Chapter 1

A Population Genomics Lexicon

Gustavo V. Barroso, Ana Filipa Moutinho, and Julien Y. Dutheil

Abstract

Population genomics is a growing field stemming from soon a 100 years of developments in population genetics. Here, we summarize the main concepts and terminology underlying both theoretical and empirical statistical population genomics studies. We provide the reader with pointers toward the original literature as well as methodological and historical reviews.

Key words Population genetics, Neutral theory, Coalescent theory, Mutation, Recombination, Selection, Lexicon

1 Genomic Variation

1.1 Loci, Alleles, and Polymorphism

Population genomics studies the evolution of genome variants in populations. A *locus* (*pl. loci*) refers to a given location in the genome. The particular sequence at a given locus may vary between individuals, each variant being termed an *allele*. We call loci with at least two alleles *polymorphic* and invariant loci *monomorphic*. The term *polymorphism* refers to the presence of multiple alleles but is commonly used as a countable noun as a substitute for “polymorphic locus” (*one polymorphism, several polymorphisms*).

Alleles may differ because of the nucleotide content, but also in length, as a result of nucleotide insertions or deletions (*a.k.a. indels*). Variable loci of length one can have up to four distinct alleles (A, C, G, or T) and are termed *single nucleotide polymorphisms* (SNPs). SNPs constitute, so far, the majority of the data accounted for by population genetic models.

1.2 Mutations

Molecular events altering the genome are termed *mutations*. Mutations include substitution of a nucleotide into another one, removal or addition of one or several nucleotides, as well as multiplication of some part of the genome. Mutation is the process by which new

Authors Gustavo V. Barroso and Ana Filipa Moutinho contributed equally to this work.

Julien Y. Dutheil (ed.), *Statistical Population Genomics*, Methods in Molecular Biology, vol. 2090, https://doi.org/10.1007/978-1-0716-0199-0_1, © The Author(s) 2020

alleles are formed. The *infinite site model* assumes that during the timeframe of evolution modeled, each locus have undergone at most one mutation [1–3]. This model also implies that each mutation creates a new allele in the population and that there is no “backward” or “reverse” mutation. The infinite site model is a generally reasonable assumption as the mutation rate is typically low and genomes are large. It might be locally invalidated, however, in case of mutation hotspots or when larger evolutionary timescales are considered. Under this premise, at most two alleles are expected per locus. Loci with two alleles are termed *diallelic* or *biallelic*, the first term having historical precedence and being more accurate [4], while the second is more commonly used since the 1990s. Furthermore, in a population genomic dataset, a sampled diallelic locus is called a *singleton* if one of the two alleles is present in only one haploid genome, and a *doubleton* if it is present in precisely two haploid genomes.

1.3 The Wright–Fisher Model

The simplest process of allele evolution within a single population is named the *Wright–Fisher model*. It describes the evolution of alleles in a population of fixed and constant size, where all alleles have the same fitness, and therefore the same chance to be transmitted to the next generation (*neutral evolution*). The population is assumed to be *panmictic*, that is, individuals are randomly mating. Time is discretized in *non-overlapping generations* so that the alleles in the current generation are a random sample of the alleles from the previous generation, without new alleles being generated by mutation. Under such conditions, allelic frequencies evolve only because of the stochasticity in the sampling of gametes that will contribute to the next generation, a process termed *genetic drift*. Because populations are of finite size, alleles will be sampled at their actual frequencies on average only and the ultimate fate of any allele is either to reach frequency zero in the population and be lost, when by chance no individual carrying this allele has any descendant in the next generation or to become fixed when all other alleles have been lost. The time until fixation depends on the population size: smaller populations will show a stronger sampling effect and shorter times to fixation. When genetic drift is the only force acting on a population, the number of alleles at a given locus is necessarily decreasing over time.

