestion and amylase: association between copy number variants at human salivary (AMY1) and pancreatic (AMY2) amylase genes. *Human Molecular Genetics*, 24(12):3472–3480.

Carvalho, C. M. B. and Lupski, J. R. (2016). Mechanisms underlying structural variant formation in genomic disorders. *Nature Reviews Genetics*, 17(4):224–238.

Castro, L. F. C., Goncalves, O., Mazan, S., Tay, B.-H., Venkatesh, B., and Wilson, J. M. (2014). Recurrent gene loss correlates with the evolution of stomach phenotypes in gnathostome history. *Proceedings of the Royal Society B: Biological Sciences*, 281(1775):20132669.

Chao, L. (1988). Evolution of sex in RNA viruses. *Journal of Theoretical Biology*, 133(1):99–112.

Charlesworth, B. and Charlesworth, D. (2016). Population genetics from 1966 to 2016. *Heredity*, 118(1):2–9.

Chen, Q., Hakimi, M., Wu, S., Jin, Y., Cheng, B., Wang, H., Xie, G., Ganz, T., Linzmeier, R. M., and Fang, X. (2010). Increased genomic copy number of DEFA1/DEFA3 is associated with susceptibility to severe sepsis in chinese han population. *Anesthesiology*, 112(6):1428–1434.

Chen, Q., Yang, Y., Hou, J., Shu, Q., Yin, Y., Fu, W., Han, F., Hou, T., Zeng, C., Nemeth, E., Linzmeier, R., Ganz, T., and Fang, X. (2019). Increased gene copy number of DEFA1/DEFA3 worsens sepsis by inducing endothelial pyroptosis. *Proceedings of the National Academy of Sciences*, 116(8):3161–3170.

Dajani, R., Sharp, S., Graham, S., Bethell, S. S., Cooke, R. M., Jamieson, D. J., and Coughtrie, M. W. (1999). Kinetic properties of human dopamine sulfotransferase (SULT1a3) expressed in prokaryotic and eukaryotic systems: Comparison with the recombinant enzyme purified fromEscherichia coli. *Protein Expression and Purification*, 16(1):11–18.

Darwin, C. (1845). *Journal of researches into the natural history and geology of the countries visited during the voyage of H.M.S. Beagle round the world, under the Command of Capt. Fitz Roy, R.N. 2d edition.* J. Murray.

Darwin, C. (1859). *On the Origin of Species by Means of Natural Selection.* Murray, London. or the Preservation of Favored Races in the Struggle for Life.

de Bakker, P. I. W., McVean, G., Sabeti, P. C., Miretti, M. M., Green, T., Marchini, J., Ke, X., Monsuur, A. J., Whittaker, P., Delgado, M., Morrison, J., Richardson, A., Walsh, E. C., Gao, X., Galver, L., Hart, J., Hafler, D. A., Pericak-Vance, M., Todd, J. A., Daly, M. J., Trowsdale, J., Wijmenga, C., Vyse, T. J., Beck, S., Murray, S. S., Carrington, M., Gregory, S., Deloukas, P., and Rioux, J. D. (2006). A high-resolution HLA and SNP haplotype map for disease association studies in the extended human MHC. *Nature Genetics*, 38(10):1166–1172.

de Weyer, A.-L. V., Monteiro, F., Furzer, O. J., Nishimura, M. T., Cevik, V., Witek, K., Jones, J. D., Dangl, J. L., Weigel, D., and Bemm, F. (2019). A species-wide inventory of NLR genes and alleles in arabidopsis thaliana. *Cell*, 178(5):1260–1272.e14.

Demuth, J. P. and Hahn, M. W. (2009). The life and death of gene families. *Bioessays*, 31(1):29–39.

Donsker, M. (1951). *An Invariance Principle for Certain Probability Limit Theorems.* American Mathematical Society. Memoirs.

Dudek, K., Gaczorek, T. S., Zieliński, P., and Babik, W. (2019). Massive introgression of major histocompatibility complex (MHC) genes in newt hybrid zones. *Molecular Ecology*, 28(21):4798–4810.

Durrett, R. (2008). *Probability Models for DNA Sequence Evolution.* Springer New York.

Eichler, E. E. (2008). Copy number variation and human disease. *Nature Education*, 1(3):1.

Ejsmond, M. J. and Radwan, J. (2009). MHC diversity in bottlenecked populations: a simulation model. *Conservation Genetics*, 12(1):129–137.

Ekblom, R., Saether, S. A., Jacobsson, P., Fiske, P., Sahlman, T., Grahn, M., Kålås, J. A., and Höglund, J. (2007). Spatial pattern of MHC class II variation in the great snipe (gallinago media). *Molecular Ecology*, 16(7):1439–1451.

Ewens, W. (1972). The sampling theory of selectively neutral alleles. *Theoretical Population Biology*, 3(1):87–112.