The *Wright–Fisher model with mutation* extends the Wright–Fisher model by introducing new alleles in the population, at a given rate. As the mutation rate is low, new mutations appear in a single copy, their initial frequency is then $1/2N$ in a diploid population. Mutation and drift act in opposite direction and a *mutation-drift equilibrium* is reached when the rate of allele creation by mutation equals the rate of allele loss by drift. The genetic diversity is then determined by the sole product of the population size N and

the mutation rate u . Under the infinite site model, the expected heterozygosity at a locus in a population of diploid individuals is approximated by [1]

$$\hat{h} = \frac{4 \cdot N \cdot u}{4 \cdot N \cdot u + 1}$$

while the expected number of distinct alleles and their respective frequencies can be estimated using *Ewens's sampling formula* [5].

A *substitution* occurs when a new mutation has spread in the population, increasing from frequency $1/(2N)$ to 1 (*see Note 1*). Kimura showed that the average time to fixation of a new mutation is $4N$ in a population of diploid individuals [6]. Furthermore, as a neutral mutation has a probability of reaching fixation equal to $1/(2N)$ and given that there are $2N \cdot u$ new mutations per generation, in a purely neutrally evolving population, the expected number of substitutions per generation is equal to $2N \cdot u \cdot 1/(2N) = u$. The substitution rate is therefore independent of the population size and, assuming that the mutation rate is constant in time, the number of substitutions between two populations is a direct measure of the number of generations separating them, a phenomenon termed *molecular clock* [7].

1.4 The Backward Wright–Fisher Model: The Standard Coalescent

While the Wright–Fisher process naturally describes the evolution of sequences within populations one generation after the other, population genetic data typically represent individuals sampled at a given time point. For inference purposes, it is therefore convenient to model the history of the genetic material that gave rise to the sample. The modelization of the ancestry of a sample (also known as the *genealogy*) is typically done backward in time, as every locus find a common ancestor in the past, until the *most recent common ancestor (MRCA)* of the sample. The merging of two lineages in the past is called a *coalescence event*, and the set of mathematical tools describing this process under a variety of demographic models is referred to as the *coalescence theory*. Kingman [8] first described the *standard coalescent*, the genealogical model corresponding to the Wright–Fisher model (but *see refs. 9 and 10* for a historical perspective). The standard coalescent is, therefore, also referred to as the *Kingman's coalescent*.

2 Beyond the Wright–Fisher Model

The Wright–Fisher model has been extended in several ways to include more realistic assumptions on the underlying evolutionary process. These extensions led to the concept of *Effective population size* (N_e), originally defined as the number of individuals contributing to the gene pool. When a population deviates from the assumptions of the Wright–Fisher model, N_e is no longer equal to the census population size (N). Often (but not always) in such cases,

N_e can be obtained by a linear scaling of N such that it reflects the number of individuals from an idealized Wright–Fisher population that would display the same genetic diversity as the actual population under study [11].

2.1 Demography

A possible deviation from the Wright–Fisher assumptions happens when the population size is not constant across generations. The term *demographic history* generally refers to the collection of demographic parameters (effective sizes, growth rates) that describes the history of the population until its most recent common ancestor [12]. When population size varies in a cyclic manner with relatively small period n generations, the resulting genealogies can be modeled by a Wright–Fisher process with a population size equal to the harmonic mean of the historical population sizes, so that

$$N_e = \frac{n}{\sum_i^n \frac{1}{N_i}},$$

where N_i refer to the i th population size [13]. More drastic demographic effects include *genetic bottlenecks*, corresponding to a sharp decrease (shrinkage) in population size.

2.2 Population Structure

In the absence of *panmixia*, genetic exchanges occur more often between certain individuals, resulting in *population structure* with several subpopulations. Population structure may occur for different reasons such as overlapping generations, assortative mating, or geographic isolation [12]. *Assortative mating* occurs when individuals choose their mates according to some similarity between their phenotypes. If the phenotype is genetically determined, assortative mating can influence the level of heterozygosity in the population [14].

Gene flow describes the migration of genetic variants between subpopulations under a scenario of population structure. It reduces genetic differentiation among subpopulations [15]. Ultimately, subpopulations can diverge and become genetically isolated, a process called *speciation*. The simplest speciation processes involve spontaneous isolation (*isolation model*) or spontaneous isolation followed by a period of gene flow (*isolation with migration model*) [16].