Falchi, M., Moustafa, J. S. E.-S., Takousis, P., Pesce, F., Bonnefond, A., Andersson-Assarsson, J. C., Sudmant, P. H., Dorajoo, R., Al-Shafai, M. N., Bottolo, L., Ozdemir, E., So, H.-C., Davies, R. W., Patrice, A., Dent, R., Mangino, M., Hysi, P. G., Dechaume, A., Huyvaert, M., Skinner, J., Pigeyre, M., Caiazzo, R., Raverdy, V., Vaillant, E., Field, S., Balkau, B., Marre, M., Visvikis-Siest, S., Weill, J., Poulain-Godefroy, O., Jacobson, P., Sjostrom, L., Hammond, C. J., Deloukas, P., Sham, P. C., McPherson, R., Lee, J., Tai, E. S., Sladek, R., Carlsson, L. M. S., Walley, A., Eichler, E. E., Pattou, F., Spector, T. D., and Froguel, P. (2014). Low copy number of the salivary amylase gene predisposes to obesity. *Nature Genetics*, 46(5):492–497.

Fay, J. C. and Wu, C.-I. (2000). Hitchhiking under positive darwinian selection. *Genetics*, 155(3):1405–1413.

Fink, G. A. (2008). Hidden markov models. In *Markov Models for Pattern Recognition*, pages 61–93. Springer Berlin Heidelberg.

Fisher, R. A. (1930). *The genetical theory of natural selection*. Clarendon Press.

Frankham, R., Bradshaw, C. J., and Brook, B. W. (2014). Genetics in conservation management: Revised recommendations for the 50/500 rules, red list criteria and population viability analyses. *Biological Conservation*, 170:56–63.

Fu, Y. (1995). Statistical properties of segregating sites. *Theoretical Population Biology*, 48(2):172–197.

Fu, Y.-X. (1997). Statistical tests of neutrality of mutations against population growth, hitchhiking and background selection. *Genetics*, 147(2):915–925.

Fulton, J. E., McCarron, A. M., Lund, A. R., Pinegar, K. N., Wolc, A., Chazara, O., Bed'Hom, B., Berres, M., and Miller, M. M. (2016). A high-density SNP panel reveals extensive diversity, frequent recombination and multiple recombination hotspots within the chicken major histocompatibility complex b region between BG2 and CD1a1. *Genetics Selection Evolution*, 48(1).

Ganz, T., Selsted, M. E., Szklarek, D., Harwig, S. S., Daher, K., Bainton, D. F., and Lehrer, R. I. (1985). Defensins. natural peptide antibiotics of human neutrophils. *Journal of Clinical Investigation*, 76(4):1427–1435.

Golicz, A. A., Bayer, P. E., Bhalla, P. L., Batley, J., and Edwards, D. (2020). Pangenomics comes of age: from bacteria to plant and animal applications. *Trends in Genetics*, 36(2):132–145.

Grimwood, J., Gordon, L. A., Olsen, A., Terry, A., Schmutz, J., Lamerdin, J., Hellsten, U., Goodstein, D., Couronne, O., Tran-Gyamfi, M., et al. (2004). The DNA sequence and biology of human chromosome 19. *Nature*, 428(6982):529–535.

Haigh, J. (1978). The accumulation of deleterious genes in a population muller's ratchet. *Theoretical Population Biology*, 14(2):251–267.

Haldane, J. B. S. (1927). A mathematical theory of natural and artificial selection, part v: Selection and mutation. *Mathematical Proceedings of the Cambridge Philosophical Society*, 23(7):838–844.

Haldane, J. B. S. (1937). The effect of variation of fitness. *The American Naturalist*, 71(735):337–349.

Hanikenne, M., Talke, I. N., Haydon, M. J., Lanz, C., Nolte, A., Motte, P., Kroymann, J., Weigel, D., and Krämer, U. (2008). Evolution of metal hyperaccumulation required cis-regulatory changes and triplication of HMA4. *Nature*, 453(7193):391–395.

Hardy, G. H. (1908). Mendelian proportions in a mixed population. *Science*, 28(706):49–50.

Hartl, D. and Clark, A. (2007). *Principles of Population Genetics*. Sinauer.

Herdegen, M., Babik, W., and Radwan, J. (2014). Selective pressures on MHC class II genes in the guppy (poecilia reticulata) as inferred by hierarchical analysis of population structure. *Journal of Evolutionary Biology*, 27(11):2347–2359.

Hess, C. M. (2000). MHC class II pseudogene and genomic signature of a 32-kb cosmid in the house finch (carpodacus mexicanus). *Genome Research*, 10(5):613–623.

Hildebrandt, M. A., Salavaggione, O. E., Martin, Y. N., Flynn, H. C., Jalal, S., Wieben, E. D., and Weinshilboum, R. M. (2004). Human SULT1a3 pharmacogenetics: gene duplication and functional genomic studies. *Biochemical and Biophysical Research Communications*, 321(4):870–878.

Högstrand, K. and Böhme, J. (1999). Gene conversion can create new MHC allees. *Immunological Reviews*, 167(1):305–317.

Hollox, E. (2004). Evolutionary genetics: Genetics of lactase persistence – fresh lessons in the history of milk drinking. *European Journal of Human Genetics*, 13(3):267–269.

Hosomichi, K., Miller, M. M., Goto, R. M., Wang, Y., Suzuki, S., Kulski, J. K., Nishibori, M., Inoko, H., Hanzawa, K., and Shiina, T. (2008). Contribution of mutation, recombination, and gene conversion to chicken mhc-b haplotype diversity. *The Journal of Immunology*, 181(5):3393–3399.

Howe, K., Schiffer, P. H., Zielinski, J., Wiehe, T., Laird, G. K., Marioni, J. C., Soylemez, O., Kondrashov, F., and Leptin, M. (2016). Structure and evolutionary history of a large family of NLR proteins in the zebrafish. *Open Biology*, 6(4):160009.

Hübner, S. (2022). Are we there yet? driving the road to evolutionary graph-pangenomics. *Current Opinion in Plant Biology*, 66:102195.

Hudson, R. R. (1983). Properties of a neutral allele model with intragenic recombination. *Theoretical Population Biology*, 23(2):183–201.

Hudson, R. R. (1998). Island models and the coalescent process. *Molecular Ecology*, 7(4):413–418.

Hui, Y. and Liu, M.-C. (2015). Sulfation of ritodrine by the human cytosolic sulfotransferases (SULTs): Effects of SULT1a3 genetic polymorphism. *European Journal of Pharmacology*, 761:125–129.

Inchley, C. E., Larbey, C. D. A., Shwan, N. A. A., Pagani, L., Saag, L., Antao, T., Jacobs, G., Hudjashov, G., Metspalu, E., Mitt, M., Eichstaedt, C. A., Malyarchuk, B., Derenko, M., Wee, J., Abdullah, S., Ricaut, F.-X., Mormina, M., Magi, R., Villems, R., Metspalu, M., Jones, M. K., Armour, J. A. L., and Kivisild, T. (2016). Selective sweep on human amylase genes postdates the split with neanderthals. *Scientific Reports*, 6(1).

Ingvarsson, P. K. and Whitlock, M. C. (2000). Heterosis increases the effective migration rate. *Proceedings of the Royal Society of London. Series B: Biological Sciences*, 267(1450):1321–1326.

Innan, H. (2009). Population genetic models of duplicated genes. *Genetica*, 137(1):19–37.

Innan, H. and Kondrashov, F. (2010). The evolution of gene duplications: classifying and distinguishing between models. *Nature Reviews Genetics*, 11(2):97–108.

Iskow, R. C., Gokcumen, O., and Lee, C. (2012). Exploring the role of copy number variants in human adaptation. *Trends in Genetics*, 28(6):245–257.

Jain, K. (2008). Loss of least-loaded class in asexual populations due to drift and epistasis. *Genetics*, 179(4):2125–2134.

Jespersgaard, C., Fode, P., Dybdahl, M., Vind, I., Nielsen, O. H., Csillag, C., Munkholm, P., Vainer, B., Riis, L., Elkjaer, M., Pedersen, N., Knudsen, E., and Andersen, P. S. (2011). Alpha-defensin DEFA1a3 gene copy number elevation in danish crohn's disease patients. *Digestive Diseases and Sciences*, 56(12):3517–3524.

Jones, J., Vance, R., and Dangl, J. (2016). Intracellular innate immune surveillance devices in plants and animals. *Science*, 354(6316).

Kim, Y. and Stephan, W. (2002). Detecting a local signature of genetic hitchhiking along a recombining chromosome. *Genetics*, 160(2):765–777.

Kimura, M. (1962). On the probability of fixation of mutant genes in a population. *Genetics*, 47(6):713–719.

Kimura, M. (1964). Diffusion models in population genetics. *Journal of Applied Probability*, 1(2):177–232.

Kimura, M. (1969). The number of heterozygous nucleotide sites maintained in a finite population due to steady flux of mutations. *Genetics*, 61(4):893–903.

Kimura, M. (1971). Theoretical foundation of population genetics at the molecular level. *Theoretical Population Biology*, 2(2):174–208.

Kimura, M. and Crow, J. F. (1964). The number of alleles that can be maintained in a finite population. *Genetics*, 49(4):725–738.

Kingman, J. F. C. (1982a). The coalescent. *Stochastic Processes and their Applications*, 13(3):235–248.

Kingman, J. F. C. (1982b). On the genealogy of large populations. *Journal of Applied Probability*, 19(A):27–43.

Kondrashov, A. S. (1982). Selection against harmful mutations in large sexual and asexual populations. *Genetics Research*, 40(3):325–332.

Kondrashov, A. S. (1994). Mullers ratchet under epistatic selection. *Genetics*, 136(4):1469–1473.

Kondrashov, F. A. (2012). Gene duplication as a mechanism of genomic adaptation to a changing environment. *Proceedings of the Royal Society B: Biological Sciences*, 279(1749):5048–5057.

Kondrashov, F. A. and Koonin, E. V. (2004). A common framework for understanding the origin of genetic dominance and evolutionary fates of gene duplications. *Trends in Genetics*, 20(7):287–290.