When speciation events occur in a short timeframe and ancestral population sizes are large, ancestral polymorphism may persist in the ancestral species, a phenomenon called *incomplete lineage sorting (ILS)* [17]. The expected amount of ILS depends on the number of generations between two isolation events (Δ_T) and the ancestral effective population size N_{e_A} [18]:

$$\Pr(ILS) = \frac{2}{3} e^{\left(-\frac{2 \cdot \Delta_T}{N_{e_A}}\right)}$$

The term *introgression* is used to depict the transfer of genetic material between diverged populations or species through secondary contact [19]. As a result, extant lineages share a common ancestor that predates the two isolation or speciation events. The resulting genealogy may, therefore, be incongruent with the phylogeny defined by the two splits, depending on the order of coalescence events between lineages [20].

3 Statistics on Nucleotide Diversity

Statistics are needed to infer population genetics parameters from polymorphism data. The *site frequency spectrum* (SFS) describes the empirical distribution of allele frequencies across segregating sites of a given (set of) loci in a population sample. For a sample of n sequences (in n haploid individuals or $n/2$ diploid individuals), the so-called unfolded SFS is the set of counts of derived alleles $X = (X_1, X_2, \dots, X_{n-1})$, where sample configurations X_i denote the number of sites that have $n - i$ ancestral and i derived alleles. The ancestral state is usually estimated using an outgroup sequence. In cases where we cannot assess the ancestral allele, the folded site frequency spectrum, X' , may be calculated instead. X' represents the distribution of the minor allele frequencies, such as $X'_i = X_i + X_{n-i}$ for $i < n/2$ and $X'_{n/2} = X_{n/2}$ [13, 21, 22]. The shape of the SFS is affected by underlying population genetic processes, such as demography and selection, and therefore serves as the input of many population genetics methods [23] (see Fig. 1).

Watterson's theta, here noted $\hat{\theta}_S$, is an estimator of the population mutation rate $\theta = 4Ne \cdot u$, where Ne is the (diploid) effective population size and u the mutation rate. It is derived from the number of segregating sites S_n of a sample of size n [25]. Assuming an infinite sites model, S_n is equal to the product of u and the expected time to coalescence, corrected by the sample size:

$$E[S_n] = u \cdot 4 \cdot Ne \sum_{i=1}^{n-1} i.$$

Since $4Ne \cdot u = \theta$ the equation may be written as $E[S_n] = \theta \cdot a_n$, where $a_n = \sum_{i=1}^{n-1} i$. The proposed estimator of θ for the sample is

$$\hat{\theta}_S = \frac{\hat{S}_n}{a_n} = \frac{\hat{S}_n}{\left(1 + \frac{1}{2} + \dots + \frac{1}{n-1}\right)},$$

Fig. 1 Effect of demography on the shape of the site frequency spectrum (SFS). The figure depicts four scenarios: constant population size, exponential growth, genetic bottleneck, and population structure. The red curve shows the expectation under a constant population size. In the case of exponential growth or a genetic bottleneck, the SFS displays an excess of low-frequency variants. Population structure, here simulated as two subpopulations exchanging migrants at a low rate, results in an excess of intermediate frequency variant when we reconstruct a single SFS from the two subpopulations. Simulations were performed using the `msprime` software [24] (see also Chapter 9 and the online companion material)

where $\hat{S}_n$ is the observed number of segregating sites in the sample. In order to be comparable, values of θ are usually reported per site, and $\hat{\theta}_S$ is then further divided by the sequence length L . This estimator is unbiased when the data is generated from a Wright–Fisher process but is not robust to deviations from it, due to selection or demography [26].

Tajima's π , the *average pairwise heterozygosity* is a measure of nucleotide diversity defined as the number of pairwise differences between a set of sequences [27]. Under the infinite sites model, the number of mutations separating two orthologous chromosomes D_{ij} is equal to the number of nucleotide differences between

sequences i and j . As the expectation of the average pairwise nucleotide differences between all pairs of sequences in a sample is equal to $\theta = 4Ne \cdot u$ [28], Tajima's estimator of θ is:

$$\hat{\theta}_{\pi} = \frac{2}{n(n-1) \cdot L} \sum_{i=1}^{n-1} \sum_{j=i+1}^n D_{ij},$$

where L is the total sequence length.