Kondrashov, F. A., Rogozin, I. B., Wolf, Y. I., and Koonin, E. V. (2002). *Genome Biology*, 3(2):research0008.1.

Kreitman, M. (1983). Nucleotide polymorphism at the alcohol dehydrogenase locus of drosophila melanogaster. *Nature*, 304(5925):412–417.

Krüger, J. and Vogel, F. (1975). Population genetics of unequal crossing over. *Journal of Molecular Evolution*, 4(3):201–247.

Lam, T. H., Shen, M., Chia, J.-M., Chan, S. H., and Ren, E. C. (2013). Population-specific recombination sites within the human MHC region. *Heredity*, 111(2):131–138.

Lenz, T. L., Wells, K., Pfeiffer, M., and Sommer, S. (2009). Diverse MHC IIB allele repertoire increases parasite resistance and body condition in the long-tailed giant rat (leopoldamys sabanus). *BMC Evolutionary Biology*, 9(1):269.

Levin, D. A. and Peres, Y. (2017). *Markov chains and mixing times*, volume 107. American Mathematical Soc.

Li, H. and Durbin, R. (2011). Inference of human population history from individual whole-genome sequences. *Nature*, 475(7357):493–496.

Liao, W.-W., Asri, M., Ebler, J., Doerr, D., Haukness, M., Hickey, G., Lu, S., Lucas, J. K., Monlong, J., Abel, H. J., Buonaiuto, S., Chang, X. H., Cheng, H., Chu, J., Colonna, V., Eizenga, J. M., Feng, X., Fischer, C., Fulton, R. S., Garg, S., Groza, C., Guarracino, A., Harvey, W. T., Heumos, S., Howe, K., Jain, M., Lu, T.-Y., Markello, C., Martin, F. J., Mitchell, M. W., Munson, K. M., Mwaniki, M. N., Novak, A. M., Olsen, H. E., Pesout, T., Porubsky, D., Prins, P., Sibbesen, J. A., Siren, J., Tomlinson, C., Villani, F., Vollger, M. R., Antonacci-Fulton, L. L., Baid, G., Baker, C. A., Belyaeva, A., Billis, K., Carroll, A., Chang, P.-C., Cody, S., Cook, D. E., Cook-Deegan, R. M., Cornejo, O. E., Diekhans, M., Ebert, P., Fairley, S., Fedrigo, O., Felsenfeld, A. L., Formenti, G., Frankish, A., Gao, Y., Garrison, N. A., Giron, C. G., Green, R. E., Haggerty, L., Hoekzema, K., Hourlier, T., Ji, H. P., Kenny, E. E., Koenig, B. A., Kolesnikov, A., Korbel, J. O., Kordosky, J., Koren, S., Lee, H., Lewis, A. P., Magalhaes, H., Marco-Sola, S., Marijon, P., McCartney, A., McDaniel, J., Mountcastle, J., Nattestad, M., Nurk, S., Olson, N. D., Popejoy, A. B., Puiu, D., Rautiainen, M., Regier, A. A., Rhie, A., Sacco, S., Sanders, A. D., Schneider, V. A., Schultz, B. I., Shafin, K., Smith, M. W., Sofia, H. J., Tayoun, A. N. A., Thibaud-Nissen, F., Tricomi, F. F., Wagner, J., Walenz, B., Wood, J. M. D., Zimin, A. V., Bourque, G., Chaisson, M. J. P., Flicek, P., Phillippy, A. M., Zook, J. M., Eichler, E. E., Haussler, D., Wang, T., Jarvis, E. D., Miga, K. H., Garrison, E., Marschall, T., Hall, I. M., Li, H., and Paten, B. (2023). A draft human pangenome reference. *Nature*, 617(7960):312–324.

Linnenbrink, M., Teschke, M., Montero, I., Vallier, M., and Tautz, D. (2018). Meta-populational demes constitute a reservoir for large MHC allele diversity in wild house mice (mus musculus). *Frontiers in Zoology*, 15(1).

Liu, L., Yu, L., Kalavacharla, V., and Liu, Z. (2011). A bayesian model for gene family evolution. *BMC bioinformatics*, 12(1):1–10.

Lohmueller, K. E. (2014). The impact of population demography and selection on the genetic architecture of complex traits. *PLoS Genetics*, 10(5):e1004379.

Magadum, S., Gangapur, U. B. P. M. D., and Ravikesavan, R. (2013). Gene duplication as a major force in evolution. *Journal of Genetics*, 92(1):155–161.

Manczinger, M., Boross, G., Kemény, L., Müller, V., Lenz, T. L., Papp, B., and Pál, C. (2019). Pathogen diversity drives the evolution of generalist MHC-II alleles in human populations. *PLOS Biology*, 17(1):e3000131.

Martinsohn, J. T., Sousa, A. B., Guethlein, L. A., and Howard, J. C. (1999). The gene conversion hypothesis of MHC evolution: a review. *Immunogenetics*, 50(3-4):168–200.

Mason, R. A. B., Browning, T. L., and Eldridge, M. D. B. (2009). Reduced MHC class II diversity in island compared to mainland populations of the black-footed rock-wallaby (petrogale lateralis lateralis). *Conservation Genetics*, 12(1):91–103.

Maxam, A. M. and Gilbert, W. (1977). A new method for sequencing DNA. *Proceedings of the National Academy of Sciences*, 74(2):560–564.

McVean, G. A. and Cardin, N. J. (2005). Approximating the coalescent with recombination. *Philosophical Transactions of the Royal Society B: Biological Sciences*, 360(1459):1387–1393.

Menashe, I., Aloni, R., and Lancet, D. (2006). A probabilistic classifier for olfactory receptor pseudogenes. *BMC Bioinformatics*, 7(1).

Mendel, G. (1865). Versuche uber pflanzen-hybriden. *Vorgelegt in den Sitzungen*.

Milesi, P., Weill, M., Lenormand, T., and Labbé, P. (2017). Heterogeneous gene duplications can be adaptive because they permanently associate overdominant alleles. *Evolution letters*, 1(3):169–180.

Miller, H. C. and Lambert, D. M. (2004). Genetic drift outweighs balancing selection in shaping post-bottleneck major histocompatibility complex variation in new zealand robins (petroicidae). *Molecular Ecology*, 13(12):3709–3721.

Muller, H. (1964). The relation of recombination to mutational advance. *Mutation Research/Fundamental and Molecular Mechanisms of Mutagenesis*, 1(1):2–9.

Nadeau, J. H. and Sankoff, D. (1997). Comparable rates of gene loss and functional divergence after genome duplications early in vertebrate evolution. *Genetics*, 147(3):1259–1266.

Nassar, H., Lavi, E., Akkawi, S., Bdeir, K., Heyman, S. N., Raghunath, P., Tomaszewski, J., and Higazi, A. A.-R. (2007). α-defensin: Link between inflammation and atherosclerosis. *Atherosclerosis*, 194(2):452–457.

Nordborg, M. (1997). Structured coalescent processes on different time scales. *Genetics*, 146(4):1501–1514.

Noskova, E., Ulyantsev, V., Koepfli, K.-P., O'Brien, S. J., and Dobrynin, P. (2020). GADMA: Genetic algorithm for inferring demographic history of multiple populations from allele frequency spectrum data. *GigaScience*, 9(3).

Notohara, M. (1990). The coalescent and the genealogical process in geographically structured population. *Journal of mathematical biology*, 29(1):59–75.

Ohno, S. (1970). *Evolution by gene duplication*. Springer, 1st edition.

Ohta, T. (1973). Slightly deleterious mutant substitutions in evolution. *Nature*, 246(5428):96–98.

Ohta, T. (1976). Simple model for treating evolution of multigene families. *Nature*, 263(5572):74–76.

Ohta, T. (1979). An extension of a model for the evolution of multigene families by unequal crossing over. *Genetics*, 91(3):591–607.

Ohta, T. (1984). Multigene families and their implications for evolutionary theory. In *Synergetics — From Microscopic to Macroscopic Order*, pages 133–139. Springer Berlin Heidelberg.

Ohta, T. (1987). Simulating evolution by gene duplication. *Genetics*, 115(1):207–213.

Ohta, T. (1988). Further simulation studies on evolution by gene duplication. *Evolution*, 42(2):375–386.

Ohta, T. (2000). Evolution of gene families. *Gene*, 259(1-2):45–52.

Okazaki, A., Yamazaki, S., Inoue, I., and Ott, J. (2020). Population genetics: past, present, and future. *Human Genetics*, 140(2):231–240.

Otto, M. and Wiehe, T. (2023). The structured coalescent in the context of gene copy number variation. *Theoretical Population Biology*, 154:67–78.

Otto, M., Zheng, Y., Grablowitz, P., and Wiehe, T. (2023). Distinguishing the roles of adaptation and demography in gene copy number changes in human populations.

Otto, M., Zheng, Y., and Wiehe, T. (2022). Recombination, selection, and the evolution of tandem gene arrays. *Genetics*, 221(3).

Pajic, P., Pavlidis, P., Dean, K., Neznanova, L., Romano, R.-A., Garneau, D., Daugherity, E., Globig, A., Ruhl, S., and Gokcumen, O. (2019). Independent amylase gene copy number bursts correlate with dietary preferences in mammals. *eLife*, 8.

Panchy, N., Lehti-Shiu, M., and Shiu, S.-H. (2016). Evolution of gene duplication in plants. *Plant Physiology*, 171(4):2294–2316.

Perry, G. H., Dominy, N. J., Claw, K. G., Lee, A. S., Fiegler, H., Redon, R., Werner, J., Villanea, F. A., Mountain, J. L., Misra, R., Carter, N. P., Lee, C., and Stone, A. C. (2007). Diet and the evolution of human amylase gene copy number variation. *Nature Genetics*, 39(10):1256–1260.

Petit, M., Astruc, J.-M., Sarry, J., Drouilhet, L., Fabre, S., Moreno, C. R., and Servin, B. (2017). Variation in recombination rate and its genetic determinism in sheep populations. *Genetics*, 207(2):767–784.