4 Selective Processes

4.1 Protein-Coding Genes

The coding region of a protein-coding gene, also known as *Coding DNA Sequence (CDS)* is the portion of DNA, or RNA, that encodes a protein. A start and stop codons limit the coding region at the five-prime and three-prime end, respectively. In mRNAs, the CDS is bounded by the five-prime untranslated region (5'-UTR) and the three-prime untranslated region (3'-UTR), also included in the exons. Mutations within coding regions are expected to be of distinct types: *synonymous mutations* lead to no change of amino-acid at the protein level due to the redundancy of the genetic code, as opposed to *non-synonymous mutations*. Non-synonymous mutations can further be classified as *conservative* and *non-conservative* (= *radical*), whether they replace an amino-acid by a biochemically similar one or not. Because of the structure of the genetic code, the four types of mutations at one site (toward A, C, G, or T) can be in principle both synonymous and non-synonymous. Sites where n out of four possible mutations are synonymous are called *n-fold degenerated*. *Four-fold degenerated* sites only undergo synonymous mutations, while a mutation at a so-called *zero-fold degenerated* site is necessarily non-synonymous. Most of second codon positions are zero-fold degenerated, while many of the third positions are four-fold degenerated.

4.2 Fitness Effect

The resulting change of fitness at the organism level characterizes the type of mutations: neutral mutations have no impact on the fitness, while harmful or deleterious mutations induce a lower fitness. Conversely, advantageous mutations increase the fitness of the organism compared to the wild-type genotype. There is, however, a wide range of selective effects, which extends the categorization of mutations from strongly deleterious, through weakly deleterious, neutral to mildly and highly adaptive mutations. The relative frequencies of these types of mutations represent the distribution of fitness effects [29, 30].

The *selection coefficient* (s) is a measure of differences in fitness, which determines the changes in genotype frequencies that occur due to selection. It is commonly expressed as a relative fitness. If

one considers a single locus with two alleles A and a , a standard parametrization is to attribute a fitness of 1 to the homozygote AA and relative fitness of $1 + s$ for the homozygote aa . The heterozygote Aa is attributed a fitness of $1 + h \cdot s$, where h is the so-called *coefficient of dominance*. The s parameter varies between -1 and $+\infty$ (but see **Note 2**), wherein values comprised among -1 and 0 are indicative of negative selection, while positive values correspond to positive selection [13, 31]. The efficiency of selection, however, depends on both s and the effective population size, Ne , so that mutations with $Ne \cdot s \ll 1$ behave in effect like neutral mutations, whose fate is determined by genetic drift only [29].

4.3 Types of Selection

Positive selection acts on alleles that increase fitness, raising their frequency in the population over time, while *negative selection* (= *purifying selection*) decreases the frequency of alleles that impair fitness. Both positive and negative selection decrease genetic diversity. Conversely, *balancing selection* acts by maintaining multiple alleles in the gene pool of a population at frequencies higher than expected by drift alone. Three mechanisms are generally acknowledged: *heterozygous advantage*, where heterozygotes have a higher fitness than homozygotes and maintain genetic polymorphism; *frequency-dependent selection*, where the fitness of the genotype is inversely proportional to its frequency in the population; and *environment-dependent fitness* of genotypes (also known as *local adaptation*) [31, 32].

4.4 Inference of Selection in Protein-Coding Sequences

The strength and direction of selection acting on protein-coding regions may be assessed by contrasting the rate of non-synonymous (potentially under selection, dN) to synonymous (assumed to be neutral, dS , but see, for instance, [33]) substitutions between species. In a population of sequences evolving neutrally, all substitutions are neutral and the two rates are equal, leading to a dN/dS ratio equal to one on average. Assuming non-synonymous mutations are either neutral or deleterious while synonymous mutations are always neutral, the rate of non-synonymous substitutions will be lower than the rate of synonymous substitutions, and the dN/dS ratio will be lower than one. Conversely, if non-synonymous mutations are positively selected, their rate of fixation may exceed the rate of synonymous mutation, leading to a higher substitution rate and a dN/dS ratio higher than one.