Qian, W. and Zhang, J. (2014). Genomic evidence for adaptation by gene duplication. *Genome Research*, 24(8):1356–1362.

Rabiner, L. R. (1989). A tutorial on hidden markov models and selected applications in speech recognition. *Proceedings of the IEEE*, 77(2):257–286.

Rafajlović, M., Klassmann, A., Eriksson, A., Wiehe, T., and Mehlig, B. (2014). Demography-adjusted tests of neutrality based on genome-wide snp data. *Theoretical Population Biology*, 95:1–12.

Rasmussen, M. D., Hubisz, M. J., Gronau, I., and Siepel, A. (2014). Genome-wide inference of ancestral recombination graphs. *PLoS Genetics*, 10(5):e1004342.

Raymond, M., Berticat, C., Weill, M., Pasteur, N., and Chevillon, C. (2001). Insecticide resistance in the mosquito culex pipiens: What have we learned about adaptation? In *Microevolution Rate, Pattern, Process*, pages 287–296. Springer Netherlands.

Redon, R., Ishikawa, S., Fitch, K. R., Feuk, L., Perry, G. H., Andrews, T. D., Fiegler, H., Shapero, M. H., Carson, A. R., Chen, W., et al. (2006). Global variation in copy number in the human genome. *Nature*, 444(7118):444–454.

Ridley, M. (2013). *Evolution: Probleme—Themen—Fragen*. Springer-Verlag.

Roux, C., Castric, V., Pauwels, M., Wright, S. I., Saumitou-Laprade, P., and Vekemans, X. (2011). Does speciation between arabidopsis halleri and arabidopsis lyrata coincide with major changes in a molecular target of adaptation? *PLoS ONE*, 6(11):e26872.

Rouzine, I. M., Brunet, E., and Wilke, C. O. (2008). The traveling-wave approach to asexual evolution: Muller's ratchet and speed of adaptation. *Theoretical Population Biology*, 73(1):24–46.

Sanger, F., Nicklen, S., and Coulson, A. R. (1977). DNA sequencing with chain-terminating inhibitors. *Proceedings of the National Academy of Sciences*, 74(12):5463–5467.

Schaschl, H., Wandeler, P., Suchentrunk, F., Obexer-Ruff, G., and Goodman, S. J. (2006). Selection and recombination drive the evolution of MHC class II DRB diversity in ungulates. *Heredity*, 97(6):427–437.

Schierup, M. H., Vekemans, X., and Charlesworth, D. (2000). The effect of subdivision on variation at multi-allelic loci under balancing selection. *Genetical Research*, 76(1):51–62.

Schiffels, S. and Durbin, R. (2014). Inferring human population size and separation history from multiple genome sequences. *Nature Genetics*, 46(8):919–925.

Schiffels, S. and Wang, K. (2020). MSMC and MSMC2: The multiple sequentially markovian coalescent. In *Methods in Molecular Biology*, pages 147–166. Springer US.

Sebat, J., Lakshmi, B., Troge, J., Alexander, J., Young, J., Lundin, P., anér, S. M., Massa, H., Walker, M., Chi, M., Navin, N., Lucito, R., Healy, J., Hicks, J., Ye, K., Reiner, A., Gilliam, T. C., Trask, B., Patterson, N., Zetterberg, A., and Wigler, M. (2004). Large-scale copy number polymorphism in the human genome. *Science*, 305(5683):525–528.

Shen, S., Li, H., Liu, J., Sun, L., and Yuan, Y. (2020). The panoramic picture of pepsinogen gene family with pan-cancer. *Cancer Medicine*, 9(23):9064–9080.

Silver, L. (2001). Evolution of gene families. In Brenner, S. and Miller, J. H., editors, *Encyclopedia of Genetics*, pages 666–669. Academic Press, New York.

Singleton, A. B., Farrer, M., Johnson, J., Singleton, A., Hague, S., Kachergus, J., Hulihan, M., Peuralinna, T., Dutra, A., Nussbaum, R., Lincoln, S., Crawley, A., Hanson, M.,

Maraganore, D., Adler, C., Cookson, M. R., Muenter, M., Baptista, M., Miller, D., Blancato, J., Hardy, J., and Gwinn-Hardy, K. (2003). α-synuclein locus triplication causes parkinson's disease. *Science*, 302(5646):841–841.

Smith, G. P. (1974). Unequal crossover and the evolution of multigene families. In *Cold Spring Harbor Symposia on Quantitative Biology*, volume 38, pages 507–513. Cold Spring Harbor Laboratory Press.

Smith, J. M. and Haigh, J. (1974). The hitch-hiking effect of a favourable gene. *Genetical Research*, 23(1):23–35.

Spence, J. P. and Song, Y. S. (2019). Inference and analysis of population-specific fine-scale recombination maps across 26 diverse human populations. *Science Advances*, 5(10):eaaw9206.

Stajich, J. E. (2004). Disentangling the effects of demography and selection in human history. *Molecular Biology and Evolution*, 22(1):63–73.

Stephan, W. and Hörger, A. C. (2019). *Molekulare Populationsgenetik*. Springer Berlin Heidelberg.