At the population level, the ratio of non-synonymous (pN) and synonymous (pS) polymorphism is indicative of the strength of purifying selection acting on a protein. Because non-synonymous mutations are more likely to have a negative fitness effect and be counter-selected, they tend to be removed from the population by purifying selection or segregate at low-frequency. We can estimate the synonymous and non-synonymous genetic diversity by computing the average pairwise heterozygosity π separately for

non-synonymous and synonymous mutations, noted π_N and π_S , respectively. The π_N/π_S ratio is therefore generally below one, the stronger the purifying selection, the closer the ratio is to zero.

Contrasting the dN/dS and pN/pS ratios allows to test the selection regime acting on the sequences [34]. If mutations are all neutral or deleterious, we expect the ratios dN/dS and pN/pS to be equal. Positively selected mutations will tend to quickly rise to fixation and will not be observed as polymorphism, leading to an increased dN/dS ratio higher than pN/pS . Conversely, balancing selection will lead to an excess of polymorphism detectable as $dN/dS < pN/pS$ [35]. A simple measure of the proportion of amino-acid substitutions resulting from positive selection (α) is given by $1 - (dS \cdot pN/dN \cdot pS)$ [36]. Using the complete synonymous and non-synonymous site frequency spectra, it is further possible to estimate the distribution of fitness effects and account for slightly deleterious and slightly advantageous mutations when estimating the rate of adaptive substitutions (*see* Chapter 5) [37].

5 Linkage and Recombination

5.1 The Coalescent with Recombination

In sexually reproducing species, *recombination* refers to both the shuffling of non-homologous chromosomes and the rearrangement of homologous chromosomes during meiosis. Such cross-over events cause each chromosome to have two parent chromosomes in the previous generation, which are themselves the products of recombination events in the previous generations. Therefore, any chromosome in the current generation can be viewed as a mosaic of chromosomes that existed in the past (*see* Fig. 2) [38]. The collection of coalescence and recombination events that describes the history of sampled chromosomes until the most recent common ancestor of each non-recombining block is reached (*see* Fig. 2) is called the *ancestral recombination graph* (ARG) [39]. Compared to a tree-like genealogy of a sample without recombination, whose complexity depends only on the sample size, the complexity of the ARG grows with the sample size and the number of recombination events in the ancestry of the sample.

Backward-in-time, the *most recent common ancestor* (MRCA) denotes the first individual where the entire sample (population) coalesces for a particular non-recombining block. The *TMRCA* notes the timing of such event. DNA sequences provide no information beyond the MRCA in a sample of genomes since all individuals will share any mutation that happens further back in time [40]. In the presence of recombination, different parts of the genome will have different MRCA. In this case, all ancestral material is eventually found as a contiguous sequence in the *grand most*

Fig. 2 An ancestral recombination graph. An ancestral recombination graph is a collection of recombination (1–2) and coalescence (3–5) events. In each depicted chromosome, white bars represent segregating ancestral material, black bars represent coalesced ancestral material, and thin lines represent non-ancestral material. The asterisk denotes trapped non-ancestral material. Note that “1” does not impact the sample because the resulting segments are joined back together in “4” before coalescing in “5.” There are thus only two relevant TMRCA’s in the ARG, separated at position x

recent common ancestor (GMRCa) of the sample (*see* Fig. 2). If the GMRCa is not an MRCA for any nucleotide, this individual does not have any significance for DNA sequences [39].

In the ARG, nucleotide segments that are found both in past chromosomes and in contemporary samples are termed *ancestral genetic material* (*see* Fig. 2). Conversely, *non-ancestral genetic material* refers to segments that are found in past chromosomes but not in contemporary samples. Furthermore, non-ancestral genetic material flanked on both sides by ancestral genetic material is referred to as *trapped genetic material*. In this setting, recombination events that happen in trapped genetic material can affect linkage disequilibrium between present-day nucleotides (*see* Fig. 2). Thus the existence of trapped genetic material introduces long-range correlations between genealogies rendering the coalescent with recombination a non-Markovian process along chromosomes

[41]. The *Sequentially Markov coalescent (SMC)* is an approximation to the coalescent with recombination whereby recombination events are assumed to happen only within ancestral material. This approximation allows the use of efficient algorithms in both simulation and data analysis [42, 43].