Sudmant, P. H., Kitzman, J. O., Antonacci, F., Alkan, C., Malig, M., Tsalenko, A., Sampas, N., Bruhn, L., Shendure, J., and and, E. E. E. (2010). Diversity of human copy number variation and multicopy genes. *Science*, 330(6004):641–646.

Sudmant, P. H., Mallick, S., Nelson, B. J., Hormozdiari, F., Krumm, N., Huddleston, J., Coe, B. P., Baker, C., Nordenfelt, S., Bamshad, M., Jorde, L. B., Posukh, O. L., Sahakyan, H., Watkins, W. S., Yepiskoposyan, L., Abdullah, M. S., Bravi, C. M., Capelli, C., Hervig, T., Wee, J. T. S., Tyler-Smith, C., van Driem, G., Romero, I. G., Jha, A. R., Karachanak-Yankova, S., Toncheva, D., Comas, D., Henn, B., Kivisild, T., Ruiz-Linares, A., Sajantila, A., Metspalu, E., Parik, J., Villems, R., Starikovskaya, E. B., Ayodo, G., Beall, C. M., Rienzo, A. D., Hammer, M. F., Khusainova, R., Khusnutdinova, E., Klitz, W., Winkler, C., Labuda, D., Metspalu, M., Tishkoff, S. A., Dryomov, S., Sukernik, R., Patterson, N., Reich, D., and Eichler, E. E. (2015a). Global diversity, population stratification, and selection of human copy-number variation. *Science*, 349(6253).

Sudmant, P. H., Rausch, T., Gardner, E. J., Handsaker, R. E., Abyzov, A., Huddleston, J., Zhang, Y., Ye, K., Jun, G., Fritz, M. H.-Y., Konkel, M. K., Malhotra, A., Stutz, A. M., Shi, X., Casale, F. P., Chen, J., Hormozdiari, F., Dayama, G., Chen, K., Malig, M., Chaisson, M. J. P., Walter, K., Meiers, S., Kashin, S., Garrison, E., Auton, A., Lam, H. Y. K., Mu, X. J., Alkan, C., Antaki, D., Bae, T., Cerveira, E., Chines, P., Chong, Z., Clarke, L., Dal, E., Ding, L., Emery, S., Fan, X., Gujral, M., Kahveci, F., Kidd, J. M., Kong, Y., Lameijer, E.-W., McCarthy, S., Flicek, P., Gibbs, R. A., Marth, G., Mason, C. E., Menelaou, A., Muzny, D. M., Nelson, B. J., Noor, A., Parrish, N. F., Pendleton, M., Quitadamo, A., Raeder, B., Schadt, E. E., Romanovitch, M., Schlattl, A., Sebra, R., Shabalin, A. A., Untergasser, A., Walker, J. A., Wang, M., Yu, F., Zhang, C., Zhang, J., Zheng-Bradley, X., Zhou, W., Zichner, T., Sebat, J., Batzer, M. A., McCarroll, S. A., Mills, R. E., Gerstein, M. B., Bashir, A., Stegle, O., Devine, S. E., Lee, C., Eichler, E. E., and and, J. O. K. (2015b). An integrated map of structural variation in 2, 504 human genomes. *Nature*, 526(7571):75–81.

Sugino, R. and Innan, H. (2006). Selection for more of the same product as a force to enhance concerted evolution of duplicated genes. *Trends in Genetics*, 22(12):642–644.

Swallow, D. M. (2003). Genetics of lactase persistence and lactose intolerance. *Annual Review of Genetics*, 37(1):197–219.

Taggart, R. T., Mohandas, T. K., Shows, T. B., and Bell, G. I. (1985). Variable numbers of pepsinogen genes are located in the centromeric region of human chromosome 11 and determine the high-frequency electrophoretic polymorphism. *Proceedings of the National Academy of Sciences*, 82(18):6240–6244.

Tajima, F. (1989). Statistical method for testing the neutral mutation hypothesis by DNA polymorphism. *Genetics*, 123(3):585–595.

Takahata, N. (1981). A mathematical study on the distribution of the number of repeated genes per chromosome. *Genetical Research*, 38(1):97–102.

Takahata, N. (1988). The coalescent in two partially isolated diffusion populations. *Genetical Research*, 52(3):213 – 222.

Tellier, A., Moreno-Gámez, S., and Stephan, W. (2014). Speed of adaptation and genomic footprints of host-parasite coevolution under arms race and trench warfare dynamics. *Evolution*, pages 2211–2224.

Thomae, B. A., Rifki, O. F., Theobald, M. A., Eckloff, B. W., Wieben, E. D., and Weinshilboum, R. M. (2004). Human catecholamine sulfotransferase (SULT1a3) pharmacogenetics: functional genetic polymorphism. *Journal of Neurochemistry*, 87(4):809–819.