5.2 Impact of Linkage on Selection

An excess of linkage between loci compared to a random association is termed *linkage disequilibrium (LD)*. LD arises from genetic drift, population admixture, and selection, but is reduced by recombination each generation. It is, therefore, higher between close loci and decays with increasing physical distance [44].

Linked selection refers to the reduction of diversity at neutral sites that happens as a result of their physical linkage to variants under selection [45]. In the absence of recombination, all variants segregating in a chromosome would undergo the same shift in frequency as the selected variant. However, recombination creates new allelic combinations and reduces this correlation as the physical distance from the selected locus increases (see Fig. 3).

Background selection refers to a form of linked selection where the reduction of diversity at neutral loci results from linkage to a locus under purifying selection [46], and *genetic hitchhiking* is commonly used to depict linked selection due to linkage to a locus under positive selection [47], where a new beneficial mutation will rise in frequency in a population. As the new positively selected allele increases its frequency, nearby linked alleles on the chromosome will “hitchhike” along with it, also growing in frequency, thus producing a *selective sweep* of genetic diversity (see Fig. 3d). *Hard sweeps* occur when a new mutation is positively selected and is therefore exclusively associated with the genetic background where it arose. Conversely, *soft sweeps* occur when a mutation is already segregating in the population at the onset of selection. This mutation may exist in several genetic backgrounds and therefore does not prompt a complete loss of genetic variation after the selective sweep [47] (see Fig. 3a–c).

Linkage of two or more loci can also impair the efficacy of positive selection, a phenomenon termed *Hill–Robertson interference (HRI)* [48]. When two advantageous mutations at distinct loci in distinct individuals segregate in the population, one will be lost unless a recombination event brings them together. In the absence of recombination between the selected loci, only the unlikely event of recurrent mutations can generate the optimal haplotypic combination [49] (see Fig. 3e).

A) Incomplete, then complete hard sweep

B) Incomplete, then complete soft sweep from standing genetic variation

C) Incomplete, then complete soft sweep from recurrent mutations

D) Background selection

E) Hill-Robertson interference

Fig. 3 Impact of selection on genetic diversity. Black lines represent individual genomes. SNP variants are displayed by filled circles. Distinct variants at the same position are depicted with different colors: neutral variants in gray, positive variants in red or yellow, and negative variant in blue. (a) A positively selected new variant spreads in the population and removes genetic diversity at linked loci, generating a hard selective sweep. (b and c) Segregation of several positively selected variants in different genetic backgrounds, either from standing variation or recurrent mutations, resulting in a soft selective sweep. (d) Reduction of neutral diversity because of linkage to deleterious mutations (background selection). (e) Competitive segregation of positively selected variant at distinct loci, resulting in the loss of advantageous variants (Hill–Robertson interference)

6 Notes

1. The use of the term *substitution* differs in population genetics and molecular biology. In the latter case, it describes a particular type of mutation where a single nucleotide replaces a distinct one (as opposed to insertions/deletions, for instance).
2. In some instances, s is substituted by $-s$, so that the relative fitnesses become $\omega_{AA} = 1$, $\omega_{Aa} = 1 - h \cdot s$ and $\omega_{aa} = 1 - s$.