Tishkoff, S. A., Reed, F. A., Ranciaro, A., Voight, B. F., Babbitt, C. C., Silverman, J. S., Powell, K., Mortensen, H. M., Hirbo, J. B., Osman, M., Ibrahim, M., Omar, S. A., Lema, G., Nyambo, T. B., Ghori, J., Bumpstead, S., Pritchard, J. K., Wray, G. A., and Deloukas, P. (2006). Convergent adaptation of human lactase persistence in africa and europe. *Nature Genetics*, 39(1):31–40.

Traherne, J. A. (2008). Human MHC architecture and evolution: implications for disease association studies. *International Journal of Immunogenetics*, 35(3):179–192.

Tuzun, E., Sharp, A. J., Bailey, J. A., Kaul, R., Morrison, V. A., Pertz, L. M., Haugen, E., Hayden, H., Albertson, D., Pinkel, D., Olson, M. V., and E, E. E. (2005). Fine-scale structural variation of the human genome. *Nature Genetics*, 37(7):727–732.

Usher, C. L., Handsaker, R. E., Esko, T., Tuke, M. A., Weedon, M. N., Hastie, A. R., Cao, H., Moon, J. E., Kashin, S., Fuchsberger, C., Metspalu, A., Pato, C. N., Pato, M. T., McCarthy, M. I., Boehnke, M., Altshuler, D. M., Frayling, T. M., Hirschhorn, J. N., and McCarroll, S. A. (2015). Structural forms of the human amylase locus and their relationships to SNPs, haplotypes and obesity. *Nature Genetics*, 47(8):921–925.

Vahdati, A. R. and Wagner, A. (2016). Parallel or convergent evolution in human population genomic data revealed by genotype networks. *BMC evolutionary biology*, 16(1):1–19.

Veitia, R. A. (2005). Gene dosage balance: deletions, duplications and dominance. *Trends in Genetics*, 21(1):33–35.

Wagner, G. P. and Gabriel, W. (1990). Quantitative variation in finite parthenogenetic populations: What stops muller's ratchet in the absence of recombination? *Evolution*, 44(3):715–731.

Wahlund, S. (1928). Zusammensetzung von populationen und korrelationserscheinungen vom standpunkt der vererbungslehre aus betrachtet. *Hereditas*, 11(1):65–106.

Wakeley, J. (2001). The coalescent in an island model of population subdivision with variation among demes. *Theoretical population biology*, 59(2):133–144.

Wakeley, J. (2016). *Coalescent Theory: An Introduction*. Macmillan Learning.

Wan, T., Liu, Z., Leitch, I., Xin, H., Maggs-Kolling, G., Gong, Y., Li, Z., Marais, E., Liao, Y., Dai, C., Liu, F., Wu, Q., Song, C., Zhou, Y., Huang, W., Jiang, K., Wang, Q., Yang, Y., Zhong, Z., Yang, M., Yan, X., Hu, G., Hou, C., Su, Y., Feng, S., Yang, J., Yan, J., Chu, J., Chen, F., Ran, J., Wang, X., Van de Peer, Y., Leitch, A., and Wang, Q. (2021). The welwitschia genome reveals a unique biology underpinning extreme longevity in deserts. *Nat Commun*, 12(1):4247.

Wang, Y., Zhao, Z., Miao, X., Wang, Y., Qian, X., Chen, L., Wang, C., and Li, S. (2022). eSMC: a statistical model to infer admixture events from individual genomics data. *BMC Genomics*, 23(S4).

Watterson, G. (1975). On the number of segregating sites in genetical models without recombination. *Theoretical Population Biology*, 7(2):256–276.

Webel, K. and Wied, D. (2016). *Stochastische Prozesse*. Springer Fachmedien Wiesbaden.

Weinberg, W. (1908). Uber den nachweis der vererbung beim menschen. *Jh. Ver. vaterl. Naturk. Wurttemb.*, 64:369–382.

Wiehe, T., Mountain, J., Parham, P., and Slatkin, M. (2000). Distinguishing recombination and intragenic gene conversion by linkage disequilibrium patterns. *Genetics Research*, 75(1):61–73.

Wilkinson-Herbots, H. M. (1998). Genealogy and subpopulation differentiation under various models of population structure. *Journal of Mathematical Biology*, 37(6):535–585.

Wright, S. (1931). Evolution in mendelian populations. *Genetics*, 16(2):97–159.

Wright, S. (1940). Breeding structure of populations in relation to speciation. *The American Naturalist*, 74(752):232–248.

Wright, S. (1951). The genetical structure of populations. *Annals of Eugenics*, 15(1):323–354.

Zeyl, C., Mizesko, M., and Visser, J. A. G. M. D. (2007). Mutational meltdown in laboratory yeast populations. *Evolution*, 55(5):909–917.

Zhang, J. (2003). Evolution by gene duplication: an update. *Trends in Ecology; Evolution*, 18(6):292–298.

Zhao, J., Gu, Q., Wang, L., Xu, W., Chu, L., Wang, Y., Li, Z., Wu, S., Xu, J., Hu, Z., Shu, Q., and Fang, X. (2018). Low-copy number polymorphism in DEFA1/DEFA3 is associated with susceptibility to hospital-acquired infections in critically ill patients. *Mediators of Inflammation*, 2018:1–8.