References

1. Kimura M, Crow JF (1964) The number of alleles that can be maintained in a finite population. *Genetics* 49:725–738
2. 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
3. Crow JF (1989) Twenty-five years ago in genetics: the infinite allele model. *Genetics* 121(4):631–634
4. Elston RC, Satagopan J, Sun S (2017) Statistical genetic terminology. *Methods Mol Biol* 1666:1–9. https://doi.org/10.1007/978-1-4939-7274-6_1
5. Ewens WJ (1972) The sampling theory of selectively neutral alleles. *Theor Popul Biol* 3(1):87–112
6. Kimura M (1970) The length of time required for a selectively neutral mutant to reach fixation through random frequency drift in a finite population. *Genet Res* 15(1):131–133
7. Kimura M (1983) *The neutral theory of molecular evolution*. Cambridge University Press, Cambridge. <http://ebooks.cambridge.org/ref/id/CBO9780511623486>
8. Kingman JFC (1982) The coalescent. *Stoch Process Appl* 13(3):235–248. [https://doi.org/10.1016/0304-4149\(82\)90011-4](https://doi.org/10.1016/0304-4149(82)90011-4)
9. Barton NH (2016) Richard Hudson and Norman Kaplan on the coalescent process. *Genetics* 202(3):865–866. <https://doi.org/10.1534/genetics.116.187542>
10. Kingman JFC (2000) Origins of the Coalescent: 1974–1982. *Genetics* 156(4):1461–1463. <http://www.genetics.org/content/156/4/1461>
11. Sjödin P, Kaj I, Krone S, Lascoux M, Nordborg M (2005) On the meaning and existence of an effective population size. *Genetics* 169(2):1061–1070. <https://doi.org/10.1534/genetics.104.026799>
12. Wakeley J (2008) *Coalescent theory: an introduction*, 1st edn. Roberts and Company Publishers, Reading
13. Wright S (1938) Size of population and breeding structure in relation to evolution. *Science* 87:430–431
14. Jiang Y, Bolnick DI, Kirkpatrick M (2013) Assortative mating in animals. *Am Nat* 181(6):E125–138. <https://doi.org/10.1086/670160>
15. Sousa V, Hey J (2013) Understanding the origin of species with genome-scale data: modeling gene flow. *Nat Rev Genet* 14(6):404–414. <https://doi.org/10.1038/nrg3446>
16. Hey J, Nielsen R (2004) Multilocus methods for estimating population sizes, migration rates and divergence time, with applications to the divergence of *Drosophila pseudoobscura* and *D. persimilis*. *Genetics* 167(2):747–760. <https://doi.org/10.1534/genetics.103.024182>
17. Duthel JY, Hobolth A (2012) Ancestral population genomics. *Methods Mol Biol* 856:293–313. https://doi.org/10.1007/978-1-61779-585-5_12
18. Hobolth A, Christensen OF, Mailund T, Schierup MH (2007) Genomic relationships and speciation times of human, chimpanzee, and gorilla inferred from a coalescent hidden Markov model. *PLoS Genet* 3(2):e7. <https://doi.org/10.1371/journal.pgen.0030007>
19. Martin SH, Jiggins CD (2017) Interpreting the genomic landscape of introgression. *Curr Opin Genet Dev* 47:69–74. <https://doi.org/10.1016/j.gde.2017.08.007>
20. Mailund T, Munch K, Schierup MH (2014) Lineage sorting in Apes. *Annu Rev Genet*

- <https://doi.org/10.1146/annurev-genet-120213-092532>
21. Bustamante CD, Wakeley J, Sawyer S, Hartl DL (2001) Directional selection and the site-frequency spectrum. *Genetics* 159 (4):1779–1788
 22. Wright S (1968) Evolution and the genetics of populations, vol 2. The theory of gene frequencies. The University of Chicago Press, Chicago
 23. Schraiber JG, Akey JM (2015) Methods and models for unravelling human evolutionary history. *Nat Rev Genet* 16(12):727–740. <https://doi.org/10.1038/nrg4005>
 24. Kelleher J, Etheridge AM, McVean G (2016) Efficient coalescent simulation and genealogical analysis for large sample sizes. *PLoS Comput Biol* 12(5):e1004842. <https://doi.org/10.1371/journal.pcbi.1004842>
 25. Watterson GA (1975) On the number of segregating sites in genetical models without recombination. *Theor Popul Biol* 7(2):256–276
 26. Tajima F (1989) Statistical method for testing the neutral mutation hypothesis by DNA polymorphism. *Genetics* 123(3):585–595
 27. Nei M, Tajima F (1981) Genetic drift and estimation of effective population size. *Genetics* 98(3):625–640. <http://www.genetics.org/content/98/3/625>
 28. Tajima F (1983) Evolutionary relationship of DNA sequences in finite populations. *Genetics* 105(2):437–460
 29. Eyre-Walker A, Keightley PD (2007) The distribution of fitness effects of new mutations. *Nat Rev Genet* 8(8):610–618. <https://doi.org/10.1038/nrg2146>
 30. Orr HA (2009) Fitness and its role in evolutionary genetics. *Nat Rev Genet* 10(8):531–539. <https://doi.org/10.1038/nrg2603>
 31. Gillespie JH (2004) Population genetics: a concise guide. JHU Press, Baltimore
 32. Nielsen R (2005) Molecular signatures of natural selection. *Annu Rev Genet* 39:197–218. <https://doi.org/10.1146/annurev.genet.39.073003.112420>
 33. Pouyet F, Bailly-Bechet M, Mouchiroud D, Guéguen L (2016) SENCA: a multilayered codon model to study the origins and dynamics of codon usage. *Genome Biol Evol* 8 (8):2427–2441. <https://doi.org/10.1093/gbe/evw165>
 34. McDonald JH, Kreitman M (1991) Adaptive protein evolution at the *Adh* locus in *Drosophila*. *Nature* 351(6328):652–654. <https://doi.org/10.1038/351652a0>
 35. Parsch J, Zhang Z, Baines JF (2009) The influence of demography and weak selection on the McDonald-Kreitman test: an empirical study in *Drosophila*. *Mol Biol Evol* 26(3):691–698. <https://doi.org/10.1093/molbev/msn297>
 36. Smith NGC, Eyre-Walker A (2002) Adaptive protein evolution in *Drosophila*. *Nature* 415 (6875):1022–1024. <https://doi.org/10.1038/4151022a>
 37. Keightley PD, Eyre-Walker A (2007) Joint inference of the distribution of fitness effects of deleterious mutations and population demography based on nucleotide polymorphism frequencies. *Genetics* 177 (4):2251–2261. <https://doi.org/10.1534/genetics.107.080663>
 38. Stumpf MPH, McVean GAT (2003) Estimating recombination rates from population-genetic data. *Nat Rev Genet* 4(12):959–968. <https://doi.org/10.1038/nrg1227>
 39. Hein J, Schierup MH, Wiuf C (2005) Gene genealogies, variation and evolution: a primer in coalescent theory. Oxford University Press, Oxford
 40. Rosenberg NA, Nordborg M (2002) Genealogical trees, coalescent theory and the analysis of genetic polymorphisms. *Nat Rev Genet* 3 (5):380–390. <https://doi.org/10.1038/nrg795>
 41. Rasmussen MD, Hubisz MJ, Gronau I, Siepel A (2014) Genome-wide inference of ancestral recombination graphs. *PLoS Genet* 10(5): e1004342. <https://doi.org/10.1371/journal.pgen.1004342>
 42. McVean GAT, Cardin NJ (2005) Approximating the coalescent with recombination. *Philos Trans R Soc Lond B Biol Sci* 360 (1459):1387–1393. <https://doi.org/10.1098/rstb.2005.1673>
 43. Marjoram P, Wall JD (2006) Fast “coalescent” simulation. *BMC Genet* 7:16. <https://doi.org/10.1186/1471-2156-7-16>
 44. Slatkin M (2008) Linkage disequilibrium—understanding the evolutionary past and mapping the medical future. *Nat Rev Genet* 9 (6):477–485. <https://doi.org/10.1038/nrg2361>
 45. Cutter AD, Payseur BA (2013) Genomic signatures of selection at linked sites: unifying the disparity among species. *Nat Rev Genet* 14

- (4):262–274. <https://doi.org/10.1038/nrg3425>
46. Charlesworth B, Morgan MT, Charlesworth D (1993) The effect of deleterious mutations on neutral molecular variation. *Genetics* 134 (4):1289–1303
47. Maynard Smith J, Haigh J (1974) The hitchhiking effect of a favourable gene. *Genet Res* 23(1):23–35
48. Hill WG, Robertson A (1966) The effect of linkage on limits to artificial selection. *Genet Res* 8(3):269–294
49. Roze D, Barton NH (2006) The Hill-Robertson effect and the evolution of recombination. *Genetics* 173(3):1793–1811. <https://doi.org/10.1534/genetics.106.058586>

Open Access This chapter is licensed under the terms of the Creative Commons Attribution 4.0 International License (<http://creativecommons.org/licenses/by/4.0/>), which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license and indicate if changes were made.

The images or other third party material in this chapter are included in the chapter's Creative Commons license, unless indicated otherwise in a credit line to the material. If material is not included in the chapter's Creative Commons license and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder.