

Adaptations of neutrality tests

Inaugural-Dissertation

zur

Erlangung des Doktorgrades

der Mathematisch-Naturwissenschaftlichen Fakultät

der Universität zu Köln

vorgelegt von

Alexander Klassmann

aus Köln

Köln, 2018

Berichterstatter: Prof. Dr. Thomas Wiehe
Prof. Dr. Martin Lercher

Tag der mündlichen Prüfung: 20. Juni 2018

Abstract

Most of the genetic variation observed within a biological species is generally thought to be evolutionary “neutral” in the sense that it is irrelevant for an individual whether its genome contains one particular variant or another. Evolutionary biologists, and in the case of the human species anthropologists and medical scientists as well, are by contrast interested in variants which do influence on an individual’s survival and/or its ability to reproduce. Population geneticists try to find such variants by purely statistical methods in the form of *tests on neutrality* or shortly *neutrality tests*.

In this thesis four publications are reprinted and discussed which are concerned with modifications of existing neutrality tests. Three of them deal with a class of tests relying on the so-called *site frequency spectrum*. It was shown previously that some of these tests, originally designed on models of constant population size, can be adapted to allow for changes in population size. This is generalized in the first publication to all tests of similar structure. Another aspect of these tests is that they are ignorant with respect to which variant in a sample might evolve non-neutrally. If instead a particular variant is suspected a priori, the tests have to allow for this information by conditioning on the existence of a variant with the observed frequency. The second and third article introduce the concept of a *conditional frequency spectrum* and derive its first resp. second moments which are necessary for an appropriate extension of the above-mentioned class of tests. The fourth article presents an algorithmic improvement of a neutrality test of a different kind. Here, primarily computational speed was of concern, in order to bear comparison with competing software.

Solely applications on human data are presented, which is available in unrivalled abundance, owing to several large-scale genotyping and sequencing projects. The applicability of neutrality tests, however, is not confined to any particular species.

Zusammenfassung

Der größte Teil genetischer Variation, die man innerhalb einer biologischen Art findet, wird allgemein als "neutral" angesehen, in dem Sinn, dass es für ein Individuum unerheblich ist, ob sein Genom eine bestimmte Variante enthält oder eine andere. Evolutionsbiologen, und im Falle der menschlichen Spezies auch Anthropologen und Mediziner, sind dagegen gerade an den Varianten interessiert, die einen Einfluß auf das Überleben und/oder Fortpflanzungsfähigkeit eines Individuums haben. Populationsgenetiker versuchen solche Varianten mit rein statistischen Methoden zu finden und zwar in der Form von *Tests auf Neutralität*, oder kurz *Neutralitätstests*.

In dieser Arbeit werden vier Artikel wiedergegeben und diskutiert, die sich alle mit Anpassungen bereits bestehender Neutralitätstests befassen. Drei davon handeln von einer Klasse von Tests, die auf dem sogenannten *Frequenzspektrum* von Varianten beruhen. Bereits zuvor war gezeigt worden, dass einige von diesen Tests, ursprünglich entwickelt anhand von Modellen mit konstanter Populationsgröße, so angepasst werden können, dass sie Änderungen in der Populationsgröße berücksichtigen. Dies wird im ersten Artikel verallgemeinert auf alle Tests mit ähnlicher Struktur. Ein anderer Aspekt dieser Tests ist, dass sie keine Vorannahmen machen, welche Variante in einer Stichprobe nicht neutral sein könnte. Wird dies jedoch von einer bestimmten Variante vermutet, müssen die Tests diese Information berücksichtigen, indem sie die Existenz einer Variante mit der beobachteten Frequenz als Bedingung enthalten. Der zweite und dritte Artikel führen das Konzept eines *bedingten Frequenzspektrums* ein und leiten seine ersten und zweiten Momente ab, die für eine geeignete Erweiterung der oben genannten Klasse von Tests benötigt werden. Der vierte Artikel präsentiert eine algorithmische Verbesserung eines anders gearteten Neutralitätstests. Sie diene hauptsächlich einer Erhöhung der Rechengeschwindigkeit um mit konkurrierenden Programmen Schritt zu halten.

Es werden ausschließlich Anwendungen auf humangenetische Daten vorgestellt, wo die Datenlage, auf Grund mehrerer großer Genotypisierungs- und Sequenzierungsprojekte, am besten ist. Die Anwendbarkeit von Neutralitätstests ist jedoch nicht auf irgendeine bestimmte biologische Art beschränkt.

Contents

1	Introduction	1
1.1	Motivation	1
1.2	Population genetic models	1
1.2.1	Genetic forces	1
1.2.2	Transmission models	2
1.2.3	Molecular signatures of selection	3
1.3	The site frequency spectrum	4
1.4	Tests on neutrality	6
1.4.1	Tests based on variant frequency differences between subpopulations	6
1.4.2	Tests based on the site frequency spectrum	6
1.4.3	Tests based on haplotypes	11
1.4.4	A combined test	12
1.5	Selection scans in humans	13
1.5.1	An early scan using F_{ST}	13
1.5.2	Scans for positive selection	13
1.5.3	Scans for balancing selection	15
1.6	Coalescent mathematics	17
1.7	Adaptations of tests on neutrality	20
2	Publications	23
2.1	Demography-adjusted tests of neutrality based on genome-wide SNP data	23
2.2	The neutral frequency spectrum of linked sites	36
2.3	The third moments of the site frequency spectrum	46
2.4	REHH 2.0: a reimplementaion of the R package REHH to detect positive selection from haplotype structure	59
3	Discussion	73
3.1	Demography-adjusted tests of neutrality	73
3.2	The neutral frequency spectrum of linked sites	74
3.3	The third moments of the site frequency spectrum	77
3.4	A reimplementaion of the R package REHH	78
3.5	Conclusions	79
3.5.1	Whole genome scans help to assess test values at individual loci	80
3.5.2	Recent completed selective sweeps are rare in humans	80
3.5.3	How neutral is the human genome?	83
4	References	87
5	Eigene Beteiligung an den Publikationen	95
6	Erklärung	97

1 Introduction

1.1 Motivation

Telling apart the essential from the accidental is involved in many a human task. In molecular biology, the ease with which genomes can be sequenced mismatches the difficulty in understanding the meaning contained in them. The sheer amount of genetic information precludes any piecemeal and exhaustive investigation by experiment. Theoretical evolutionary biologists try to interpret comparisons of multiple sequences with the aim to prioritize genomic regions which may warrant a closer examination in the laboratory. One approach is to search for *evolutionary conserved* sequences from diverse branches of the tree of life. These are likely to be of fundamental biological importance for each organism. On the other end of the evolutionary scale, genetic variation among individuals of a single species is the subject of inquiry - the realm of *Population Genetics*. Variation within a species is often classified into short variants such as single nucleotide polymorphisms (SNPs), short tandem repeats ("micro-satellites"), insertions and deletions of a few nucleotides and *structural variation* such as deletions, single or multiple duplications, translocations and inversions of larger chromosomal segments. Nowadays SNPs are by far the most preferred type of variation to perform population genetic inferences on. They can be ascertained in large numbers and represent the biggest share of independently occurred mutations, an important issue for the application of statistical tests. Conspicuous values of such tests form the basis for delineating *candidate regions* of some particular biological interest.

1.2 Population genetic models

1.2.1 Genetic forces

Population geneticists try to understand the evolution of a single species by investigation of its genetic composition which is thought to result mainly from the evolutionary forces *mutation, genetic drift, selection, recombination, population splits* and *migration*.

The simplest model of variation consists of a single hypothetical *locus* in a genome where two different variants are observed in a population. This is sufficient to classify selection into different *modes*:

- a variant is referred to be under *positive* or *Darwinian* selection, if its carrier individuals have a consistent advantage in viability and/or fecundity and leave on average more offspring than other individuals.
- a variant is referred to be under *negative* or *purifying* selection, if its carrier individuals have a consistent disadvantage and leave on average less offspring than other individuals.

Both kinds of selection are *directional* since the population frequency of the selected variant tends to steadily increase resp. decrease until the variant is fixed in the population or lost, respectively. By contrast,

- *balancing* selection refers to any mechanism promoting the co-existence of two (or more) variants.

Two variants that are not under selection are referred to as *evolutionary neutral*. The time course of their population frequencies follows purely random fluctuations or *genetic drift* which, again, in any finite

1 Introduction

population will eventually lead to fixation of one variant and loss of the other.

While drift and selection eliminate or at best maintain variation, the origin of variants lies always in mutation. This force can be modelled as a random process with a constant rate μ over time and along the genome. The observed mutation rate is mostly low, in case of single nucleotide mutations in humans 10^{-8} per nucleotide per generation [The 1000 Genomes Project Consortium, 2012] and can be well described by the *infinite sites model* formulated by Kimura and Ohta [1969]. The model posits that every mutation happens at a hitherto un-mutated position in the genome, and as a consequence, at each position at most two variants co-exist in a population, one ancestral and the other derived [Tajima, 1996].

Sexually reproducing diploid organisms turn over to their offspring only half of their genetic material. For each parent it is random which of two homologous chromosomes is transmitted to the child, entailing the Mendelian law of independent assortment. Variants on the same chromosome would be inherited always together were it not for the phenomenon of recombination which exchanges pieces between homologous chromosomes. In humans there is about one recombination per chromosome per generation [Dumont and Payseur, 2008] which, per base pair, is similar to the mutation rate. Despite recombination events occurring very inhomogeneously along a chromosome [McVean et al., 2004; The 1000 Genomes Project Consortium, 2012], for lack of better knowledge and the sake of simplicity, in population genetic models its rate is usually assumed to be constant in time and space, or, for short regions, neglected altogether.

A population split can arise by geographical isolation. Variant frequencies in separated subpopulations diverge by genetic drift, supplemented possibly by selection caused by different environments, ushering in *population structure*. Subsequent migration may lead in the short run to genetically inhomogeneous or *admixed* subpopulations.

1.2.2 Transmission models

In order to describe changes of variant frequencies in a population, a transmission model for variants is necessary. The WRIGHT-FISHER model is one of them and supposes in its simplest form a population of N individuals which reproduce simultaneously in non-overlapping generations. Variant frequency changes arise out of different reproductive success of individuals, allowing for both random events as well as selection due to their genetic variants. N most often cannot be equated with the census number of individuals in a population at a certain time point. Instead, it represents a synthetic number, the *effective population size*, influenced by various factors such as mating behaviours and past population size changes. Consequently it cannot be measured directly, but has to be inferred. The effective population size of the human species has been estimated to be of order 10^4 [Tenesa et al., 2007].

The product $N \cdot \mu$ appears in many population genetic equations dealing with neutral evolution and for a diploid species it is abbreviated by $\theta = 4N\mu$. Formally only a parameter, θ should be interpreted as a measure for population variation as motivated by its various estimators presented in section 1.4.2.

For a large population and time scales of many generations, the WRIGHT-FISHER model can be approximated by continuous partial differential equations, called *diffusion approximation* in analogy to the eponymous process in physics [Kimura, 1964]. The WRIGHT-FISHER model and its approximation has been used to derive quantitative properties of variation. Important results include the time variants circulate in the population until they become fixed or lost. This time can differ enormously between neutral variants and those under selection: if a variant gets fixed by chance alone, it needs on average on the order of N generations [Kimura and Ohta, 1969], while a variant under (strong) positive selection needs on average only a time proportional to the logarithm of the population size and inversely proportional to its selective advantage [Ewens, 2004, section 5.4].

The WRIGHT-FISHER model looks “forward-in time” and projects a given state into the future. Kingman [1982] invented another model, the *coalescent tree*. This model looks “backwards-in-time”, from the present into the past. Although it is particularly well apt only for modeling neutral evolution, it rapidly gained popularity over the WRIGHT-FISHER model. Its success is intimately related to the development of DNA sequencing technologies. On the one hand, the obtained sequences convinced most researchers of the *neutral theory* which states that a large amount of the observed molecular variation evolves neutrally. On the other hand, the availability of sequence data made it possible to infer population characteristics, hence the necessity to develop appropriate statistical models. Coalescent trees are well suited for this task. They are bifurcating trees with a sample of present day sequences at the leaves and their most recent common ancestor at the root. As such, they resemble phylogenetic trees, however they are not intended to describe particular relationships among specific individuals, but as a mean to calculate statistical averages serving as background for inferences about the entire population [Wakeley, 2008].

1.2.3 Molecular signatures of selection

The strategy to detect the supposedly few variants experiencing selection among the much more numerous neutral variants is to search for distinctive patterns of variation caused by selection, its *molecular signature*.

As stated above, the population frequency of a variant under positive selection can change much faster than that of neutral variants. However, neutral mutations in the vicinity of a selected one will share its fate unless a recombination event separates them onto different chromosomes - they “hitch-hike”. Sometimes the metaphor is extended to refer to variants as *driver* and *passenger*, respectively [Bozic et al., 2010]. If a notable amount of neutral mutations reaches fixation together with the selected variant, the reservoir of variants is depleted by a *selective sweep* (Figure 1.1). The founders of the mathematical formulation of this scenario, Maynard Smith and Haigh [1974], expounded, that the ratio of selective strength to recombination rate determines the strength of the depletion. Others stressed the importance of their absolute values [Przeworski, 2002]. In any case, a low number of variants, all of low population frequency, possibly surrounded by high frequency variants that “escaped” from fixation by recombination, are the hallmarks of a completed selective sweep’s aftermath.

Neutral variants may also hitch-hike with negatively selected variants and hence are driven to extinction. A characteristic pattern of this *background selection* is that the number of variants as well as their population frequency are reduced in comparison with purely neutral evolution [Charlesworth et al., 1993]. Background selection thus partially confounds the signature of a selective sweep, causing a long-standing controversy about their respective share in human and other genomes [Stephan, 2010; Hernandez et al., 2011].

Balancing selection is supposed to lead to a distinctive molecular signature only if it operates on evolutionary long times, while variants that not yet or only recently reached their equilibrium frequency yield patterns similar to a partial selective sweep [Charlesworth, 2006]. Neutral variants in the vicinity of a variant under balancing selection are hindered in their way to fixation and hence are “trapped” and accumulate. There is no mathematical model of similar fame as the above-mentioned for selective sweeps, but the model of Hudson and Kaplan [1988] may worth mentioning, since it is the only one so far explicitly referred to in a genomic scan for balancing selection [DeGiorgio et al., 2014]. It presupposes a balancing selection between two variants strong enough to yield a permanently fixed equilibrium frequency. Recombination prevents an infinite accumulation of neutral variants by “migrating” them from chromosomes belonging to one selected variant to chromosomes of the other. Irrespective of this specific model, long-term balancing selection is expected to yield in the vicinity of the selected variants a surplus of neutral variants with similar population frequency. The detectable region, though, may be quite narrow, since evolutionary long times provide much opportunity for recombination to erode the pattern [Charlesworth,

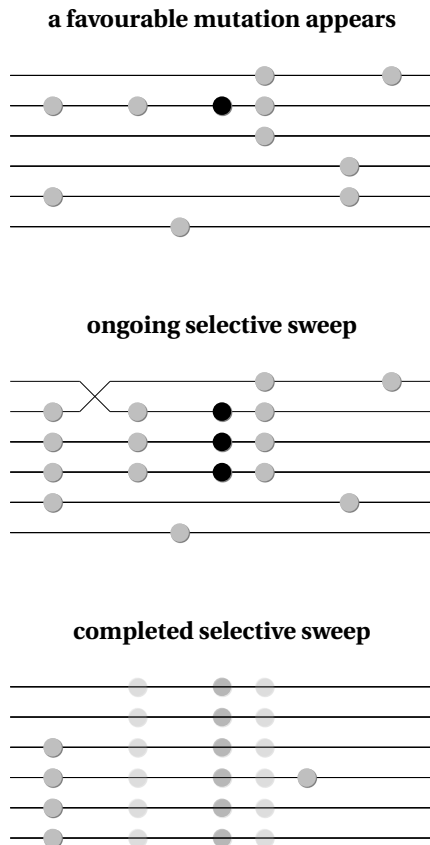


Figure 1.1: Schema of a selective sweep. The lines symbolize sequences, the circles mutations, synonymous with derived variants. The upper panel shows a region with neutral mutations (grey) in equilibrium, together with a newly arising favourable mutation (black). Thanks to this mutation the second sequence quickly replaces other sequences and neutral mutations on it “hitch-hike” to higher frequencies. At a certain time (middle panel) the first two sequences recombine between the leftmost mutation and the remainder, swapping that mutation from the second to the first sequence. In the third panel, the sweep is completed, i.e. the favourable mutation and two neutral ones have become fixed in the population, while most other variants have been “swept” away. Only two variable sites remain: the high-frequency leftmost mutation and a low-frequency mutation, that appeared during or shortly after the sweep.

2006].

Molecular signatures of selection have the caveat that non-selective forces like demography and substructure can produce similar patterns of variation. For example, a growing population will mimic a selective sweep, since the latter can be imagined as a growing subpopulation containing the favoured variant. Similarly, a subpopulation with a few migrants will have a surplus on variants with low frequency, too. A strongly admixed population, on the contrary, may exhibit patterns of molecular variation similar to balancing selection. In principle, selection acts on particular variants and hence leaves its footprints only on certain genomic regions, while demography and population substructure are expected to affect the whole genome. However, they may not do so uniformly, increasing the amount of neutral “noise”, thus bedeviling the detection of selection signals. A particularly worrisome demographic scenario is a *population bottleneck*, a sudden reduction of population size, followed by a rapid expansion, which may cause many spurious signatures of selective sweeps [Jensen et al., 2005]. It is generally agreed that all non-African human populations experienced a bottleneck during their migration out of Africa [Marth et al., 2004; Stajich and Hahn, 2005; Liu et al., 2006; Gutenkunst et al., 2009].

1.3 The site frequency spectrum

The complete pattern of variation which appears in aligned sequences is hard to interpret. Hence, it is helpful to extract the relevant information into appropriate *summary statistics*. An easy-to-calculate and nevertheless very versatile summary statistic is the *site frequency spectrum*. It can be used to estimate mutation [Liu et al., 2009] or recombination rates [Lachance and Tishkoff, 2014] and past population size changes [Adams and Hudson, 2004; Liu and Fu, 2015; Lapierre et al., 2017], but it is its property as

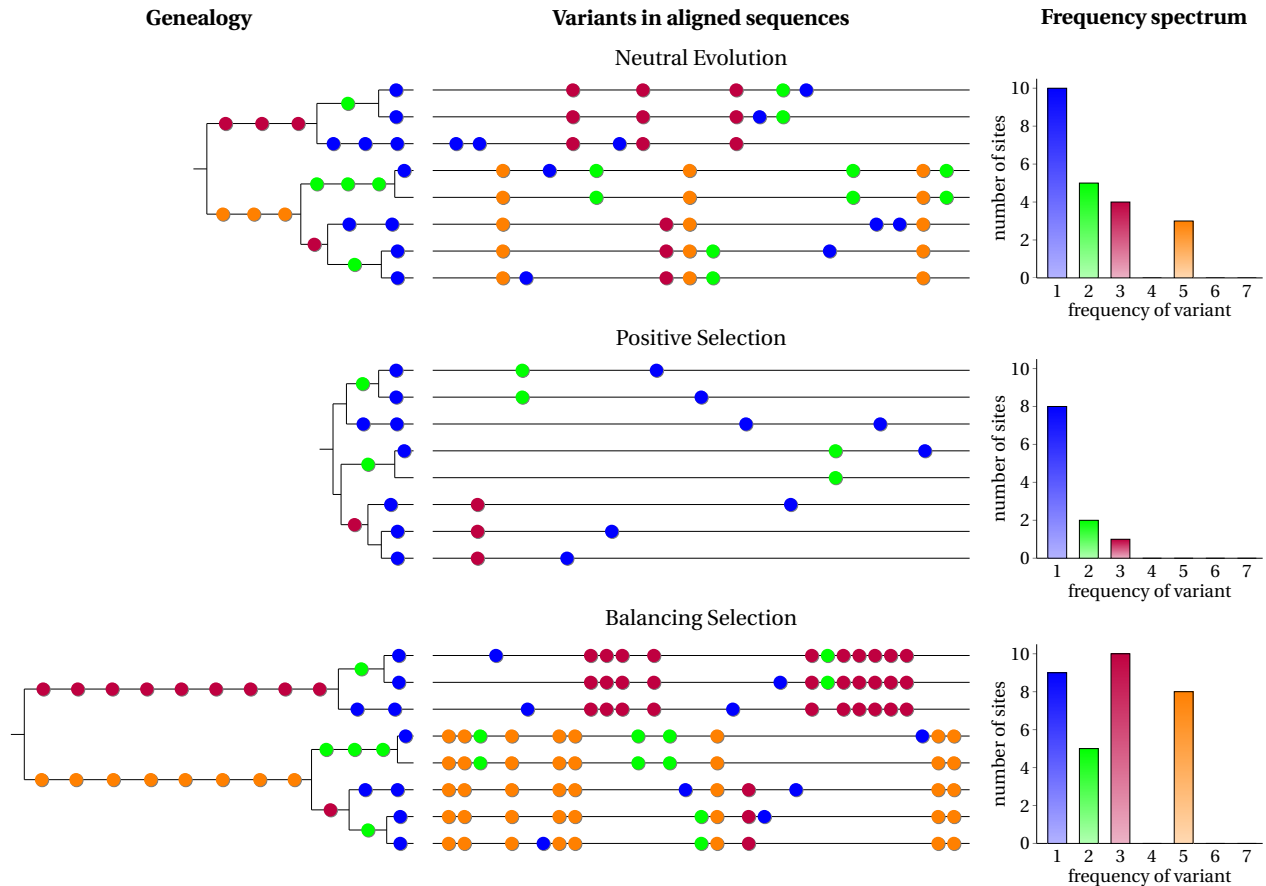


Figure 1.2: Evolutionary scenarios for a sample of eight sequences in a non-recombining genomic region. Top panels: neutral evolution, in an equilibrium between mutation and genetic drift. Middle panels: positive selection, the situation after a completed selective sweep. Bottom panels: balancing selection, the long-time co-existence of two variants. Left: the genealogy of the region. Centre: the pattern of variation in the aligned sequences of the region. Right: the frequency spectrum of derived variants. The colours represent the frequency of the mutation in the sample. Modified after [Bamshad and Wooding, 2003].

an indicator for natural selection that is of most importance within this thesis. Figure 1.2 illustrates the underlying concept. In the middle panels, hypothetical variation in aligned sequences from a short non-recombining genomic region is depicted. These patterns are the result of the genealogical trees on the left side, representing different evolutionary scenarios. In practice, sequence variation is known and genealogies are not. Although attempts have been made to infer the true genealogies from sequence data [Rasmussen et al., 2014], the methods are computationally demanding and the precision of the results hard to ascertain. Instead, most wanted is a “simple” assignment of the observed variation to one of the evolutionary scenarios. This can be achieved to some extent with the help of the frequency spectrum depicted on the right. The absolute frequency of a variant in a sample is called its *size*, identified in the picture by its colour. Counting all same-size variants yields the frequency spectrum. Since its expected form under neutrality is surprisingly simple (Figure 1.3), deviations from it can be quantified and associated with different evolutionary scenarios. Positive selection is associated with an over-representation of low frequency variants and balancing selection is thought to yield a preponderance of middle frequency variants. If ancestral and derived variant cannot or need not be distinguished, the *folded* frequency spectrum can be formed by considering at each site the variant with minor frequency.

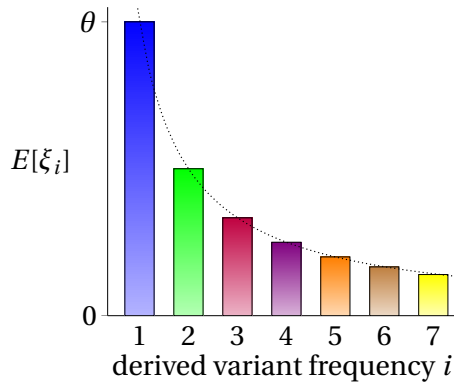


Figure 1.3: The expected frequency spectrum under neutrality and constant population size is given by $E\{\xi_i\} = \frac{\theta}{i}$. The dashed line shows its continuous extension $\frac{\theta}{x}$.

1.4 Tests on neutrality

Throughout the last two decades a steady stream of newly proposed neutrality tests has come to the fore, almost all aimed at detecting positive selection in form of selective sweeps, reviewed by Nielsen [2001, 2005]; Thornton et al. [2007]; Oleksyk et al. [2010]; Vitti et al. [2013]; Booker et al. [2017]; Pavlidis and Alachiotis [2017]. Not all, however, reached a high profile beyond population genetic specialists. For instance, a guideline for biologists [Cadzow et al., 2014] recommends the usage of F_{ST} , TAJIMA's D , FAY&WU's H , iHS and rsb . An overview over the field is given below.

1.4.1 Tests based on variant frequency differences between subpopulations

Subpopulations that became separated geographically or otherwise will diverge by genetic drift and selection. Variants with extremely high or low difference in population frequency can be suspected to be under regional directional selection or global balancing selection, respectively. Among various, often similar measures for population subdivision, F_{ST} is most widely used. It compares variation within subpopulations with that of a hypothetical non-divided total population. It can be calculated for every site with two variants (with extensions to include multiple variants) and yields 0 if the variant frequencies are the same in subpopulations and 1 if different variants are fixed in different subpopulations (reviewed by Holsinger and Weir [2009]). F_{ST} can be used as a test statistic if simulations of neutrally evolving sequences provide critical values for significance.

The related quantities p_{excess} [The international HapMap Consortium, 2005, Supplementary information] and PBS [Yi et al., 2010] measure the difference of F_{ST} in one population to one or several other populations serving as putatively neutral references. The 1000 Genomes Project Consortium [2012] used the difference in the frequency of derived variants ΔDAF as measure for population differentiation.

1.4.2 Tests based on the site frequency spectrum

Since the frequency spectrum is a central topic of this thesis, the development of associated tests will be described in detail.

The MAN-WHITNEY U test seems to be the only test from standard statistic theory that has been applied to frequency spectra [Akashi, 1999; Andrés et al., 2009; DeGiorgio et al., 2014]. It is a non-parametric test to handle ordinal scaled data, in this case the size of mutations. For instance, the frequency spectrum for positive selection in Figure 1.2 is represented by the set of numbers $\{1, 1, 1, 1, 1, 1, 1, 1, 2, 2, 3\}$ which can be compared by the test with any other observed spectrum to indicate general trends such as that one sample contains more lower frequency variants than another. If applied to the folded spectrum, it can

detect an over- or under-representation of middle-sized frequency variants.

Quite a few tests rely on the comparison of estimators for the parameter θ as defined in section 1.2.2. Watterson [1975] found an easy-to-calculate estimator for θ by counting the number of segregating sites S within a sample and correcting for the sample size n by a harmonic number $H_n = \sum_{i=1}^n \frac{1}{i}$,

$$\hat{\theta}_S = \frac{S}{H_{n-1}}. \quad (1.1)$$

This estimator has come to be known as WATTERSON's estimator and is often symbolized by $\hat{\theta}_W$, too. He also derived the variance of this estimator. Tajima [1983] computed the variance of another estimator of θ , namely the average pairwise difference between two sequences, referred to as Π . More formally, it is defined as the number of all differences k_{ij} between pairs of sequences i and j of a sample, divided by the possible number of pairings:

$$\hat{\theta}_\Pi = \Pi = \frac{2}{n(n-1)} \sum_{i \neq j}^n k_{ij}. \quad (1.2)$$

Building on this, Tajima [1989] showed, that, although both estimators are identical for samples of size $n = 3$, their correlation decreases with increasing n . This led him to propose a test which became known as TAJIMA's D :

$$D = \frac{\hat{\theta}_\Pi - \hat{\theta}_S}{\sqrt{\text{Var}[\hat{\theta}_\Pi - \hat{\theta}_S]}}. \quad (1.3)$$

Since under neutral evolution both estimators have the same expected value, the test statistic for a sample drawn from a neutrally evolving population should yield a value near zero. If instead, for a given number of segregating sites, low-frequency variants are over-represented, the value of Π is diminished and the test statistic negative which can be taken as a signal for positive selection. Conversely, an over-representation of middle-frequency variants causes a positive value of the test statistic, possibly signalling balancing selection. For example, the values of TAJIMA's D for the scenarios in Figure 1.2 are -0.06, -1.00 and 0.73 for neutral evolution, positive selection and balancing selection, respectively. In short, specific deviations of the observed frequency spectrum from that expected under neutrality can be summarized by a single number and, as discussed below, assigned a significance level.

Fu and Li [1993] portioned coalescent trees into external branches which lead to leaves and internal branches which do not. Mutations on external branches are seen by definition only on a single sequence and are called *singletons*. It turned out that the number of singletons is an estimator for θ , too:

$$\hat{\theta}_e = \xi_1. \quad (1.4)$$

The authors proposed two new tests

$$D = \frac{\hat{\theta}_S - \hat{\theta}_e}{\sqrt{\text{Var}[\hat{\theta}_S - \hat{\theta}_e]}} \quad (1.5)$$

and

$$F = \frac{\hat{\theta}_\Pi - \hat{\theta}_e}{\sqrt{\text{Var}[\hat{\theta}_\Pi - \hat{\theta}_e]}}. \quad (1.6)$$

While WATTERSON's and TAJIMA's estimators are "symmetric" with respect to ancestral and derived variants, the last two tests are not and hence for them the variants have to be *polarized* (see eponymous box).

1 Introduction

Fu and Li [1993] proposed modifications D^* and F^* of these tests using the folded frequency spectrum and hence suitable for un-polarized variants. Since the modifications are in fact small, the tests are likely to have similar power, although formal comparisons are seemingly not published. Instead, a comparison between TAJIMA's D and FU&LI's D^* and F^* has been undertaken by Braverman et al. [1995] and Simonsen et al. [1995]. Both studies investigated the power of the tests to reject neutrality using extensive simulations of different non-neutral scenarios. The former study tested against an alternative scenario of a population bottleneck, a single hitch-hiking event and a population split. The latter study tested against neutral mutations affected by recurrent selective sweeps in their vicinity. Both found, that TAJIMA's D performs always better than the other two tests. A follow-up simulation study by Fu [1997] suggested that FU&LI's D^* and F^* are more powerful to detect background selection.

Fay and Wu [2000] proposed yet another estimator and an associated test, aiming to improve on previous tests particularly in the detection of selective sweeps:

$$\hat{\theta}_H = \frac{2}{n(n-1)} \sum_{i=1}^{n-1} i^2 \xi_i \quad (1.7)$$

is used to construct the test

$$H = \frac{\hat{\theta}_\Pi - \hat{\theta}_H}{\sqrt{\text{Var}[\hat{\theta}_\Pi - \hat{\theta}_H]}}. \quad (1.8)$$

While Tajima's estimator can be expressed as a weighted average of heterozygosity, $\hat{\theta}_\Pi = \sum_{i=1}^{n-1} 2\xi_i \frac{i(n-i)}{n(n-1)}$, the estimator $\hat{\theta}_H$ measures the average homozygosity for the derived variant. It puts hence much weight on high frequency derived variants, which should be rare in a neutrally evolving genomic region. The authors argue that only selective sweeps lead to an over-representation of such variants and consequently confounding scenarios such as bottlenecks or background selection can be ruled out. However, this specificity comes with the price of relying heavily on the correct polarization of the variants. They acknowledge this by suggesting that the estimated proportion of *ancestral variant misidentification* (see associated box) should be allowed for in the calculation of significance levels.

A simulation study [Przeworski, 2002] scrutinized the properties of FAY&WU's H and found, that this test is only powerful to detect selective sweeps in a narrow time frame around the fixation of the advantageous variant. Furthermore, it is vulnerable to population structure in the sense that a few migrants from another population may cause fixed derived variants of the main population to appear segregating at high frequency, confounding the signature of a nearly-completed selective sweep. [Zeng et al., 2006, co-authors Fu and Wu] tried to address these concerns by constructing another two tests that both are mixtures of the preceding ones. The first one uses a new estimator

$$\hat{\theta}_L = \frac{1}{2}(\hat{\theta}_H + \hat{\theta}_\Pi), \quad (1.9)$$

to yield the test statistic

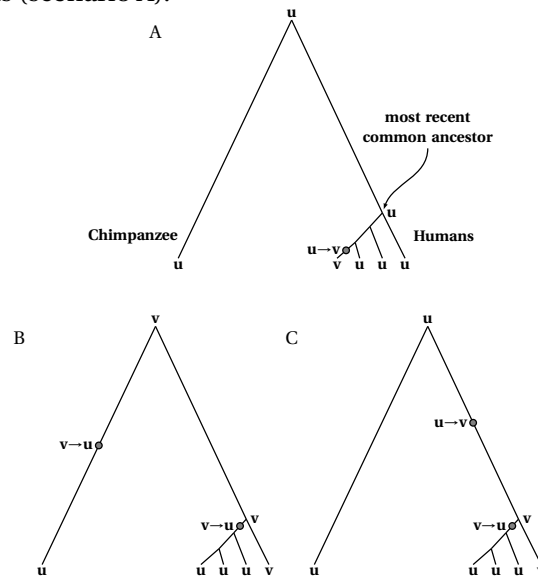
$$E = \frac{\hat{\theta}_L - \hat{\theta}_W}{\sqrt{\text{Var}[\hat{\theta}_L - \hat{\theta}_W]}} \quad (1.10)$$

while the second one, called DH , consists in a joint application of the D and H -tests. Both new tests are suggested to be less sensitive to population size changes and population structure. The authors add a ranking of the tests D , H , E and DG with respect to their sensitivity to various evolutionary scenarios.

Finally [Achaz, 2008] investigated the influence of sequence errors on the above-mentioned tests. He argued that such errors mostly lead to spurious singletons and suggested to use, in a case of doubt, an amended version of TAJIMA's D , named Y , which excludes singletons from the calculation.

Polarization of variants

Suppose that in a sample of human sequences at a certain genomic position two variants **u** and **v** are observed. To *polarize* them means to establish the direction of the mutation, or equivalently, to infer which of the two was carried by the most recent common ancestor of modern humans. If a “sister species” like Chimpanzees carries one of the variants, it is common practise to take that one as the ancestral variant of humans (scenario A).



The infinite sites model which allows only one mutation at a given site is appropriate for time scales within a species. However for the long time separating humans and chimpanzees, a second mutation has a non-negligible probability to occur on the same position. In this case the ancestral variant of humans is misspecified by the chimpanzee variant (scenarios B and C). How often does this happen? A first approximation is $\frac{d}{3}$, where d stands for the average fraction of different bases between the species and $\frac{1}{3}$ is the probability that the second mutation is a reversal of the first if we assume that all mutations are equally likely. If instead, as is well known, transitions (mutations of type $A \leftrightarrow G$ or $C \leftrightarrow T$) happen twice as often as transversions ($A \leftrightarrow C$, $A \leftrightarrow T$, $C \leftrightarrow T$ or $G \leftrightarrow T$), then the probability of a mutation $u \leftrightarrow v$ within humans being a transition is the same as being a transversion, namely $\frac{1}{2}$ and the probability of a transition or transversion occurring on the long branches is $\frac{d}{2}$ each. A second transition would restore necessarily the original variant while a second transversion would do so only in half of cases. Hence in total there is a higher probability of $\frac{1}{2} \left(\frac{d}{2} + \frac{d}{2} \cdot \frac{1}{2} \right) = \frac{3d}{8}$ for a misspecification [Fay and Wu, 2000]. Hernandez et al. [2007] extended this argument to all 12 possible mutations and allowing for context-dependence on the preceding and succeeding site. The divergence d between humans and chimpanzees is commonly stated as 1% [The Chimpanzee Sequencing and Analysis Consortium, 2005], although this number describes only base substitutions and neglects insertions, deletions and structural mutations [Cohen, 2007]. The divergence and thus the problem of mis-identification is greater in other well studied sister species like *Drosophila melanogaster* / *simulans* ($d \approx 4.1\%$) Garrigan et al. [2012] or *Arabidopsis thaliana* / *lyrata* ($d \approx 15\%$) [Hu et al., 2011]. On the other hand, more closely related species are more likely to share ancestral or *trans-species* polymorphisms, again impeding their correct polarization [Baudry and Depaulis, 2003]. Wiuf et al. [2004] calculated an approximation for this probability under neutral evolution, which in case of humans/chimps yields, rather as upper limit, $9e^{-8} \approx 0.003$, still lower than the probability of a double mutation. For all known cases of human trans-species polymorphisms, such as the binding region of MHC molecules [Klein et al., 1998], the *ABO* blood group gene [Thompson et al., 2013] and a few other genes of the immune system [Těšický and Vinkler, 2015], balancing selection is invoked.

1 Introduction

The formal similarity of the tests led to an early attempt of generalization by Fu [1996]. Fu [1995] had shown that for every $i = 1, \dots, n-1$ the product $i\xi_i$ can be taken as an estimator $\hat{\theta}_i$ for θ . He observed that $\hat{\theta}_S, \hat{\theta}_\Pi, \hat{\theta}_e$ and in fact the later defined $\hat{\theta}_H$ can be described essentially as members of a one-parameter family of estimators given by

$$\hat{\theta}(r) = \frac{1}{\sum_{i=1}^{n-1} \left(\frac{1}{i}\right)^r} \sum_{i=1}^{n-1} \left(\frac{1}{i}\right)^r i \xi_i \quad (1.11)$$

and hence the corresponding tests can be identified by two parameters r and r' . Achaz [2009] found another, more flexible, characterization of the tests. He parametrized not the estimators, but their differences. Let $\Omega_i, i, \dots, n-1$ be numbers with the condition that $\sum_{i=1}^{n-1} \Omega_i = 0$ and write $\Omega = (\Omega_1, \dots, \Omega_{n-1})$ resp. $\hat{\Theta} = (\hat{\theta}_1, \dots, \hat{\theta}_{n-1})$ as vectors. Then every of the above-mentioned tests is identified by its *weighting scheme* Ω and can be written in the form

$$T_\Omega = \frac{\hat{\Theta} \cdot \Omega}{\sqrt{\text{Var}[\hat{\Theta} \cdot \Omega]}}. \quad (1.12)$$

Using this notation the variances in the nominator, once painstakingly calculated for each test independently, can be subsumed into a single formula [Achaz, 2009, Eq. (9)].

The existence of a family of tests defined by weights provoked the question: which of its members can best discriminate neutral evolution against a given alternative scenario? The answer was given partially by Ferretti et al. [2010b]. They formulated a consistency requirement that a weighting scheme for any test should scale with the sample size in the sense that the same relative parts of the frequency spectrum are contrasted by the test. They showed that although TAJIMA'S D does not fulfil strictly this criterion, its dependency on sample size is not strong, while the tests of FU&LI do not scale well, since they oppose for all sample sizes the class of singletons to the remaining $n-2$ classes. More importantly, Ferretti et al. [2010b] showed that, taking the maximization of the expectation value as a criterion for optimality, it is easy to generate an "optimal" test distinguishing two given scenarios: the weights Ω must be chosen as the expected differences of the two corresponding frequency spectra.

All tests have a practical problem in common, namely the establishment of critical values for significance, because the distributions of the test statistics are not known analytically. Although "the normalization [by the nominator] is intended to standardize the variance of the test statistic and hopefully bring the statistic close to the standard normal distribution" [Fu and Li, 1993], this is not accurate enough, because the variance itself depends on an estimation of θ . [Tajima, 1989] approximated the distribution of his test statistic by a beta-distribution, well known by statisticians, yet this was founded on merely accidental visual similarity. The correct assignment of critical values is cumbersome since it involves allowing for variance of the theta estimator in the denominator. Detailed lists of critical values have been computed by Simonsen et al. [1995] and Fu [1996], however, in practise these values are obtained by approximate coalescent simulations, based on the estimated value of θ alone, conceding some imprecision arising from the neglect of its variance.

Tests relying on the frequency spectrum, but not subscribing to the above mentioned framework, have been proposed, too. In particular [Fu, 1996] discusses a test statistic of the form

$$T = \sum_{i=1}^{n-1} \sum_{j=1}^{n-1} \left(\xi_i - \frac{1}{i} \right) (\text{Cov}[\xi_i, \xi_j]^{-1})_{ij} \left(\xi_j - \frac{1}{j} \right), \quad (1.13)$$

suggested to follow a χ^2 -distribution and being sensitive to selective sweeps. However, he admitted that the effect of recombination on this test remains unclear, while it merely reduces the power of the above-mentioned tests [Wall, 1999]. A mathematical treatment on similar tests has been provided by Ferretti et al. [2010a]. Furthermore, machine learning [Ronen et al., 2013] and Bayesian [Eldon et al., 2015] methods to

discriminate between frequency spectra have been proposed. So far, none of these approaches found much application.

By contrast, the comparison of frequency spectra by likelihoods turned out to be fruitful. Assume a null- and an alternative hypothesis that a variant has a probability of p_i^0 resp. p_i^A for being of size i . A test using variants in a specific genomic region and having the form

$$T = \frac{\prod_{i=1}^{n-1} (p_i^A)^{\xi_i}}{\prod_{i=1}^{n-1} (p_i^0)^{\xi_i}} \quad (1.14)$$

is called a *composite* likelihood ratio test, because the probabilities p_i^0 and p_i^A of neighbouring variants are not independent and thus not constituting a standard likelihood. Kim and Stephan [2002] modified this basic approach by allowing the p_i^A to vary within the considered genomic region: the probabilities are specified using a model for a selective sweep with two parameters, one for recombination rate scaled by strength of selection and another for the position of the selected site, both estimated by maximizing the likelihood. This idea was implemented in the program SWEEPfinder by Nielsen [2005], with the modification of using genome-wide observed frequencies as p_i^0 instead of relying on a specific null model as proposed by Kim and Stephan [2002]. Pavlidis et al. [2013] created a competing program SWEED, supposedly an order of magnitude faster and with an option to compute p_i^0 for various demographies. Finally, the SWEEPfinder was updated to version number 2 by DeGiorgio et al. [2016] to include mutation rate variation within a genome and allowing for background selection.

1.4.3 Tests based on haplotypes

Although the site frequency spectrum has proven to be a useful summary statistic, it has the major caveat that it does not capture *linkage*, the correlation of variants at different genomic positions. Figure 1.4 shows two different patterns of variation reflecting different evolutionary scenarios, but sharing the site frequency spectrum. They can be discriminated, though, by their *haplotypes*, the succession of variants on the same sequence. While in the left alignment the four lower sequences which contain the central derived variant, show no great difference in variation to the upper sequences with the ancestral variant at the central position, in the second alignment the four lower sequences are more similar to each other than the upper sequences. Sequences 5 and 6 are even identical and said to be *homozygous*, slightly extending the standard definition, referring to a genotype of a single individual, to the population level. A seminal paper by Sabeti et al. [2002] introduced the *Extended Haplotype Homozygosity (EHH)* as a measure for the similarity of sequences around a given position. This value is calculated separately for sequences containing the derived variant and sequences containing the ancestral variant of a scrutinized SNP. Comparison of the two values at an arbitrarily preset distance from this SNP yields the *Long Range Haplotype (LRH)* test. Since recombination events tend to distribute variation among sequences, the *EHH* value is an indirect measure for the amount of recombination events that has happened since the emergence of the derived variant, and this in turn is a proxy for its age. A “young” derived variant with a high population frequency is taken as signal for an ongoing selective sweep. Many modifications have been proposed since, with names such as *LDD* [Wang et al., 2006], *iHS* [Voight et al., 2006], *XP-EHH* [Sabeti et al., 2007], *rsb* [Tang et al., 2007], *nS_L* [Ferrer-Admetlla et al., 2014]. An overview of these tests is given in the boxed area of the fourth reprinted article and therefore omitted here.

It should be added, that these tests have also disadvantages. One is technical and a consequence of current sequencing technologies which produce too short reads to collate them unambiguously to the two chromosomes of a diploid organism. Thus, haplotypes extending beyond the length of reads have to be

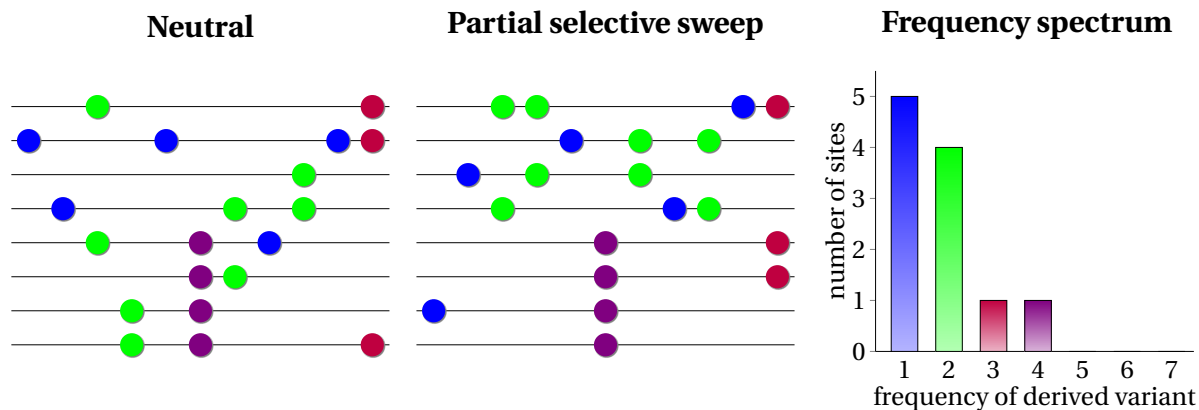


Figure 1.4: Limitations of the frequency spectrum. Notwithstanding the suggestive ordering of the central variant, the pattern on the left represents an almost perfectly neutrally evolving region, while the pattern in the middle is typical for an on-going selective sweep, since the sequences carrying the derived variant show less variation than the sequences carrying the ancestral variant (compare with Figure 1.1). The frequency spectrum is the same for both, though. The value for TAJIMA’s D is -0.20 , arguing rather for neutral evolution, and 1.02 for FAY&WU’s H , indicating the absence of derived variants with a high frequency. On the contrary, haplotype based summary statistics are able to distinguish the two patterns. For instance, the integrated EHH (iHH) for derived and ancestral variant of the central SNP yields ratios $\frac{iHH_d}{iHH_a} = \frac{3258}{2948} \approx 1.1$ for the left alignment and $\frac{8917}{1882} \approx 4.7$ for the right alignment.

reconstructed or *phased* computationally [Browning and Browning, 2011], although experimental remedies are being explored [Huang et al., 2017]. A deeper caveat with tests exploiting haplotype structure is that the signal they detect gets diluted faster than that contained in frequency spectra. Based on simulation studies, Sabeti [2006] estimated the time that a selective sweep can be detected in humans as 250.000 years by TAJIMA’s D , 80.000 years by FAY&WU’s H and less than 30.000 years by tests on haplotypes.

1.4.4 A combined test

Grossman et al. [2010, last author P. Sabeti] combined several summary statistics that showed little correlation under neutrality into the *Composite of Multiple Signals (CMS)* test. The following five statistics were chosen: three previously known (F_{ST} , iHS and $XP - EHH$) and two newly defined (ΔDAF and ΔiHH). Their distributions under neutrality and under a selective sweep were computed by simulations, allowing for demography parameters for the investigated populations. If s_i , $i = 1, \dots, 5$, are the values of the five statistics for an experimentally observed SNP, then the following product of posterior probabilities is taken as its score

$$CMS = \prod_{i=1}^5 \frac{P(s_i|selected)P(selected)}{P(s_i|selected)P(selected) + P(s_i|unselected)P(unselected)}. \quad (1.15)$$

This test was aimed to narrow the signal within an already otherwise established candidate region. The authors assumed that exactly one of the SNPs in such a region is under selection while the remainder are neutral. A uniform prior probability was chosen, meaning that each SNP has the same chance of being the selected one, $P(selected) = \frac{1}{\#SNP}$. Later, Grossman et al. [2013] adapted this test for genome-wide detection of selection signals for which the specification of prior probabilities is not possible

$$CMS_{GW} = \prod_{i=1}^5 \frac{P(s_i|selected)}{P(s_i|unselected)}. \quad (1.16)$$

1.5 Selection scans in humans

Haas and Payseur [2016] compiled a non-exhaustive list of 73 human-specific whole genome scans for selection. Here, instead, an account of only a few milestones shall be given as an overview.

A prerequisite for whole genome scans in human variation data is obviously the genome sequence itself, which, as is generally known, was presented as a draft version in the year 2000.

1.5.1 An early scan using F_{ST}

Akey [2002] performed a scan for both positive and balancing selection. He calculated F_{ST} values of 26530 genome-wide SNPs genotyped by the SNP consortium [Thorisson and Stein, 2003] in samples of European Americans, African Americans and East Asians, each consisting of 42 individuals. 8862 SNPs resided in the vicinity of a known gene. Genes with at least one SNP showing a F_{ST} value among the 2.5% highest genome-wide were proposed as likely to be under positive selection, yielding a total of 156, among them genes underlying diseases such as *CFTR*, associated with cystic fibrosis and *PPARG*, associated with diabetes type II. In order to qualify as candidate gene for balancing selection, essentially identical variant frequencies in the three populations were required. Among the 18 genes thus identified were *guanine nucleotide exchange factor for Rap1 (GFR)* and *tropomodulin 3 (TMOD3)*.

1.5.2 Scans for positive selection

The scan for positive selection of Carlson et al. [2005] used data generated by the company Perlegen Sciences[®], consisting of about 1,6 million SNPs genotyped in 71 Americans, having either European (24), African (23) and Chinese (24) ancestry [Hinds et al., 2005]. For each population group TAJIMA's D was calculated in overlapping windows of size 10^5 bases, sliding over the whole genome. In total 55 candidate regions for selection were identified, defined by multiple contiguous windows having values that belong to the lowest 1% empirical quantile. Furthermore, the paper addresses the influence of *ascertainment bias*, since the SNPs genotyped by array chips as in this case, are biased towards common variants, leading to a skewed frequency spectrum. Comparison with a set of previously fully sequenced genes as well as a re-sequencing done on some genes within the candidate regions, showed that the TAJIMA's D values for both kinds of data are significantly correlated, albeit with a merely intermediate correlation coefficient R^2 .

The first phase of the HapMap project yielded 1 million SNPs genotyped in 90 individuals with European ancestry from Utah (CEU), 90 individuals from the Yoruba in Ibadan (YRI), 45 Han Chinese from Beijing (CHB) and 44 Japanese from Tokyo (JPT). The CEU and YRI samples consisted of 30 trios (father, mother, child) while the other two samples represented unrelated individuals. In the publication associated with the data release [The international HapMap Consortium, 2005], regions with extreme values in four summary statistics were reported. The calculation of F_{ST} values yielded 926 SNPs more differentiated than the a specific variant at the *Duffy* locus which confers resistance against malaria and hence is likely to have been under selection; among them 32 were non-synonymous coding, including 6 within the gene *ALMS1*. The outliers of the *long-range haplotype (LRH)* test were headed by the *LCT* gene in the CEU sample. The supplement contains lists of regions with low heterozygosity and long haplotypes, respectively.

Voight et al. [2006] introduced the *iHS* statistic and applied it to the HapMap SNPs. They observed an excess of extreme values and furthermore a conspicuous clustering of these with respect to simulations of neutral evolution including a variety of demographies. The longest observed haplotypes with a derived allele frequency of over 50 % were found near the Gaucher disease gene *GBA* in the CHB/ JPT sample, near a gene involved in insulin regulation (*NKX2-2*) in CEU and in a region without known genes on chromosome

1 Introduction

5 in YRI. Another ranking restricted to coding regions yielded an enrichment of several categories with the most significant being “other carbohydrate metabolism” and “chromatin packaging” in CHB/JPT, “electron transport” and “MHC1-mediated immunity” in CEU, and “steroid metabolism” in YRI [Voight et al., 2006, table 2].

The publication of “HapMap 2” [The international HapMap Consortium, 2007], describing an additional 2.1 million SNPs genotyped in the same samples as in the first phase, was accompanied by a study devoted to detect selection [Sabeti et al., 2007] using *LRH*, *iHS* and a newly defined cross-population haplotype test *XP-EHH*. 22 regions were found with test statistics so extreme, that they did not occur in simulations of 10 Gigabases. These regions contained 9166 SNPs on which three filters were applied: a selection candidate had to be derived, highly differentiated among populations and belonging to known functional genetic elements. Of the remaining 41 SNPs, 8 were found to cause amino acid changes in the genes *SLC24A5*, *EDAR*, *PCDH15*, *ADAT1*, *KARS*, *HERC1*, *SLC30A9* and *BLFZ1*. Additionally, the filtering process was reversed by starting with non-synonymous coding SNPs and ranking them by the other two criteria. In the end, for each population two different candidate genes were found that were tightly connected functionally: *LARGE* and *DMD*, related to susceptibility to the Lassa virus, in the YRI sample, *SLC24A5* and *SLC45A2*, involved in skin pigmentation in the CEU sample and *EDAR* and *EDA2R*, both trans-membrane receptors, involved in the development of hair follicles, in the CHB/JPT sample.

Williamson et al. [2007, last author R. Nielsen] scanned the Perlegen[®] data with the SWEEPfinder. They used a sliding window of 201 SNPs and computed p-Values by simulations of neutral evolution incorporating the ascertainment bias of the data set and an approximate recombination rate for each window. 164 windows showed a p-Value of less than 10^{-5} , 101 of them had an annotated gene within 10^5 bases distance. The strongest signal showed the gene *DTNA* in the Chinese sample and only slightly less in the sample of European Americans. This gene is a component of the *dystrophin protein complex (DPC)*, important for the architecture of muscles; several other genes of this complex showed signals of selection in this sample. Further gene categories presented as being under selection comprised skin pigmentation, olfactory receptors, hair morphology and heat shock proteins. Several centromeres were reported to show evidence for selection. In total, 10% of the genome of Europeans and Chinese were declared to have been affected by selective sweeps, identified in the study by windows with a p-value of 0.05 or less.

Pickrell et al. [2009, last author K. Pritchard] screened a set of 657143 SNPs, genotyped by Li et al. [2008] in 938 individuals of 53 populations from the “Human Genome Diversity Project” [Cavalli-Sforza, 2005]. They scanned geographically grouped populations with *iHS*, *XP-EHH* and the SWEEPfinder and used F_{ST} for closely related single populations. Instead of an enrichment analysis for biological processes, they examined, if particular a-priori gene sets, associated with pigmentation or one of several diseases, showed stronger signals of selection than random loci in the genome. For the former category this was clearly the case, while among diseases this held only for diabetes II, with SNPs associated with the disease not matching those showing signals of selection. Apart from that, multiple genes of the pathway *NRG-ERBB4*, involved in the development of a number of tissues, displayed extreme values in several populations.

In the HapMap 3 project [The international HapMap Consortium, 2010], the samples of the previously investigated populations were enlarged and supplemented by 7 further populations: African ancestry in the Southwestern USA (ASW), Chinese in metropolitan Denver, USA (CHD), Gujarati Indians in Houston, USA (GIH), Luhya in Webuye, Kenya (LWK), Maasai in Kinyawa, Kenya (MKK), Mexican ancestry in Los Angeles, USA (MXL) and Tuscans in Italy (TSI). In total, 1.6 million SNPs were genotyped in 1184 individuals. The same tests as in HapMap2 were applied, followed by a fine-mapping of candidate regions by *CMS*. Candidate genes in the new populations were *KITLG* and *MLPH*, both involved in pigmentation, *LAMA3*, involved in wound healing and an olfactory cluster in population TSI, immune related genes *CD226*, *ITGAE* and *DPP7* in both Kenyan populations and the gene *ANKH*, having a role in bone growth, in MKK.

The 1000 genomes project began with a pilot phase, consisting of very low coverage sequencing of

179 individuals from the same populations as in HapMap 1 + 2 and yielding 15 million SNPs [The 1000 Genomes Project Consortium, 2010]. A scan using F_{ST} was performed which turned out very few fixed differences ($F_{ST} = 1$) between populations: 2 SNPs between CEU and CHB+JPT (one of them in the gene *SLC24A5*), 4 between CEU and YRI (including a mutation next to the *Duffy* blood group gene *DARC*), and 72 between CHB+JPT and YRI (24 of them clustered around the gene *EXOC6B*, necessary for exocytosis). However, 139 non-synonymous variants showed very high values ($F_{ST} \geq 0.8$), including two genes involved in meiotic recombination (*FANCA* and *TEX15*).

On this data set Grossman et al. [2013] performed a whole genome scan using CMS_{WG} to delineate candidate regions, complemented by those of The international HapMap Consortium [2007]. Fine-mapping these regions with the CMS yielded 412 regions of median length 27 kb which contained a median of 47 SNPs. These regions contained only 35 amino acid-changing variants, but 59 variants associated with expression levels measured in the cell lines used for sequencing, among them 48 long inter-genic non-coding RNAs (lincRNAs). One of the non-synonymous coding variants, L616F in the innate immune system gene *TLR5*, was experimentally shown to result in different responses to bacterial flagellin.

In phase 1 of the 1000 genomes project, samples from 14 populations were sequenced to yield 38 million SNPs [The 1000 Genomes Project Consortium, 2012]. The populations CHS, CLM, FIN, GBR, IBS and PUR were newly included with respect to HapMap samples; for abbreviations see Table 1.1. Signals of selection were searched by means of the population differentiation measure ΔDAF [The 1000 Genomes Project Consortium, 2012, Table S12]. The SNP showing the biggest derived frequency difference between the two non-admixed African samples LWK and YRI had a value of 0.475 and lies in a putative binding site of the *Neuron Restrictive Silencing Factor* (*NRSF*). Pybus et al. [2014] calculated various test statistics for the populations CEU, CHB and YRI, available via a “selection browser” (<http://hsb.upf.edu/>).

The 1000 Genomes Project Consortium [2015], concluding phase 2 and final phase 3, presented the sequences of 2504 individuals of 26 populations (Table 1.1). Among other kinds of variation, 78 million SNPs were “called”. SNPs in genes were scanned for outliers with the differentiation measure PBS , applied on populations of the same continental group with the remaining continents as out-group. The gene *SLC24A5* showed high values within all five continental groups, while SNPs in the genes *TRBV9* and *PRICKLE4* belonged to the most highly differentiated within both South Asians and Africans [The 1000 Genomes Project Consortium, 2015, Extended data Figure 8].

An altogether different scan on positive selection was performed by Mathieson et al. [2015], who used *ancient* genomes from 213 people living in Europe between 6500 and 300 BC, to compare them with present-day Europeans. In this case, frequency changes need not be inferred from present day variation, but can be measured directly. The problem is here to relate the correct populations in time, since migration confounds a purely geographic association. In order to assess the significance of frequency changes, the authors tested the hypothesis that variant frequencies in four populations of the 1000 genomes project (CEU, GBR, IBS and TSI) can be described by a linear mixture of those in the supposedly three ancestral populations “Early Farmers”, “Hunter-gatherers” and “Steppe ancestry”. Twelve signals of selection were reported, among them SNPs associated with lactose persistence (*LCT*), fatty acid metabolism (*FADS1*), Vitamin D regulation (*DHRC7*), pigmentation (*SLC45A2* and *GRM5*), innate immunity (*TLR1-6-10* cluster) and adaptive immunity (*MHC*). Two variants determining light skin (*SLC24A5*) and light eye colour (*HERC2/OCA2*), respectively, had already high frequency in one of the ancestral populations.

1.5.3 Scans for balancing selection

Bubb et al. [2006] searched for particularly diverse regions in the genome. In order to avoid false positives due to sequence errors they did not use the SNPs called by The international HapMap Consortium [2005], but instead used the primary data of the project to focus on high-quality reads showing more than aver-

1 Introduction

East Asians	CHB	Han Chinese in Beijing, China	103
	CHS	Han Chinese South, China	105
	CDX	Chinese Dai in Xishuangbanna, China	93
	JPT	Japanese in Tokyo, Japan	104
	KHV	Kinh in Ho Chi Minh City, Vietnam	99
South Asians	BEB	Bengali in Bangladesh	86
	GIH	Gujarati Indian in Houston, Texas	103
	ITU	Indian Telugu in the UK	102
	PJL	Punjabi in Lahore, Pakistan	96
	STU	Sri Lankan Tamil in the UK	102
Africans	ESN	Esan in Nigeria	99
	GWD	Gambian in Western Division, The Gambia	113
	LWK	Luhya in Webuye, Kenya	99
	MSL	Mende in Sierra Leone	85
	YRI	Yoruba in Ibadan, Nigeria	108
Europeans	CEU	Utah Residents with Northern and Western European Ancestry	99
	FIN	Finnish in Finland	99
	GBR	British in England and Scotland	91
	IBS	Iberian populations in Spain	107
	TSI	Tosceni in Italy	107
Admixed Americans	ACB	African Caribbean in Barbados	96
	ASW	African Ancestry in Southwest USA	61
	CLM	Colombian in Medellín, Colombia	94
	MXL	Mexican Ancestry in Los Angeles, California	64
	PEL	Peruvian in Lima, Peru	85
	PUR	Puerto Rican in Puerto Rico	104
		26	2504

Table 1.1: Populations and sample sizes of The 1000 Genomes Project Consortium [2015].

age deviation from the human reference sequence. The SNPs within these reads were controlled by PCR in 10 Americans of African ancestry and further filtered by searching for variation in the flanking regions, leaving in the end 16 regions of high diversity. The two most differing haplotypes within each region were manually identified and additional individuals sequenced to obtain at least three sequences of each haplotype with a length of 20kb. The 5kb part that showed the highest divergence within these re-sequenced regions was finally reported. They compared these regions with two other regions, believed to be under balancing selection: the *major histocompatibility complex (MHC)* region, of central importance in adaptive immunity, and the *ABO* gene, responsible for the major blood group system. The divergence of all newly found candidate regions was comparable to that of the *ABO* gene, while the *MHC* region stuck out by an order of magnitude. The 16 regions however, did not show any enrichment of genes or conserved non-coding sites. Furthermore, simulations showed that their levels of divergence could be well explained by neutral evolution alone. The authors concluded that they do not represent instances of balancing selection.

Andrés et al. [2009, last author R. Nielsen] used a data set of Bustamante et al. [2005], comprising fully sequenced exons of 13400 genes in 19 African Americans and 20 European Americans. After a quality filtering, 4877 genes remained, each containing at least one SNP. Two tests were applied: a one-sided *HKA*-test, which scales human diversity by divergence to chimpanzee and a one-sided *MAN-WHITNEY-U*-test on the folded frequency spectrum, used to discern genes with an excess of SNPs of intermediate frequency. 60 “extreme genes” were found that fell into the 5% significance level of both tests, with critical values established by simulations of neutral evolution. The set contained about a dozen genes involved in immunity such as three of the *MHC* and a gene *FUT2*, affecting a minor blood group, but not the *ABO* gene. A few molecular function categories were enriched, such as “extracellular matrix”, “intermediate filament” and “serine protease inhibitor”, however no biological process category was enriched in both

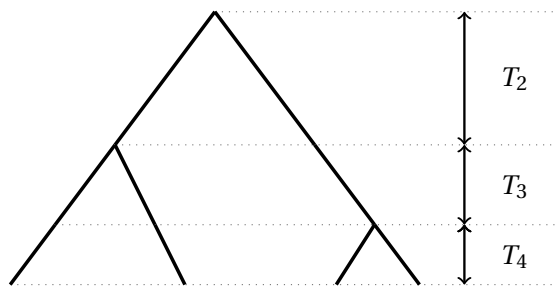


Figure 1.5: A coalescent tree for a sample of size 4. Each tip represents a present-day sequence of the sample. The sequences find their common ancestors at the bifurcation points (“coalescent events”) until the most recent common ancestor of the whole sample is found. The times between coalescent events are exponentially distributed and depend on the number of lineages in each time segment.

populations.

DeGiorgio et al. [2014, last author R. Nielsen] used high-quality whole-genome SNP data of 9 unrelated individuals from each of the populations CEU and YRI, genotyped by the company Complete Genomics® [Drmanac et al., 2010]. They devised two composite likelihood tests, applied in sliding windows over the genome. The first one compares the portion of substitutions respective to an ancestral sequence with that of polymorphisms. Under balancing selection this ratio should tilt to the latter. This ratio was derived analytically using a system of recursion equations on basis of the model by Hudson and Kaplan [1988]. The second test used the frequency spectrum extended by substitutions, i.e. derived variants of size n . These spectra were computed by simulations. The second test, being an extension of the first, performed better. Several genes of the *MHC* were detected by the tests. However, once these excluded, no specific category of genes appeared to be enriched among the candidate genes. One of the highest ranking genes was *FANK1*, suspected to distort equal segregation of chromosomes during meiosis.

1.6 Coalescent mathematics

The $\frac{1}{x}$ “law” for the expected values of the frequency spectrum has been known already a long time [Kimura, 1964]. Fu [1995] re-derived it using coalescent theory and, more importantly, calculated for the first time exact expressions for the covariances of its components. Three of the articles reprinted in this thesis deal with an extension of his work. This section introduces the mathematics involved.

The coalescent model as invented by Kingman [1982] is aimed at describing neutral genetic variation as a succession of two random processes, namely the creation of a genealogical or coalescent tree and the subsequent “addition” of mutations. If need be, these mutations can be placed randomly into a genomic region to mimic experimental sequence data. A particular coalescent tree is characterized by a series of ordered random bifurcations, *coalescent events*, defining its *topology*, and its branch lengths which symbolize elapsed time (Figure 1.5).

Starting from the present day leaves, any two lineages have the same probability to join, i.e. to coalesce. The time between such coalescence events is modelled as an exponential distribution with parameter $\lambda = \frac{i(i-1)}{4N}$,

$$T_i \sim \text{Exp}\left(\frac{i(i-1)}{4N}\right) \quad (1.17)$$

and from elementary statistical theory its expectation value is known to be $\frac{1}{\lambda}$ or

$$E[T_i] = \frac{4N}{i(i-1)}. \quad (1.18)$$

Mutations are supposed to be rare events that happen with a probability proportional to time. They are modelled by a Poisson-distribution with parameter μT (Figure 1.6).

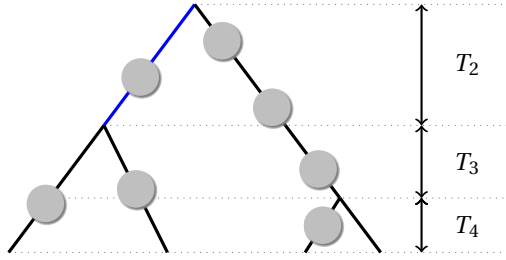
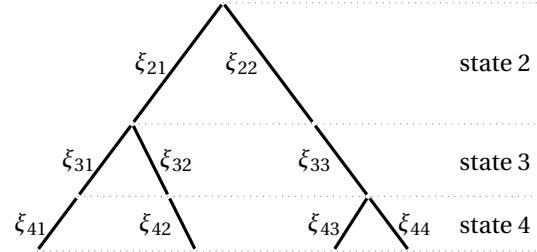


Figure 1.6: A coalescent tree with mutations. Mutations occur randomly “on” branches and follow a Poisson distribution with mean equal to the length of the branch, i.e. time. For instance, the expected number of mutations occurring on the blue branch is $E[\mu T_2] = \mu \frac{4N}{2(2-1)} = 2N\mu = \frac{\theta}{2}$.

Figure 1.7: States and lines of a coalescent. The states k are defined by the number of lineages present. The branches are split along the states into lines kl on which ξ_{kl} mutations occur. The line numbering within each state, denoted by l , is arbitrary and serves only to distinguish lines.



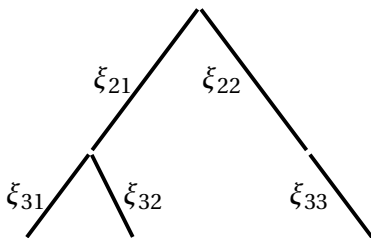
So far, this is standard coalescent theory as covered in text-books like Wakeley [2008]. The approach of Fu [1995] introduces two further concepts: *states* and *lines*. A state simply marks the number of lineages at a given time. The branches of the coalescent are subdivided along the states into lines. At each state k there are k different lines (Figure 1.7).

For any line resp. combination of lines, the expected number of mutations and higher moments can be calculated with elementary statistics. For instance, the expected value of mutations occurring on a line kl yields

$$E[\xi_{kl}] = E[T_k \mu] = \frac{4N}{k(k-1)} \mu = \frac{\theta}{k(k-1)}. \quad (1.19)$$

The expected frequency spectrum for sample size $n = 3$

For sample size $n = 3$ exists only a single tree topology. Mutations occurring on line 21 appear on two sequences and hence are of size 2. Mutations happening on all other lines are of size 1. The sum over the expected mutations on each line yields the expected neutral spectrum.

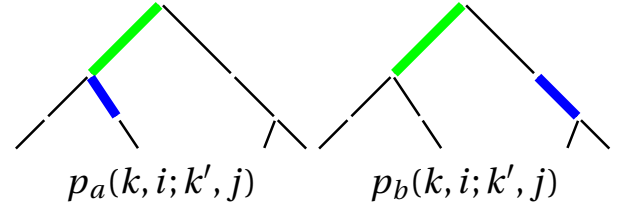


$$\begin{aligned} E[\xi_1] &= E[\xi_{22}] + E[\xi_{31}] + E[\xi_{32}] + E[\xi_{33}] \\ &= \theta \frac{1}{2(2-1)} + \theta \frac{1}{3(3-1)} + \theta \frac{1}{3(3-1)} + \theta \frac{1}{3(3-1)} \\ &= \theta \frac{1}{2} + 3\theta \frac{1}{6} \\ &= \theta \end{aligned}$$

$$\begin{aligned} E[\xi_2] &= E[\xi_{21}] \\ &= \theta \frac{1}{2(2-1)} \\ &= \frac{\theta}{2} \end{aligned}$$

For sample size $n = 3$ the site frequency spectrum can be calculated directly using Eq. (1.19), see Box. For any larger sample one has to allow for different tree topologies. In order to do this, we need the probability that a mutation occurring on line kl is of size i , or with other words, that line kl from state k has i

Figure 1.8: The two possible relationships of two lines and there associated probabilities. The green line symbolizes a line kl and the blue line a line $k'l'$, which may or may not be a descendant of line kl . The coincidence of both lines is a special case of the first type of relation. No other configurations are possible in a non-recombining tree.



descendants at state n . This probability can be taken from a specialized branch of statistics called *Pólya urn theory* [Mahmoud, 2008] and yields

$$p(k, i) = \frac{\binom{n-i-1}{k-2}}{\binom{n-1}{k-1}}. \quad (1.20)$$

With the help of Eq. (1.20), it is possible to calculate the site frequency spectrum for any sample size. All lines of all states are summed up with respect to their probabilities to yield i descendants:

$$\begin{aligned} E[\xi_i] &= \sum_{k=2}^n \sum_{l=1}^k p(k, i) E[\xi_{kl}] \\ &= \sum_{k=2}^n k p(k, i) E[\xi_{k1}] \\ &= \sum_{k=2}^n k \frac{\binom{n-i-1}{k-2}}{\binom{n-1}{k-1}} \frac{\theta}{k(k-1)} \\ &= \sum_{k=2}^n k \frac{\binom{n-k}{i-1}}{\binom{n-1}{i}} \frac{k-1}{i} \frac{\theta}{k(k-1)} \\ &= \frac{\theta}{i} \sum_{k=2}^n \frac{\binom{n-k}{i-1}}{\binom{n-1}{i}} \\ &= \frac{\theta}{i} \sum_{k=0}^{n-2} \frac{\binom{k}{i-1}}{\binom{n-1}{i}} \\ &= \frac{\theta}{i} \end{aligned} \quad (1.21)$$

The fourth step involves an easy-to-prove rearrangement of binomial coefficients and corresponds to Eq. 14 of [Fu, 1995]. The last step exploits the *hockey-stick identity* $\sum_{k=m}^n \binom{k}{m} = \binom{n+1}{m+1}$.

The covariances and any higher moments of the site frequency spectrum can be computed analogously, however the probabilities involved are joined probabilities, which get increasingly complex and the calculation of which has to be separated into cases. For the derivation of the covariances Fu [1995] needed to consider essentially two cases (Figure 1.8):

- $p_a(k, i; k', j)$ the probability that a line kl at state k has i descendants at state n and another line $k'l'$ at state $k' \geq k$ is a descendant of line kl and has j descendants at state n .
- $p_b(k, i; k', j)$ the same as above, but the line $k'l'$ at state k' is not a descendant of line kl .

Given these probabilities, the covariances can be easily written down, however the simplification of the resulting nested sums is tedious. The main result of Fu [1995] is hence the following:

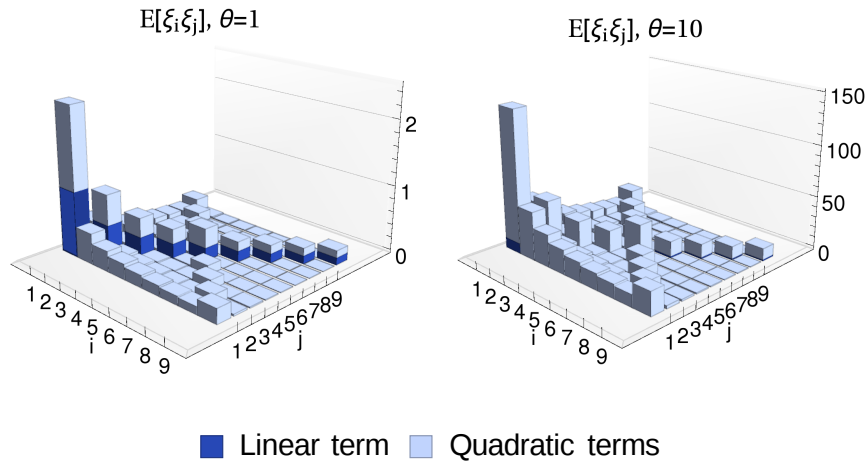


Figure 1.9: The expected values $E[\xi_i \xi_j]$ for sample size $n = 10$, coloured by the respective shares of the linear and quadratic parts of Eq. (1.22).

The second moments have the structure

$$E[\xi_i \xi_j] = \delta_{i=j} \frac{1}{i} \theta + \tau_{ij} \theta^2 \quad (1.22)$$

with

$$\tau_{ij} = t_a(i, j) + t_a(j, i) + t_b(i, j) + t_b(j, i) \quad (1.23)$$

and

$$t_a(i, j) = \begin{cases} \frac{1}{2} (\beta_n(j) - \beta_n(j+1)) & \text{if } j < i \\ \frac{1}{2} \beta_n(j) & \text{if } j = i \end{cases} \quad (1.24)$$

$$t_b(i, j) = \begin{cases} \frac{1}{ij} - \frac{1}{i(i+j)} - \frac{1}{2} (\beta_n(j) - \beta_n(j+1)) & \text{if } i + j < n \\ \alpha_n(j) - \frac{1}{2} \beta_n(j) & \text{if } i + j = n \end{cases} \quad (1.25)$$

where $\alpha_n(i)$ and $\beta_n(i)$ are simple fractions involving harmonic numbers. Figure 1.9 shows the second moments of the site frequency spectrum for sample size $n = 10$. Obviously the covariances can be calculated by $Cov[\xi_i, \xi_j] = E[\xi_i \xi_j] - E[\xi_i]E[\xi_j]$.

1.7 Adaptations of tests on neutrality

This section gives an overview over the four articles presented in the next chapter. A more profound placing into the scientific literature is given in the publications themselves and not repeated here.

The TAJIMA'S D -like tests presented in subsection 1.4.2 have been built upon the standard coalescent model which presupposes a population of constant size. If that assumption is violated, application of the tests may lead to spurious results. Modifying the coalescent model accordingly is relatively simple, since changing population size is reflected merely by changes of the time distribution between coalescent events. However, if the exponential distribution in Eq. (1.17) has to be modified or altogether replaced by another distribution, the formulas for the first and second moments get much more involved. While the first moment given by Eq. (1.21) have been generalized already by Griffiths and Tavaré [1994] to populations of varying size, Živković and Wiehe [2008] did so for the second moments, hence they adapted

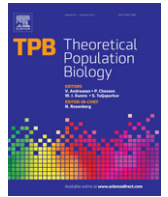
Eqs. (1.24) and (1.25). Using these results they modified some of the tests given in subsection 1.4.2 and applied them to data from two populations of fruit flies. In article 1 that adaptation to population size changes is extended to the whole class of tests within the framework of Achaz [2009]. Furthermore, the genome scan for positive selection performed by Carlson et al. [2005] was updated using standard as well as demography-adapted tests and newer data.

As explained in section 1.4.3, the frequency spectrum has its limits in representing the information contained in an alignment. It is formed by independent counts of single mutations, no matter in which configuration they are with each other. The spectrum of two linked variants, yielding a two-dimensional spectrum, is presented in the second article as a step towards a more comprehensive exploitation of the data. Its calculation is relatively straightforward, given the variances of the standard one-dimensional frequency spectrum. If in that spectrum the frequency of one variant is held fixed, one yields a conditional one-dimensional spectrum. Then, the classification of two variants with respect to their relative position in a coalescent tree into five distinct classes opens the way for finer-grained analyses. These classifications increase the information content transported by a 2-loci spectrum. Curiously, the corresponding joined probabilities were already implicit in the proofs by Fu [1995], yet never brought to much attention.

The third article is intimately related to the second. The third moments of the frequency spectrum are derived using largely the same technics as Fu [1995], yet the extra dimension brings with it rather bulky expressions. An immediate corrolate of the third moments are the covariances of a conditional spectrum as described above. In addition, the third moments enable for the first time an analytical approximation to the distributions of the TAJIMA's D -like test statistics described in section 1.4.2.

The *Extended Haplotype Homozygosity (EHH)* is a measure that is calculated from a focal SNP, whose variants define two initial haplotypes, to subsequent farther SNPs on either side. The implementation for this calculation in version 1 of the R-package REHH was very inefficient; for each further SNP the whole calculation was repeated. However, the calculation conforms to the "Markov property" in the sense that the calculation for SNP number $x + 1$ depends on the result for SNP x , but not on any of the more previous ones. Hence the calculation can be done stepwise, holding track of the different haplotypes until the current SNP. It turned out that an indexation of haplotypes is feasible that requires merely one array of integer numbers with a size equal to the amount of sequences, hence negligible memory space.

2 Publications



Demography-adjusted tests of neutrality based on genome-wide SNP data



M. Rafajlović^{a,e,1}, A. Klassmann^{b,1}, A. Eriksson^{c,d}, T. Wiehe^b, B. Mehlig^{a,e,*}

^a Department of Physics, University of Gothenburg, SE-412 96 Gothenburg, Sweden

^b Institut für Genetik, Universität zu Köln, 50674 Köln, Germany

^c Department of Zoology, University of Cambridge, CB2 3EJ Cambridge, UK

^d Integrative Systems Biology Lab, King Abdullah University of Science and Technology, Thuwal 23955-6900, Saudi Arabia

^e The Linnaeus Centre for Marine Evolutionary Biology, University of Gothenburg, SE-405 30 Gothenburg, Sweden

ARTICLE INFO

Article history:

Received 5 July 2013

Available online 6 June 2014

Keywords:

Single nucleotide polymorphism

Infinite-sites model

Site frequency spectrum (SFS)

Bottleneck

Coalescent approximation

ABSTRACT

Tests of the neutral evolution hypothesis are usually built on the standard null model which assumes that mutations are neutral and the population size remains constant over time. However, it is unclear how such tests are affected if the last assumption is dropped. Here, we extend the unifying framework for tests based on the site frequency spectrum, introduced by Achaz and Ferretti, to populations of varying size. Key ingredients are the first two moments of the site frequency spectrum. We show how these moments can be computed analytically if a population has experienced two instantaneous size changes in the past. We apply our method to data from ten human populations gathered in the 1000 genomes project, estimate their demographies and define demography-adjusted versions of Tajima's D , Fay & Wu's H , and Zeng's E . Our results show that demography-adjusted test statistics facilitate the direct comparison between populations and that most of the differences among populations seen in the original unadjusted tests can be explained by their underlying demographies. Upon carrying out whole-genome screens for deviations from neutrality, we identify candidate regions of recent positive selection. We provide track files with values of the adjusted and unadjusted tests for upload to the UCSC genome browser.

© 2014 Elsevier Inc. All rights reserved.

1. Introduction

In natural populations, genetic diversity is shaped not only by population genetic forces such as drift and natural selection, but also by geographic structure and demographic history. In order to identify genome regions affected by natural selection many statistical tests of neutrality have been designed in the past. Typically, they are based on properties of the site frequency spectrum (SFS) (e.g. Tajima's D Tajima, 1989a) or of the haplotype structure (e.g., EHH Sabeti et al., 2002, and their various derivatives) and in numerous studies they have been applied to the human genome (Akey et al., 2004; Stajich and Hahn, 2005; Carlson et al., 2005; Nielsen et al., 2005; Voight et al., 2006; Grossman et al., 2013). One of the main challenges in interpreting the results of such scans is

to distinguish between the effects of selection and of the underlying unknown demography upon genetic variation. Indeed, tests of neutrality are usually built on two null assumptions, neutrality of mutations and constancy of population size. When empirical data from only a few genomic regions are available, the quantiles of test statistics serve as a basis for detecting deviations from "neutrality". However, if the assumption of constant population size is false, both mean and variance of the test distributions can markedly differ from the theoretical expectations, even if mutations are neutral. Thus, the quantiles of the null distributions do not contain enough information to decide which of the two assumptions (or both) are violated. However, it is well recognised (see Akey et al., 2004 and references therein) that the effects of demographic history are visible on a genome-wide scale while those of natural selection are expected to be local. Hence, assuming that most of genome-wide genetic variation is neutral, the empirical test distributions should be mainly shaped by demographic history, and distortions due to selection can be ignored. Therefore, a number of authors (Stajich and Hahn, 2005; Carlson et al., 2005; Nielsen et al., 2005; Voight et al., 2006; Grossman et al., 2013) encouraged the use of

* Corresponding author at: Department of Physics, University of Gothenburg, SE-412 96 Gothenburg, Sweden.

E-mail address: Bernhard.Mehlig@physics.gu.se (B. Mehlig).

¹ These authors have equally contributed to this work.

empirical distributions from whole-genome data as a background against which to search for local deviations from neutrality. But, as pointed out by Marth et al. (2004) (see also references therein), it is not clear what the percentage of genome regions targeted by selection is, and thus it is impossible to quantify to which extent they distort the empirical whole-genome distributions. Moreover, the variance of the empirical distribution depends strongly on the underlying demography, and it is thus very difficult to quantify and compare the amount of deviation from neutrality of a given region between populations with different demographies.

In this study we show that such a comparison is possible and meaningful, provided the demography, estimated from genome-wide single nucleotide polymorphisms, is integrated into SFS-based tests, such as Tajima's D , Fay and Wu's H (Fay and Wu, 2000), and Zeng's E (Zeng et al., 2006). We call these modified tests *demography-adjusted*.

One method to estimate the demography of a given population is based on a maximum likelihood (ML) analysis applied to the spectrum of intergenic, physically distant SNPs (Nielsen, 2000; Adams and Hudson, 2004; Marth et al., 2004). We use coalescent simulations to analyse the performance of ML demography estimation in dependence on the number of SNPs used for the inference. Our analysis is focused on a piecewise constant population-size model involving at most two instantaneous population-size changes. Such a model was used (Adams and Hudson, 2004; Marth et al., 2004; Stajich and Hahn, 2005) to capture the main events of the human out-of-Africa expansion (Cavalli-Sforza and Feldman, 2003; Ramachandran et al., 2005; Liu et al., 2006; Tanabe et al., 2010; Eriksson et al., 2012).

To be able to define demography-adjusted tests, we derive analytical expressions for the first and second moments of the site frequency spectrum under the demographic model. For an idealised population of constant size ("Wright-Fisher-model") our adjusted tests are identical to the original (unadjusted) ones.

We apply both unadjusted and adjusted tests to 10 populations from the 1000 genomes project (McVean et al., 2012), version 3, released April 30th, 2012. We find that the empirical distributions of unadjusted tests substantially differ between different populations, whereas the distributions of the corresponding demography-adjusted tests are very similar between different populations. This suggests that differences in the empirical distributions of the unadjusted tests are mainly caused by the differences in the underlying demographies. Our results further show that demography adjustment is reflected in an affine linear transformation of the test statistics. Therefore, the identification of regions under selection by means of empirical quantiles is not affected by the adjustment. However, by correcting for demographic effects, the demography-adjusted tests allow one to compare the extent of deviation from neutrality between different populations. For the unadjusted tests such a comparison is ill defined. We provide unadjusted and adjusted test values as BED-files formatted for upload to the UCSC genome browser.

2. Materials and methods

2.1. Demography-adjusted tests of neutrality

Tajima (1989a) introduced a test of neutrality comparing two unbiased estimators of the scaled mutation rate $\theta = 4\mu LN$, with N denoting the diploid population size, μ the mutation rate per site, chromosome, generation, and L the number of sites in the genomic sequence. One estimator in Tajima's D test, denoted by $\hat{\theta}_S$ below, is based on the total number of segregating sites, S , and the other, denoted by $\hat{\theta}_\Pi$ below, is based on the average number of pairwise differences, Π . If mutations are neutral, and the population size

constant, Tajima (1989a) showed that the estimators $\hat{\theta}_S$, and $\hat{\theta}_\Pi$, defined as

$$\hat{\theta}_S = \frac{S}{a_n}, \quad \text{with } a_n = \sum_{i=1}^{n-1} \frac{1}{i}, \quad \text{and } \hat{\theta}_\Pi = \Pi, \quad (1)$$

have the same expected values, that is, $\langle \hat{\theta}_S \rangle = \langle \hat{\theta}_\Pi \rangle = \theta$. Tajima's D test compares these two estimators, and it is defined as (Tajima, 1989a):

$$D = \frac{\Pi - \frac{1}{a_n} S}{\sqrt{\text{Var}\left(\Pi - \frac{1}{a_n} S\right)}}. \quad (2)$$

As shown by Achaz (2009), the numerator of Eq. (2) can be written in terms of the site frequency spectrum, ξ_i ($i = 1, \dots, n-1$), as

$$\Pi - \frac{1}{a_n} S = \sum_{i=1}^{n-1} (\omega_i^\Pi - \omega_i^S) i \xi_i, \quad (3)$$

with weightings ω_i^Π, ω_i^S ($i = 1, \dots, n-1$) which satisfy $\sum_{i=1}^{n-1} \omega_i^\Pi = \sum_{i=1}^{n-1} \omega_i^S = 1$, and consequently $\sum_{i=1}^{n-1} \omega_i^\Pi - \omega_i^S = 0$. These weightings are listed (up to normalisation) in Table 1 of Achaz (2009). Using Eq. (3) it is straightforward to show that in the constant population-size case the expected value of the numerator of Eq. (2) is equal to zero, because in this case it holds that $\langle \xi_i \rangle = \theta/i$ (Fu, 1995). However, this does not hold for an arbitrary demography where in general $\langle \xi_i \rangle \neq \theta/i$. As a result, deviations from "neutrality" captured by Tajima's D (given by Eq. (2)) are caused not only by mutations that are targeted by selection, but also by a varying population-size history. Therefore, the underlying demography needs to be integrated into the estimators (and hence into the test). A method to do this is explained next.

Following the notation introduced by Achaz (2009) and Ferretti et al. (2010), we write the neutral site frequency spectrum obtained under the actual population demographic history (null demography) in the form $\langle \xi_i \rangle = \xi_i^0 \theta$, where $\xi_i^0 = \langle \xi_i \rangle|_{\theta=1}$ is equal to one half of the expected total branch length of lineages with i leaves in a gene genealogical tree. The value of ξ_i^0 depends on the sample size n and the parameters of the demography, but not on θ . It follows that in a sample of size n , the spectrum provides $n-1$ unbiased estimators $\hat{\theta}^{(i)} = \xi_i / \xi_i^0$. In fact, any linear combination of $\hat{\theta}^{(i)}$, with weightings $\tilde{\omega}_1, \dots, \tilde{\omega}_{n-1}$, can be used as an estimator of θ :

$$\hat{\theta}_{\tilde{\omega}} = c_{\tilde{\omega}} \sum_{i=1}^{n-1} \tilde{\omega}_i \hat{\theta}^{(i)}. \quad (4)$$

Here $\tilde{\omega}$ in the subscript denotes a set of weights $\tilde{\omega}_i$ ($i = 1, \dots, n-1$), and $c_{\tilde{\omega}} = (\sum_{i=1}^{n-1} \tilde{\omega}_i)^{-1}$ is the corresponding normalisation coefficient. This allows us to re-define the unbiased estimators of θ based on Π and S (see Eq. (1)) to take into account a given null demography as follows

$$\hat{\theta}_\Pi = c_\Pi \Pi = c_\Pi \sum_{i=1}^{n-1} \tilde{\omega}_i^\Pi \hat{\theta}^{(i)}, \quad \text{and} \quad (5)$$

$$\hat{\theta}_S = c_S \frac{S}{a_n} = c_S \sum_{i=1}^{n-1} \tilde{\omega}_i^S \hat{\theta}^{(i)},$$

with the normalisation constants $c_\Pi = (\sum_{i=1}^{n-1} \tilde{\omega}_i^\Pi)^{-1}$, and $c_S = (\sum_{i=1}^{n-1} \tilde{\omega}_i^S)^{-1}$, and weights $\tilde{\omega}_i^\Pi = \omega_i^\Pi i \xi_i^0$, and $\tilde{\omega}_i^S = \omega_i^S i \xi_i^0$. Here ω_i^Π , and ω_i^S are the weights when the null demography corresponds to the constant population size. Using $\hat{\theta}_\Pi$, and $\hat{\theta}_S$ given

in Eq. (5), we define the demography-adjusted Tajima's D as (cf. Zivkovic and Wiehe, 2008)

$$D = \frac{c_{\Pi} \Pi - c_S \frac{S}{a_n}}{\sqrt{\text{Var}\left(c_{\Pi} \Pi - c_S \frac{S}{a_n}\right)}}. \quad (6)$$

When mutations are neutral, the expected value of the numerator of Eq. (6) is equal to zero under a given null demography. We note that demography adjustment of any test based on the site frequency spectrum (e.g. Fu and Li, 1993b, Fay and Wu, 2000, Zeng et al., 2006, Achaz, 2008) can be done in a similar fashion (by first multiplying the weights of the unadjusted estimators with $i\xi_i^0$ and then correctly normalising them). In fact, with Ω_i denoting the difference between the i th normalised weights of the two estimators which define a given test (e.g. for Tajima's D it holds $\Omega_i = c_{\Pi} \tilde{\omega}_i^{\Pi} - c_S \tilde{\omega}_i^S$) the demography-adjusted tests based on the site frequency spectrum, denoted by T_{Ω} below, can be written in vector notation Ferretti et al. (2010, Eq. (12)) as

$$\begin{aligned} T_{\Omega} &= \frac{\sum_{i=1}^{n-1} \Omega_i \hat{\theta}^{(i)}}{\sqrt{\text{Var}\left[\sum_{i=1}^{n-1} \Omega_i \hat{\theta}^{(i)}\right]}} = \frac{\sum_{i=1}^{n-1} \Omega_i \frac{\xi_i}{\xi_i^0}}{\sqrt{\text{Var}\left[\sum_{i=1}^{n-1} \Omega_i \frac{\xi_i}{\xi_i^0}\right]}} \\ &= \frac{\mathbf{\Omega} \cdot \hat{\Theta}}{\sqrt{\text{Var}[\mathbf{\Omega} \cdot \hat{\Theta}]}}. \end{aligned} \quad (7)$$

Here $\mathbf{\Omega} \cdot \hat{\Theta} \equiv \mathbf{\Omega}^T \hat{\Theta}$ denotes the scalar product of the vectors $\mathbf{\Omega} = (\Omega_1, \dots, \Omega_{n-1})^T$, and $\hat{\Theta} = (\hat{\theta}^{(1)}, \dots, \hat{\theta}^{(n-1)})^T$. The denominator in Eq. (7) for constant population size is given by Achaz (2009, his Eq. (9)). For populations of varying size, we obtain (see Appendix A):

$$\text{Var}\left[\sum_{i=1}^{n-1} \Omega_i \hat{\theta}^{(i)}\right] = \theta \sum_{i=1}^{n-1} \frac{\Omega_i^2}{\xi_i^0} + \theta^2 \sum_{i,j=1}^{n-1} \frac{\Omega_i}{\xi_i^0} \sigma_{ij}^0 \frac{\Omega_j}{\xi_j^0}, \quad (8)$$

where $\sigma_{ij}^0 = \text{Cov}(\xi_i, \xi_j)|_{\theta=1}$ for $i \neq j$, and $\sigma_{ii}^0 = (\text{Var}(\xi_i) - \langle \xi_i \rangle)|_{\theta=1}$, as defined by Fu (1995). Note that, according to its definition, σ_{ij}^0 does not depend on θ . In the constant population-size case, it is a function of sample size n (see Fu, 1995), and for a non-constant demography it is a function of n and of the parameters of the demography.

As Eq. (8) shows, estimates of θ and θ^2 are needed to calculate the variance. For populations of constant size, Tajima (1989a) used $\hat{\theta}_S = \frac{1}{\sum_{k=1}^{n-1} \frac{1}{k}} S$. As explained above, for populations of varying size $\hat{\theta}_S$ satisfies $\hat{\theta}_S = c_S \sum_{i=1}^{n-1} \tilde{\omega}_i^S \hat{\theta}^{(i)} = \frac{1}{\sum_{i=1}^{n-1} \xi_i^0} S$. Based on $\hat{\theta}_S$, an unbiased estimator of the second moment of θ is (see Appendix A)

$$\begin{aligned} \hat{\theta}_S^2 &= \frac{\hat{\theta}_S^2 - y_n \hat{\theta}_S}{1 + z_n}, \quad \text{with } y_n = \left(\sum_{i=1}^{n-1} \xi_i^0\right)^{-1}, \\ \text{and } z_n &= \left(\sum_{i,j=1}^{n-1} \sigma_{ij}^0\right) \left(\sum_{i=1}^{n-1} \xi_i^0\right)^{-2}. \end{aligned} \quad (9)$$

For populations of constant size, $\hat{\theta}_S^2$ reduces to Eq. (34) in Tajima (1989a) since, in this case, y_n and z_n are given by

$$y_n = \left(\sum_{i=1}^{n-1} \frac{1}{i}\right)^{-1}, \quad \text{and } z_n = \sum_{i=1}^{n-1} \frac{1}{i^2} \left(\sum_{i=1}^{n-1} \frac{1}{i}\right)^{-2}. \quad (10)$$

It is known that when recombination is neglected, estimation of θ by $\hat{\theta}_S$ in the constant population-size case is efficient (i.e.

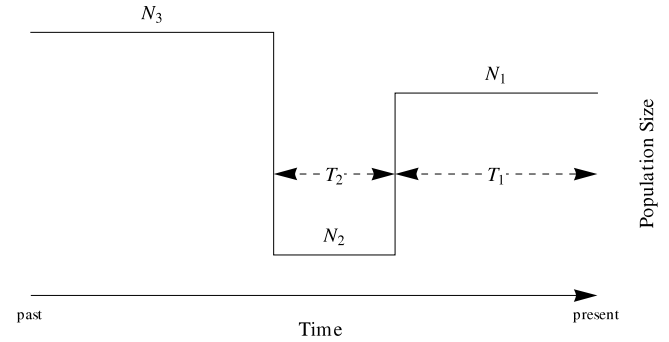


Fig. 1. Demographic model. Present population size is N_1 . In the past, two population-size changes occurred: one at T_1 generations ago from N_1 to N_2 and another one $T_1 + T_2$ generations ago from N_2 to N_3 .

the estimator has minimal variance) for small values of θ (Fu and Li, 1993a). Conversely, for high values of θ (long sequences), the variance of the estimator $\hat{\theta}_S$ decreases as the value of recombination rate along sequences increases (Fu, 1994). One can show that this holds for our extended version of $\hat{\theta}_S$ as well. We note that it is common practise to apply tests, such as Tajima's D , to recombining sequences (Akey et al., 2004; Stajich and Hahn, 2005; Carlson et al., 2005) although in their derivation recombination is neglected.

Our adjusted tests are identical to the unadjusted ones if population size is constant. In this case, expressions for ξ_i^0 and σ_{ij}^0 can be written in closed form and are given by Fu (1995). No such analytical expressions are known in general for varying population sizes. Nawa and Tajima (2008) used computer simulations to assess the first moments of the site frequency spectra under a past population-size expansion, decline, or bottleneck, and Marth et al. (2004) derived a corresponding analytical expression for $\langle \xi_i \rangle$ for piecewise constant demographies. In this study, we use results of Fu (1995) and of Eriksson et al. (2010) (see also Zivkovic and Wiehe, 2008) to compute the second moments under a piecewise constant demography shown in Fig. 1. The details of the computation are given in Appendix B.

2.2. Demographic model

As explained in the introduction, we assume a piecewise constant demography with two population-size changes in the past, because this model was used (Adams and Hudson, 2004; Marth et al., 2004; Stajich and Hahn, 2005) to capture the main events of the human out-of-Africa expansion (Cavalli-Sforza and Feldman, 2003; Ramachandran et al., 2005; Liu et al., 2006; Tanabe et al., 2010; Eriksson et al., 2012). The model is illustrated in Fig. 1. When $N_2 < N_1$ and $N_2 < N_3$ the demography represents a population bottleneck.

In the following we assume a well-mixed random mating diploid population with non-overlapping generations. We also assume that the population size is large so that gene genealogies can be modelled by the standard coalescent (Kingman, 1982). Under this model, there are four unknown parameters to be determined. Upon scaling the parameters of the model (N_1, N_2, N_3, T_1, T_2) by the present population size N_1 , the unknown parameters are the scaled population sizes $x_i = N_i/N_1$ ($i = 2, 3$), and the scaled times t_i ($i = 1, 2$) such that $T_i = \lfloor 2t_i N_1 \rfloor$.

2.3. Estimating demographic parameters using the site frequency spectrum

We use the analytical expressions for the moments of the site frequency spectrum under a given demography to compute

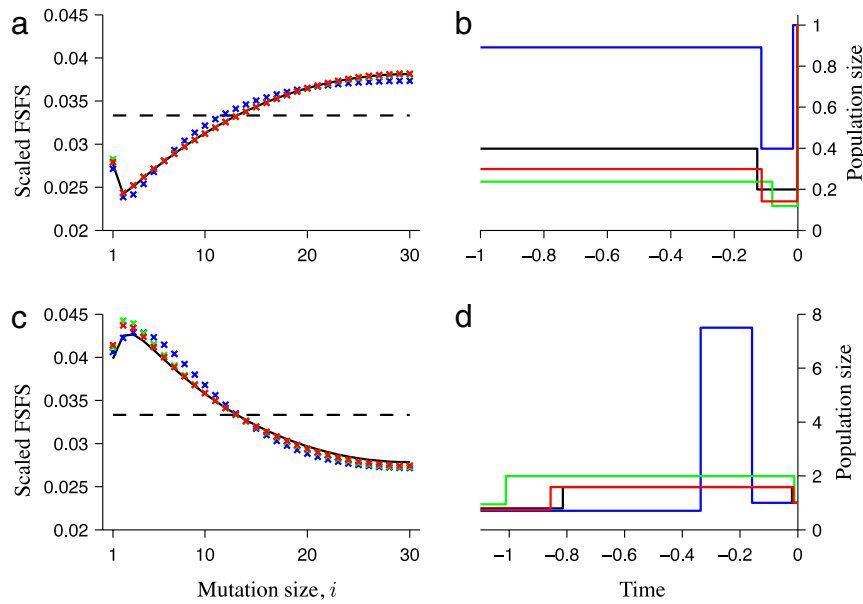


Fig. 2. (a), (c) Scaled folded site frequency spectra computed analytically. The spectra are scaled so that, in the constant population-size case, one obtains a constant equal to $1/[n/2]$ (shown by dashed lines). Analytical spectra corresponding to the actual underlying demographies (shown by black lines in panels b and d, respectively) are shown by black lines. The best-fitted spectra estimated using 10^4 SNPs are shown by blue crosses, green crosses show the best-fitted spectra estimated using 10^5 SNPs, and red crosses show the best-fitted spectra estimated using 10^6 SNPs. (b) Actual underlying demography (black line) for the spectrum shown in a by a black line (recent bottleneck). (d) Actual demography (black line) for the spectrum shown in c by a black line (past population-size expansion, followed by a recent population-size decline). In b and d the maximum likelihood histories estimated using 10^4 SNPs, 10^5 SNPs, and 10^6 SNPs are shown by blue, green, and red lines, respectively. The population size is scaled by N_1 , and the time is scaled by $2N_1$. Sample size used: $n = 60$. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

ML estimates of the parameters of our demographic model. We follow a similar approach as described in Adams and Hudson (2004). We calculate the expected spectrum for a large set of plausible parameters and choose the parameters with highest likelihood, given the data. If SNPs are assumed to be uncorrelated, the spectrum counts ξ_1, \dots, ξ_{n-1} are multinomially distributed (conditional on the total number of SNPs, S), with the parameters given by the expected values of ξ_i (Nielsen, 2000; Adams and Hudson, 2004).

Similarly, the probability to observe the *folded* site frequency spectrum $\eta_1, \dots, \eta_{[n/2]}$ in a sample of $S = \sum_{i=1}^{[n/2]} \eta_i$ polymorphic sites is multinomial with

$$\text{Prob}(\eta_1, \eta_2, \dots, \eta_{[n/2]} | S) = \binom{S}{\eta_1, \eta_2, \dots, \eta_{[n/2]}} \prod_{i=1}^{[n/2]} p_i^{\eta_i}. \quad (11)$$

In this case, the parameters p_i are given by:

$$p_i = \frac{\langle \eta_i \rangle}{\sum_{j=1}^{[n/2]} \langle \eta_j \rangle}. \quad (12)$$

As mentioned in the previous subsection, the expression for $\langle \xi_i \rangle$ (and thus for $\langle \eta_i \rangle$) under the model shown in Fig. 1 is given in Appendix B (see Eqs. (B.4)–(B.6)).

It is known that different demographies can lead to exactly the same spectra (Myers et al., 2008). Hence, cases exist in which it is difficult to distinguish the underlying demographies by their spectra. In order to obtain an estimate for the minimum number of SNPs necessary for reliable inference, we use coalescent simulations to generate spectra under two different demographic histories with two population-size changes in the past (see Fig. 2). These idealised demographies roughly represent the populations CEU and YRI. As an input for the maximum-likelihood parameter estimation, we use folded site frequency spectra (which do not require inference of the ancestral state) because assignment of

the ancestral state via an outgroup can be erroneous, and this can substantially bias demography estimation. We simulate $81 \cdot 10^6$ independent gene genealogies with $n = 60$, and $\theta = 0.01$. For such a small value of θ , genealogies rarely contain more than one mutation. For each demography, we determine three resulting spectra, one containing 10^4 SNPs, one with 10^5 SNPs, and one with 10^6 SNPs (see circles in Fig. S1 in Supplementary material). To obtain the spectra in a way consistent with practical data sampling, we randomly select exactly one SNP from randomly chosen genealogies having mutations.

Using such spectra, we compute the likelihood for our model parameters x_2, x_3, t_1 , and t_2 . The base-10 logarithms of candidate population sizes x_2 , and x_3 are taken from a grid within the interval $[-2, 2]$, and the base-10 logarithms of candidate times t_1 , and t_2 are taken from a grid within the interval $[-3, 0]$ (mesh size 0.025). Thus, for each population we test in total $121^2 \cdot 161^2 \approx 3.8 \times 10^8$ combinations of the four unknown demographic parameters. Note that the ML-estimation does not depend on the parameter θ , as Eq. (12) shows. We also investigate with simulations whether the adjusted Tajima's D can be distorted if inference is based on a (too) small number of SNPs.

2.4. Whole-genome scans with demography-adjusted tests of neutrality

We apply the above ML-procedure to spectra of ten human populations (see Table 1). Data are taken from the 1000 genomes project (McVean et al., 2012), version 3, rel. April 30th, 2012. Variants are filtered by variant type "SNP" (i.e. indels excluded). From each population, four (possibly overlapping) subsamples of 30 individuals are drawn. For demography estimation we use only SNPs from intergenic regions.

As explained above, in order to use the analytical formulae for parameter estimation, SNPs must be uncorrelated, i.e. unlinked. On the other hand, a large amount of SNPs is necessary to render the demography estimation reliable. As a compromise we collect

Table 1

Populations and the corresponding number of individuals sampled.
 Source: Data from the 1000 genomes project (McVean et al., 2012).

	Population	Sample
CEU	CEPH individuals	85
FIN	Finnish in Finland	93
GBR	British from England and Scotland	89
TSI	Toscans in Italia	98
CHB	Han Chinese in Beijing, China	97
CHS	Han Chinese South, China	100
JPT	Japanese in Tokyo, Japan	89
ASW	African ancestry in Southwest USA	61
LWK	Luhya in Webuye, Kenya	97
YRI	Yoruba in Ibadan, Nigeria	88

SNPs in the following way: from each of the 4 subsamples of 30 individuals we draw randomly 10^4 SNPs with the condition that the minimal physical distance between any pair of SNPs is $5 \cdot 10^4$ base pairs (50 kb). This is repeated 10 times for each subsample to obtain in total 40 random spectra, representing $4 \cdot 10^5$ SNPs. We perform the ML-estimation for each population by using the average of these 40 spectra.

The estimated maximum likelihood demographies allow us to obtain demography-adjusted versions of Tajima's D , Fay and Wu's H and Zeng's E . We perform whole-genome scans with both demography-adjusted and unadjusted versions of these tests, using the method of Carlson et al. (2005). We calculate the test statistics in a sliding window of size 100 kb and step size 10 kb. Windows containing less than 5 SNPs are ignored. This way, we collect about 280,000 data points. For the tests of Fay and Wu, and of Zeng it is necessary to know the ancestral allele. This information is obtained through a 6-way alignment of humans and five other primates and is included into the 1000 genomes data. In order to detect putative regions under selection, we determine so-called "contiguous regions of Tajima's D reduction (CRTR)". As in Carlson et al. (2005) we define them as a genomic region of at least 20 consecutive windows, of which at least 75% have a Tajima's D value in the lower 1% quantile.

3. Results

3.1. Test of the maximum likelihood procedure on simulated data

The results of the demography estimation based on simulated data under two reference demographies are shown in Fig. 2. As explained in Section 2, one demography corresponds to a recent bottleneck (black line in panel b) and the other to a past population-size expansion followed by a recent decline (black line in panel d). The corresponding scaled folded spectra computed analytically are shown by black lines in panels a, and c, respectively. The spectra are scaled so that in the constant population-size case one obtains a constant value (independent of i) equal to $1/\lfloor n/2 \rfloor$ (dashed lines in Fig. 2(a), (c)). The demography estimation is based on the folded spectra obtained using coalescent simulations with 10^4 , or 10^5 , or 10^6 SNPs (see blue, green, and red circles in Fig. S1(b), (d) in Supplementary material). As expected, by comparing the actual underlying histories to the estimated ones (see Fig. 2(b), and (d)), we find that by increasing the number of SNPs, the deviation of the parameters corresponding to the maximum likelihood demography from those of the actual demography decreases. In all cases, the parameter with largest deviation from its actual value is t_1 , because both reference demographies assume a very recent population-size change. In particular, the demographies estimated using 10^4 SNPs deviate strongly from the actual ones. Indeed, in this case the estimated time t_1 deviates from its actual value by 650% (Fig. 2(b)), or by 695% (Fig. 2(d)). However, by using 10^5 SNPs, the deviation of

this parameter is drastically reduced to 44% (Fig. 2(b)), or 41% (Fig. 2(d)). Therefore, one can conclude that 10^4 SNPs are not enough for a reliable demography estimation. In contrast, the described procedure works well when the estimation is based on spectra with at least 10^5 SNPs.

In order to further assess the consistency of our demography estimation, we compute the marginal probability distributions of the four unknown parameters under the estimations based on 10^4 , or 10^5 , or 10^6 SNPs (see Figs. S2–S5). These figures show a comparison of the results obtained under the estimation based on folded spectra (panels c and d) with those based on unfolded spectra (panels a and b). As can be seen from the figures, the estimation based on 10^4 SNPs usually results in long-tailed distributions, and as a consequence, the mean value of a given parameter (weighted by the likelihoods of the candidate values) is substantially shifted from its actual value. In addition, the marginal probability distribution in this case usually has a maximum which is substantially shifted from the actual value (or the distribution is bimodal). However, by increasing the number of SNPs the marginal distributions become narrower and their maxima approach the corresponding actual values. Our results suggest that at least about 10^5 SNPs are needed to obtain reliable estimation results. Indeed, as Figs. S2–S5 show, the marginal probability distributions in this case (and for 10^6 SNPs) are substantially narrower (short-tailed, except in Figs. S3(c) and S5(c)) and their mean and maxima are centered sufficiently close to the actual values of the corresponding parameters.

Figs. S2–S5 also show that the marginal probability distributions of the parameters are narrower when the estimation is based on unfolded than on folded spectra, suggesting that the procedure is more stable in the former than in the latter case. However, information about the ancestral state in real data is prone to mis-specifications. An error in the estimation caused due to mis-specifications may be larger than the improvement due to unfolding of the spectra.

3.2. Neutrality tests adjusted to the estimated demographies of simulated data

The demographic parameters estimated using the maximum likelihood procedure serve as an input for computing demography-adjusted tests of neutrality (as described in Section 2). In Fig. 3 we compare the distribution of Tajima's D adjusted to the actual demography, and that adjusted to the estimated demography. It can be seen that the test adjusted to the estimated demography using 10^4 SNPs (blue circles) deviates from the test adjusted to the actual demography (grey region) especially in the tails. By contrast, the quantiles of the test distribution adjusted to the demography estimated with 10^5 SNPs are similar to the quantiles of the distribution of the test adjusted to the actual demography (see also Table 2 which lists the first four moments of the distributions obtained). Note that these results further support our finding that at least 10^5 SNPs are needed for reliable demography estimation and consequently reliable adjustment of the tests. In addition, Table 2 compares the moments of the distributions adjusted to the actual demographies to those of the null distribution of the original Tajima's D (under the constant population size). As this table shows, the first two moments of our adjusted tests are close to those of the null distribution of the original test. However, slight deviations appear in higher moments of the distributions, which was already observed in Zivkovic and Wiehe (2008).

3.3. Estimated human demographies

We applied our maximum likelihood procedure to the human genome data. The spectra used for the estimation contain $4 \cdot 10^5$ SNPs (see Section 2). Our demography estimation shows

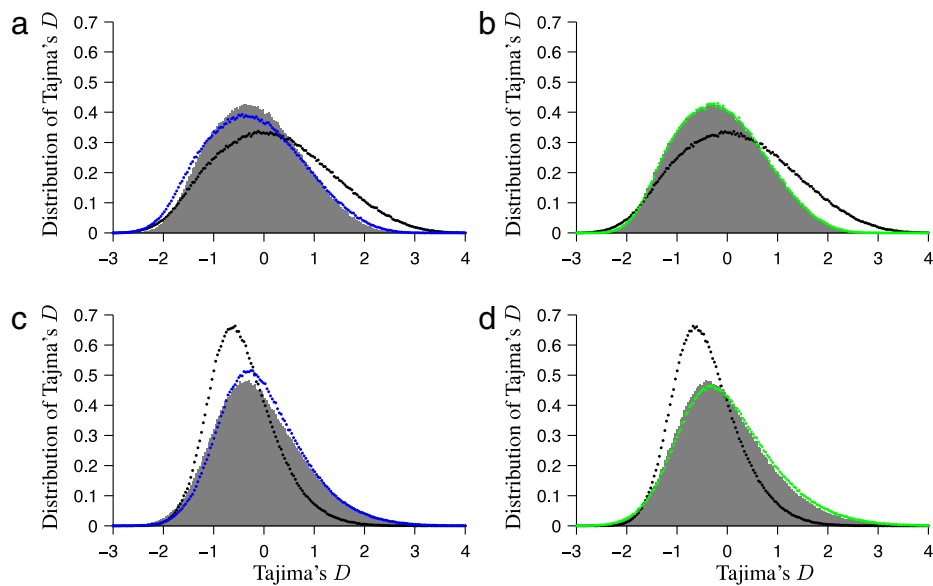


Fig. 3. (a), (b) Numerically computed distributions of Tajima's D for demographic histories shown in Fig. 2(b). Grey region shows the distribution of Tajima's D adjusted to the actual underlying demography, black circles show the unadjusted test, and coloured circles show the test adjusted to the maximum likelihood demographies (for a given number of SNPs). Results of the estimation based on 10^4 SNPs are shown in panel a, and on 10^5 SNPs in panel b. (c)–(d) Same as in panels a, b, respectively, but for demographic histories shown in Fig. 2(d). Remaining parameters used: sample size $n = 60$, scaled mutation rate $\theta = 100$. Number of independent gene genealogies simulated: 10^6 . (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

Table 2

Moments of the curves in Fig. 3, as well as the moments of the null distribution of the original Tajima's D .

	Panel	Mean	Variance	Skewness	Kurtosis
Unadjusted	a, b	0.20	1.25	0.23	2.58
Adjusted to the ML demography	a	-0.18	0.91	0.23	2.57
Adjusted to the actual demography	b	-0.13	0.76	0.23	2.57
Unadjusted	a, b	-0.15	0.76	0.23	2.57
Unadjusted	c, d	-0.41	0.43	0.56	3.46
Adjusted to the ML demography	c	-0.02	0.71	0.56	3.47
Adjusted to the actual demography	d	0.00	0.88	0.56	3.47
Adjusted to the actual demography	c, d	-0.07	0.82	0.57	3.47
Original null		-0.11	0.77	0.40	2.97

(see Fig. 4 and Table S1 in Supplementary material) that the frequency spectra of the non-African populations are consistent with a population bottleneck. By contrast, the spectrum of the African population ASW is consistent with two population-size expansions, and those of LWK and YRI are consistent with an ancestral population-size expansion followed by a recent decline ('inverse bottleneck').

3.4. Neutrality tests adjusted to the estimated human demographies

We show in Fig. 5 (upper panels) genome-wide values of Tajima's D , Fay and Wu's H , and Zeng's E for Europeans (CEU), Asians (CHB) and Africans (YRI) from the 1000 genomes project (McVean et al., 2012). As Fig. 5 (upper panels) shows, the empirical distributions of the tests differ substantially between different populations. The empirical distributions of demography-adjusted tests are, however, similar between different populations (see Fig. 5, lower panels), and hence only demography-adjusted tests make a reasonable comparison between populations possible. This further suggests that most of the differences in the distributions of unadjusted tests are due to the distinct underlying demographies.

Indeed, the inclusion of demography into the tests essentially results in an affine linear transformation of the empirical test values (coefficient of determination $R^2 > 0.999$). Note that the Eq. (8) for a given test depends only (via θ) on the number of segregating sites S . For spectra with the same S unadjusted and

Table 3

Mean values of empirical test distributions shown in Fig. 5.

	Population	Tajima's D	Fay and Wu's H	Zeng's E
Unadjusted tests	CEU	0.26	-0.65	0.87
	CHB	0.38	-0.83	1.14
	YRI	-0.44	-0.10	-0.31
Adjusted tests	CEU	-0.09	-0.31	0.25
	CHB	-0.09	-0.31	0.26
	YRI	-0.12	-0.45	0.29

adjusted test values differ only in the linear weightings of the two θ -estimators (and the different constant in the denominator). The transformation of spectra with vastly different numbers of segregating sites can in principle show deviations from linearity, in particular for small absolute values of S , however these are negligible in comparison with our observed inner-population variance (see below). In the human genome data, we found that the value of θ (and consequently S) per window is relatively large ($\theta > 50$ for almost all windows). The observed scattering appears to concern primarily windows containing very few SNPs which do not yield extreme test values.

Table 3 shows that the mean values of adjusted Tajima's D correspond very well to those of the original test under standard neutrality. The empirical distributions of the other two tests are not centered at zero, due to their sensitivity to high-frequency derived SNPs. These occur in excess, which is a known phenomenon. For

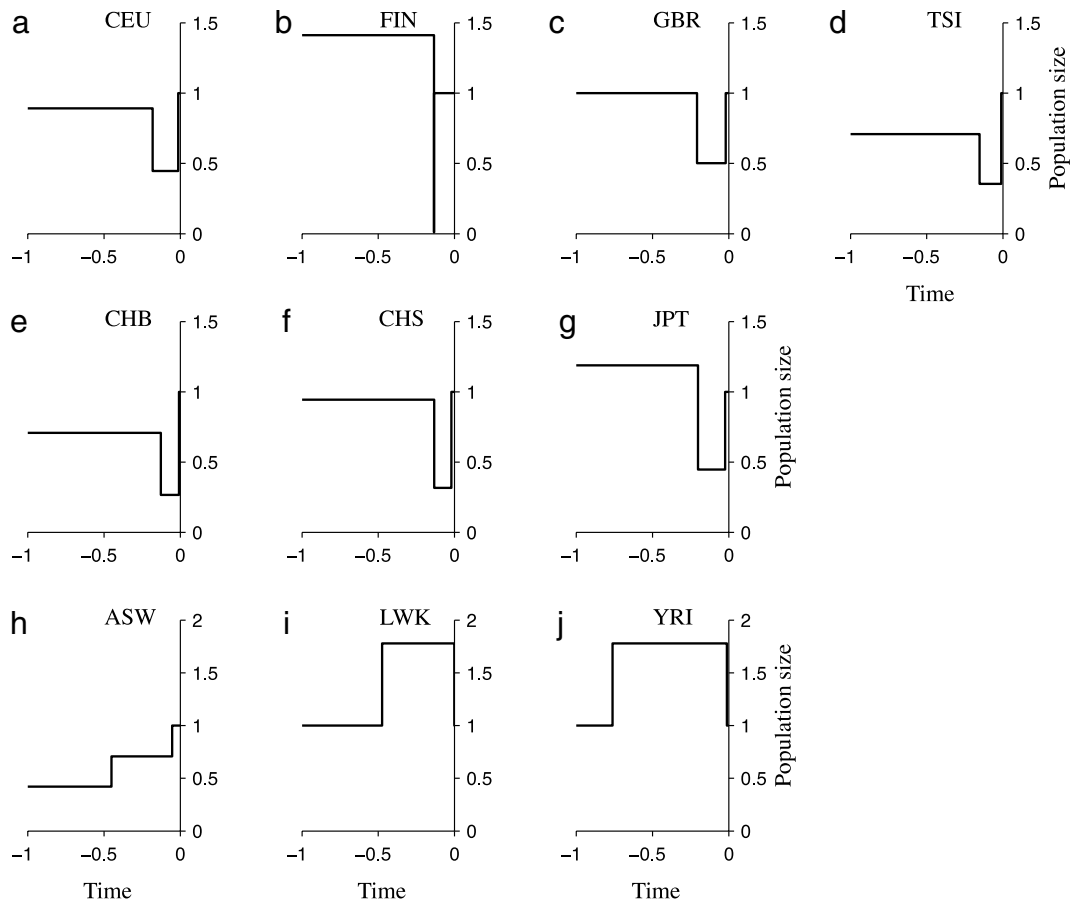


Fig. 4. Estimated demographies for 10 human populations. Note that the demographies of LWK and YRI have identical shape (inverse bottleneck). However, in both cases the population-size decline is so recent, that it cannot be seen on this scale. In each panel, the size is scaled by N_1 , and time is scaled by $2N_1$.

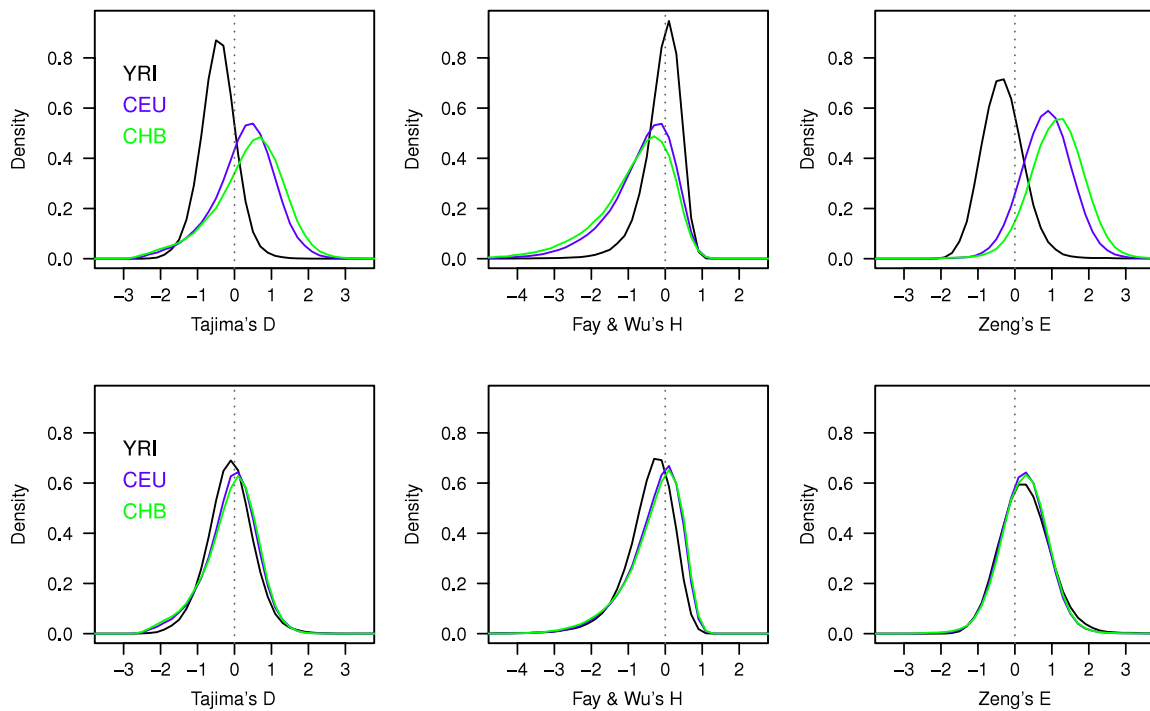


Fig. 5. Distribution of test values over all sliding windows. Top row: unadjusted tests. Bottom row: demography-adjusted tests.

instance, Fig. 2(c) of McVean et al. (2012) shows that derived alleles of very high frequency are found more than four times as often as expected under neutrality. However, the most likely cause is mis-assignment of the state of the ancestral allele. This can have several reasons, e.g. polymorphisms in the outgroup species or recurrent mutations. Hernandez et al. (2007) developed a formula (and program) to correct for the latter effect. Upon applying this procedure, we could explain part of the excess. Another possible source of error is sequencing errors of fixed derived alleles, appearing thus as very high frequency polymorphic sites.

3.5. Identifying candidate regions of positive selection

We compared Tajima's D between the different subsamples of the same population and obtained a coefficient of determination of $R^2 \approx 0.8$ in all populations. For subsamples from different populations, the highest correlation shows CHB with CHS ($R^2 \approx 0.73$), and CEU with GBR ($R^2 \approx 0.71$). The lowest correlation shows LWK or YRI compared with the Asian populations ($R^2 \approx 0.1$). Note that the adjusted and unadjusted tests yield essentially the same correlations, because they are linearly related (see above).

Concerning the contiguous regions of Tajima's D reduction, we find that they vary considerably among subsamples of the same population. We therefore add a condition and require the test statistic of a particular window to be in the 1%-quantile in each of the four subsamples. From these windows we construct CRTRs as described above. The additional constraint reduces the number of CRTRs by more than 50%. For the populations CEU, CHB and YRI the obtained regions are depicted in Fig. 6. We find 7 CRTRs for population CEU, 10 for CHB and 8 for YRI, respectively. This differs from the results by Carlson et al. (2005). Using the SNP array data available at that time, they found 7 CRTRs for the African, 23 for the European and 29 for the Chinese population samples, which only partially overlap with ours. These differences are caused most likely by the distinct population samples and by the more exhaustive SNP set which we used. In the supplement we list CRTRs of all 10 populations analysed in the current study. The program used to calculate the adjusted test statistics is available as C++ source code on <http://ntx.sourceforge.net/> and tracks for the UCSC browser containing test values (unadjusted as well as adjusted) for all ten populations are available at <http://jakob.genetik.uni-koeln.de/data/>.

4. Discussion and conclusions

It is common practise to use quantiles of empirical whole-genome distributions of neutrality tests to detect regions under selection. However, tests are usually defined using constant population size as a null assumption, and it is presumed (but not tested) that empirical whole-genome distributions are mainly shaped by the underlying demography. Moreover, since the variances of empirical distributions are strongly affected by the underlying demography of an analysed population, it is very difficult to quantify and compare deviations from neutrality at a given genome region between different populations. In order to solve these issues, we defined in this study demography-adjusted tests of neutrality by directly integrating the effects of the actual (or, in practise, estimated) demography into SFS-based tests. A necessary step towards defining demography-adjusted tests is to compute the first two moments of the SFS under the estimated demography. In this study we derived exact analytical expressions for these moments under a demographic model allowing for two population-size changes, by combining the results of Fu (1995)

with those of Eriksson et al. (2010). Such a model is believed to capture the essence (Adams and Hudson, 2004; Marth et al., 2004; Stajich and Hahn, 2005) of the out-of-Africa expansion of humans (Cavalli-Sforza and Feldman, 2003; Ramachandran et al., 2005; Liu et al., 2006; Tanabe et al., 2010; Eriksson et al., 2012). Note that our expressions for the first two moments of the SFS are also helpful to find optimal tests of neutrality under piecewise constant demographies (Ferretti et al., 2010). For populations of constant size, our 'adjusted' tests are identical to the original (unadjusted) ones. Our procedure generalises previous results regarding demography-adjustment of Tajima's D (Zivkovic and Wiehe, 2008).

In order to estimate the demography of a given population, we applied a ML procedure to single nucleotide polymorphisms (SNPs) sampled at physically distant sites, which are largely independent from each other, as proposed by Nielsen (2000). Because of the independence of SNPs used for demography estimation, the bins of the SFS are populated according to a multinomial distribution, which simplifies the mathematical treatment. To test how sensitive ML-estimates are with respect to the number of SNPs used for estimation, we performed a series of computer experiments. We fitted folded site frequency spectra simulated under two reference demographies, one a recent bottleneck, and the other a past population-size expansion followed by a recent decline. As expected, we found that ML-estimation of demography is consistent: the estimated parameters converge to those of the true demography with increasing number of SNPs. The spectrum corresponding to the ML-demography is almost indistinguishable from the spectrum corresponding to the actual underlying demography if the estimation is based on more than 100,000 SNPs. We confirmed this finding for our two reference demographies by comparing Tajima's D adjusted to the actual underlying demography with that adjusted to the ML-demography.

After confirming the validity of the ML-procedure, we applied our demography-adjustment procedure to data from the 1000 genomes project (McVean et al., 2012). We sampled the folded frequency spectra of ten human populations from physically distant, presumably neutral (Adams and Hudson, 2004), intergenic regions in order to estimate the ML-parameters of the piecewise constant demographic model allowing for population size parameters to change by at most two orders of magnitude (Marth et al., 2004). The time parameters were allowed to vary by three orders of magnitude (i.e. from -3 to 0 on logarithmic scale). The lower bound for the times corresponds to 10 generations (about 200–250 years). This is about the minimum time to leave an imprint on the frequency spectrum. The upper bound for the times was chosen to correspond to the emergence of anatomically modern humans about 200,000 years ago (Cavalli-Sforza and Feldman, 2003).

Our results are consistent with the results of Adams and Hudson (2004) and of Marth et al. (2004): the ML-demographies of non-African populations correspond to a bottleneck, and the ML-demography of one of the sampled African populations (ASW) corresponds to two subsequent population-size expansions. The spectra of the remaining two African populations (LWK and YRI) gave rise to demographies with a distant population-size expansion followed by a population-size decline.

Myers et al. (2008) argued that inference of demography from the frequency spectrum may not be possible at all, because very different demographies can lead to the same frequency spectrum. Still, we found the ML-parameter estimation to be consistent for our simple demographic model, albeit sensitive to small changes in the frequency spectrum. Notably, our calculations in Appendix B show that the first two moments of the SFS under a bottleneck depend both on the duration and the size of the bottleneck, and this dependence cannot be expressed in terms of a single parameter (duration divided by the size of the bottleneck).



Fig. 6. Contiguous regions of Tajima's D reduction ("CRTR") from Carlson et al. (2005) compared with those derived from our demography-adjusted test. From above to beneath: Carlson: African descent (grey); ASW (grey) and YRI (black); Carlson: European-descent (blue); CEU; Carlson: Chinese-descent (green); CHB. The regions found by Carlson et al. have been translated from hg17 to hg19 coordinates. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

In order to detect regions under selection, we performed a genome screen with three tests of neutrality. We found that the empirical distributions of the adjusted tests are very similar to each other, suggesting that the differences between the unadjusted empirical distributions are mainly caused by the different demographies. The linearity of the transformation causes the empirical quantiles of the adjusted tests to be shifts of the unadjusted ones. Consequently, the candidate regions for selection do not change.

Neither unadjusted nor adjusted test statistics take recombination into account. It is well-known that recombination decreases the variances of test distributions (Tajima, 1989a), but it is unclear how heterogeneity of scaled recombination rates among genomic regions and populations affects the linear relationship between demography-adjusted and unadjusted tests.

When we compared our Tajima's D values with the ones calculated from SNP array data by Carlson et al. (2005), we found

only modest correlation (data not shown). As a consequence, the candidate regions of selection show little overlap. One reason lies in the different population samples. Another reason is the low robustness of CRTRs, as defined by Carlson et al. (2005): ‘long stretches of low Tajima’s D ’ are easily disrupted if only a few measurements within the stretch change. As a slight modification of Carlson et al.’s definition, we required windows to belong to the respective lower 1%-quantile in several subsamples of the same population. This reduced considerably the number of candidate regions, but made them more robust.

A somewhat complementary approach to allow for demography in tests on the frequency spectrum was recently taken by Ronen et al. (2013). They performed simulations of selective sweeps including (a given) demography, and using machine learning methods they generated weights (a coarser version of our Ω_i ’s) to find a test with optimal power for the given evolutionary scenario. In contrast, our strategy is to adjust existing tests by analytically integrating demographic effects.

In conclusion, the demography-adjusted tests introduced here serve as a basis for disentangling the effects of selection from those of demography, and they facilitate a direct comparison between populations with different demographies. It is, however, not yet clear if inhomogeneity of recombination rates along the genome affects differently the distributions of adjusted and unadjusted tests. This remains to be answered in future work.

Acknowledgments

This work was financially supported by grants from Vetenskaprådet, from the Göran Gustafsson Foundation for Research in Natural Sciences and Medicine, through the Linnaeus Centre for Marine Evolutionary Biology (CeMEB, www.cemeb.science.gu.se) to BM, by a grant of the German Science Foundation (DFG-SFB680) to TW, and AE by the Leverhume Trust and the Biotechnology and Biological Sciences Research Council (Grant BB/H005854/1).

Appendix A. The denominator of Eq. (7)

The numerator of Eq. (7) depends on the first moment of the spectrum under a given demography. The denominator of Eq. (7) depends on the second moment. We find:

$$\begin{aligned} \text{Var}\left[\sum_{i=1}^{n-1} \Omega_i \hat{\theta}^{(i)}\right] &= \text{Var}\left[\sum_{i=1}^{n-1} \Omega_i \frac{\xi_i}{\xi_i^0}\right] \\ &= \sum_{i=1}^{n-1} \text{Var}\left(\Omega_i \frac{\xi_i}{\xi_i^0}\right) + \sum_{\substack{i,j=1 \\ i \neq j}}^{n-1} \text{Cov}\left(\Omega_i \frac{\xi_i}{\xi_i^0}, \Omega_j \frac{\xi_j}{\xi_j^0}\right) \\ &= \sum_{i=1}^{n-1} \left(\frac{\Omega_i}{\xi_i^0}\right)^2 \text{Var}(\xi_i) + \sum_{\substack{i,j=1 \\ i \neq j}}^{n-1} \frac{\Omega_i}{\xi_i^0} \text{Cov}(\xi_i, \xi_j) \frac{\Omega_j}{\xi_j^0} \\ &= \sum_{i=1}^{n-1} \left(\frac{\Omega_i}{\xi_i^0}\right)^2 (\theta \xi_i^0 + \theta^2 \sigma_{ii}^0) + \theta^2 \sum_{\substack{i,j=1 \\ i \neq j}}^{n-1} \frac{\Omega_i}{\xi_i^0} \sigma_{ij}^0 \frac{\Omega_j}{\xi_j^0} \\ &= \theta \sum_{i=1}^{n-1} \Omega_i^2 \frac{1}{\xi_i^0} + \theta^2 \sum_{i,j=1}^{n-1} \frac{\Omega_i}{\xi_i^0} \sigma_{ij}^0 \frac{\Omega_j}{\xi_j^0}. \end{aligned} \quad (\text{A.1})$$

Here one has $\sigma_{ij}^0 = \text{Cov}(\xi_i, \xi_j)|_{\theta=1}$, for $i \neq j$, and $\sigma_{ii}^0 = (\text{Var}(\xi_i) - \langle \xi_i \rangle)|_{\theta=1}$. Eq. (A.1) corresponds to Eq. (8) given in the main text. Note that for the constant population size one has $\xi_i^0 = 1/i$, and σ_{ij}^0 is given by Fu (1995). Thus, Eq. (A.1) reduces to Eq. (9) in Achaz (2009).

To evaluate Eq. (A.1) using the observed spectrum, one needs an estimate of θ^2 . If $\hat{\theta}_\omega = \sum_{i=1}^{n-1} \omega_i \xi_i / \xi_i^0$, then

$$\begin{aligned} \langle \hat{\theta}_\omega^2 \rangle &= \text{Var}[\hat{\theta}_\omega] + \langle \hat{\theta}_\omega \rangle^2 = \theta \sum_{i=1}^{n-1} \frac{\omega_i^2}{\xi_i^0} + \theta^2 \sum_{i,j=1}^{n-1} \frac{\omega_i}{\xi_i^0} \sigma_{ij}^0 \frac{\omega_j}{\xi_j^0} + \theta^2 \\ &= y_n \theta + (1 + z_n) \theta^2, \end{aligned}$$

with

$$y_n = \sum_{i=1}^{n-1} \frac{\omega_i^2}{\xi_i^0} \quad \text{and} \quad z_n = \sum_{i,j=1}^{n-1} \frac{\omega_i}{\xi_i^0} \sigma_{ij}^0 \frac{\omega_j}{\xi_j^0}.$$

It follows that

$$\langle \hat{\theta}_\omega^2 \rangle - y_n \langle \hat{\theta}_\omega \rangle = \theta^2 (1 + z_n).$$

Solving the latter with respect to θ^2 yields:

$$\theta^2 = \frac{\langle \hat{\theta}_\omega^2 \rangle - y_n \langle \hat{\theta}_\omega \rangle}{1 + z_n}.$$

Hence, as an estimator for θ^2 we take

$$\hat{\theta}_\omega^2 = \frac{\hat{\theta}_\omega^2 - y_n \hat{\theta}_\omega}{1 + z_n}.$$

This expression corresponds to Eq. (9).

Appendix B. The first two moments of the site frequency spectrum

Now, we compute the first two moments of the spectrum, $\langle \xi_i \rangle$ and $\langle \xi_i \xi_j \rangle$. We consider a randomly mating diploid population with varying population size and the infinite sites model with mutation rate μ per generation per site. The scaled mutation rate per sequence of length L is given by $\theta = 4\mu N_1 L$, where N_1 denotes the present population size. We consider the spectrum for gene genealogies of n individuals. Upon scaling time in units of $2N_1$ generations, we denote by τ_k the time interval during which gene genealogies have exactly $k \leq n$ lineages.

The first two moments of the spectrum can then be expressed as (Fu, 1995)

$$\langle \xi_i \rangle = \frac{\theta}{2} \sum_{k=2}^n kp(k, i) \langle \tau_k \rangle, \quad (\text{B.1})$$

$$\begin{aligned} \langle \xi_i \xi_j \rangle &= \delta_{i,j} \sum_{k=2}^n kp(k, i) \left(\frac{\theta}{2} \langle \tau_k \rangle + \frac{\theta^2}{4} \langle \tau_k^2 \rangle \right) \\ &\quad + \frac{\theta^2}{4} \left\{ \sum_{k=2}^n k(k-1) p(k, i; k, j) \langle \tau_k^2 \rangle \right. \\ &\quad \left. + \sum_{k < m} km (p(k, i; m, j) + p(k, j; m, i)) \langle \tau_k \tau_m \rangle \right\}, \end{aligned} \quad (\text{B.2})$$

where

$$\delta_{i,j} = \begin{cases} 1, & \text{for } i = j, \\ 0, & \text{for } i \neq j, \end{cases}$$

$$p(k, i) = \frac{\binom{n-k}{i-1} k - 1}{\binom{n-1}{i}},$$

$$p(k, i; k, j) = \begin{cases} \frac{\binom{n-i-j-1}{k-3}}{\binom{n-1}{k-1}}, & \text{for } k > 2, \\ p(k, i), & \text{for } k = 2, \text{ and } i + j = n, \\ 0, & \text{for } k = 2, \text{ and } i + j \neq n, \text{ and} \end{cases}$$

$$p(k, i; m, j) = (\delta_{\lfloor i/j \rfloor, 0} + \delta_{i,j}) p_a(k, i; m, j) + (\delta_{\lfloor (i+j)/n \rfloor, 0} + \delta_{i+j, n}) p_b(k, i; m, j). \quad (\text{B.3})$$

The probabilities $p_a(k, i, m, j)$, and $p_b(k, i, m, j)$ in Eq. (B.3) are given by Fu (1995)

$$p_a(k, i, m, j) = \begin{cases} \sum_{t=2}^{\min(m-k+1, i-j+1)} \frac{\binom{m-k}{t-1}}{\binom{m-1}{t}} \frac{k-1}{m} \frac{\binom{i-j-1}{t-2} \binom{n-i-1}{m-t-1}}{\binom{n-1}{m-1}}, & \text{for } j < i \\ \frac{k-1}{m(m-1)} \frac{\binom{n-i-1}{m-2}}{\binom{n-1}{m-1}}, & \text{for } i = j, \end{cases}$$

$$p_b(k, i, m, j) = \begin{cases} \sum_{t=1}^{\min(m-2, m-k+1, i)} \frac{\binom{m-k}{t-1}}{\binom{m-1}{t}} \frac{(k-1)(m-t)}{tm} \frac{\binom{i-1}{t-1} \binom{n-i-j-1}{m-t-2}}{\binom{n-1}{m-1}}, & \text{for } k > 2 \\ \frac{1}{jm} \frac{\binom{n-m}{j-1}}{\binom{n-1}{j}}, & \text{for } k = 2, \text{ and } i + j = n. \end{cases}$$

In the limit $\theta \rightarrow 0$, Eq. (B.2) reduces to:

$$\langle \xi_i^2 \rangle = \frac{\theta}{2} \langle \xi_i \rangle, \quad \text{and} \quad \langle \xi_i \xi_{j \neq i} \rangle = 0 \quad \text{for } \theta \rightarrow 0.$$

In other words, in this limit the spectrum counts are multinomially distributed, as explained in Section 2.

For constant population size, it follows from Eq. (B.1) that $i \langle \xi_i \rangle = \theta$, independently of i . In contrast, for the demographic history shown in Fig. 1, this is not true. Using the results of Eriksson et al. (2010), in this case we find:

$$\langle \xi_i \rangle = \frac{\theta}{2} \sum_{m_1=2}^n a_{m_1}^{(ni)} f_{m_1}, \quad \text{for } i = 1, \dots, n-1, \quad (\text{B.4})$$

where $a_{m_1}^{(ni)}$, and f_{m_1} are:

$$a_{m_1}^{(ni)} = \sum_{k=2}^{m_1} k c_{nkm_1} p(k, i), \quad (\text{B.5})$$

$$f_{m_1} = b_{m_1}^{-1} \left(1 - (1 - x_2) e^{-b_{m_1} t_1} + (x_3 - x_2) e^{-b_{m_1} t_1} e^{-b_{m_1} s_2} \right). \quad (\text{B.6})$$

Here, $x_2 = N_2/N_1$, $x_3 = N_3/N_1$, $s_2 = t_2/x_2$, $b_{m_1} = \binom{m_1}{2}$, and c_{nkm_1} is given by Eq. (11) in Eriksson et al. (2010). This result is consistent with Eq. (1) in Marth et al. (2004), assuming $M = 3$ in the model of Marth et al. (2004).

In what follows, we list our results for $\langle \xi_i \xi_j \rangle$ under the demographic history shown in Fig. 1. We find:

$$\langle \xi_i \xi_j \rangle = \delta_{i,j} \left(\langle \xi_i \rangle + \frac{\theta^2}{4} \sum_{m_1=2}^n \sum_{k=2}^{m_1} a_{m_1 k}^{(nij)} f_{m_1 k} \right) + \frac{\theta^2}{4} \sum_{m_1=2}^n \left(\sum_{k=2}^{m_1} g_{m_1 k}^{(nij)} f_{m_1 k} + \sum_{m_2=2}^{m_1} h_{m_1 m_2}^{(nij)} f_{m_1 m_2} \right), \quad (\text{B.7})$$

where

$$a_{m_1 k}^{(nij)} = 2k c_{nkm_1} c_{kkk} p(k, i), \quad (\text{B.8})$$

$$g_{m_1 k}^{(nij)} = 2k(k-1) c_{nkm_1} c_{kkk} p(k, i; k, j) \quad (\text{B.9})$$

$$h_{m_1 m_2}^{(nij)} = \sum_{l=m_2}^{m_1} l c_{nlm_1} \sum_{k=2}^{m_2} k c_{lkm_2} [p(k, i; l, j) + p(k, j; l, i)]. \quad (\text{B.10})$$

For the terms f_{m_1, m_2} in Eq. (B.7), we consider separately the cases $m_1 \neq m_2$, and $m_1 = m_2$. For the case $m_1 \neq m_2$, we find

$$f_{m_1 m_2} = \frac{1}{b_{m_2}} \left\{ \frac{1 - e^{-b_{m_1} t_1} [1 - x_2^2 + (x_2^2 - x_3^2) e^{-b_{m_1} s_2}]}{b_{m_1}} - [1 - x_2 + (x_2 - x_3) e^{-b_{m_2} s_2}] \frac{e^{-b_{m_2} t_1} - e^{-b_{m_1} t_1}}{b_{m_1} - b_{m_2}} + x_2(x_3 - x_2) e^{-b_{m_1} t_1} \frac{e^{-b_{m_2} s_2} - e^{-b_{m_1} s_2}}{b_{m_1} - b_{m_2}} \right\}. \quad (\text{B.11})$$

For the case $m_1 = m_2$, we obtain:

$$f_{m_1 m_1} = \frac{1}{b_{m_1}} \left\{ \frac{1 - e^{-b_{m_1} t_1} [1 - x_2^2 + (x_2^2 - x_3^2) e^{-b_{m_1} s_2}]}{b_{m_1}} - t_1 [1 - x_2 + (x_2 - x_3) e^{-b_{m_2} s_2}] e^{-b_{m_1} t_1} + x_2(x_3 - x_2) s_2 e^{-b_{m_1} t_1} e^{-b_{m_1} s_2} \right\}. \quad (\text{B.12})$$

Eqs. (B.4)–(B.6) are used to find the demographic parameters that correspond to empirical data in terms of the maximum likelihood approach. Eqs. (B.7)–(B.12) are used to compute the tests of neutrality under the estimated demographics. The results are shown in Results.

Appendix C. Supplementary material

Supplementary material related to this article can be found online at <http://dx.doi.org/10.1016/j.tpb.2014.05.002>.

References

Achaz, G., 2008. Testing for neutrality in samples with sequencing errors. *Genetics* 179 (3).

Achaz, G., 2009. Frequency spectrum neutrality tests: one for all and all for one. *Genetics* 183 (1), 249–258.

Adams, A.A., Hudson, R.R., 2004. Maximum-likelihood estimation of demographic parameters using the frequency spectrum of unlinked single-nucleotide polymorphisms. *Genetics* 168, 1699–1712.

Akey, J.M., Eberle, M.A., Rieder, M.J., Carlson, C.S., Shriver, M.D., Nickerson, D.A., Kruglyak, L., 2004. Population history and natural selection shape patterns of genetic variation in 132 genes. *PLoS Biol.* 2 (10), e286.

Carlson, C.S., Thomas, D.J., Eberle, M.A., Swanson, J.E., Livingston, R.J., Rieder, M.J., Nickerson, D.A., 2005. Genomic regions exhibiting positive selection identified from dense genotype data. *Genome Res.* 15, 1553–1565.

Cavalli-Sforza, L.L., Feldman, M.W., 2003. The application of molecular genetic approaches to the study of human evolution. *Nat. Genet. (Suppl.)* 33, 266–275.

Eriksson, A., Betti, L., Friend, A.D., Lycett, S.J., Singarayer, J.S., von Cramon-Taubadel, N., Valdes, P.J., Balloux, F., Manica, A., 2012. Late Pleistocene climate change and the global expansion of anatomically modern humans. *Proc. Natl. Acad. Sci.* 109 (40), 16089–16094.

Eriksson, A., Mehlig, B., Rafajlović, M., Sagitov, S., 2010. The total branch length of sample genealogies in populations of variable size. *Genetics* 186 (2), 601–611.

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

Ferretti, L., Perez-Enciso, M., Ramos-Onsins, S., 2010. Optimal neutrality tests based on the frequency spectrum. *Genetics* 186 (1), 353–365.

Fu, Y.-X., 1994. Estimating effective population size or mutation rate using the frequencies of mutations of various classes in a sample of DNA sequences. *Genetics* 138, 1375–1386.

Fu, Y.X., 1995. Statistical properties of segregating sites. *Theor. Popul. Biol.* 48 (2), 172–197.

Fu, Y.-X., Li, W.-H., 1993a. Maximum likelihood estimation of population parameters. *Genetics* 134 (4), 1261–1270.

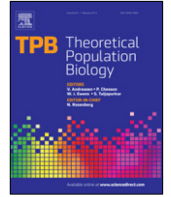
Fu, Y.X., Li, W.H., 1993b. Statistical tests of neutrality of mutations. *Genetics* 133 (3), 693–709.

- Grossman, S.R., Andersen, K.G., Shlyakhter, I., Tabrizi, S., Winnicki, S., Yen, A., Park, D.J., Griesemer, D., Karlsson, E.K., Wong, S.H., Cabili, M., Adegbola, R.A., Bamezai, R.N., Hill, A.V., Vannberg, F.O., Rinn, J.L., Lander, E.S., Schaffner, S.F., Sabeti, P.C., 2013. Identifying recent adaptations in large-scale genomic data. *Cell* 152 (4), 703–713.
- Hernandez, R.D., Williamson, S.H., Bustamante, C.D., 2007. Context dependence, ancestral misidentification, and spurious signatures of natural selection. *Mol. Biol. Evol.* 24 (8), 1792–1800.
- Kingman, J.F.C., 1982. The coalescent. *Stochastic Process. Appl.* 13 (3), 235–248.
- Liu, H., Prugnolle, F., Manica, A., Balloux, F., 2006. A geographically explicit genetic model of worldwide human-settlement history. *Am. J. Hum. Genet.* 79 (2), 230–237.
- Marth, G.T., Czabarka, E., Murvai, J., Sherry, S.T., 2004. The allele frequency spectrum in genome-wide human variation data reveals signals of differential demographic history in three large world populations. *Genetics* 166 (1), 351–372.
- McVean, , et al., 2012. An integrated map of genetic variation from 1092 human genomes. *Nature* 491, 56–65.
- Myers, S., Fefferman, C., Patterson, N., 2008. Can one learn history from the allelic spectrum? *Theor. Popul. Biol.* 73, 342–348.
- Nawa, N., Tajima, F., 2008. Simple method for analyzing the pattern of DNA polymorphism and its application to SNP data of human. *Genes Genet. Syst.* 83 (4), 353–360.
- Nielsen, R., 2000. Estimation of population parameters and recombination rates from single nucleotide polymorphisms. *Genetics* 154, 931–942.
- Nielsen, R., Williamson, S., Kim, Y., Hubisz, M.J., Clark, A.G., Bustamante, C., 2005. Genomic scans for selective sweeps using SNP data. *Genome Res.* 15, 1566–1575.
- Ramachandran, S., Deshpande, O., Roseman, C., Rosenberg, N., Feldman, M., Cavalli-Sforza, L., 2005. Support from the relationship of genetic and geographic distance in human populations for a serial founder effect originating in Africa. *Proc. Natl. Acad. Sci. USA* 102 (44), 15942–15947.
- Ronen, R., Udpa, N., Halperin, E., Bafna, V., 2013. Learning natural selection from the site frequency spectrum. *Genetics* 195 (1), 181–193.
- Sabeti, P.C., Reich, D.E., Higgins, J.M., Levine, H.Z.P., Richter, D.J., Schaffner, S.F., Gabriel, S.B., Platko, J.V., Patterson, N.J., McDonald, G.J., Ackerman, H.C., Campbell, S.J., Altshuler, D., Cooper, R., Kwiatkowski, D., Ward, R., Lander, E.S., 2002. Detecting recent positive selection in the human genome from haplotype structure. *Nature* 419, 832–837.
- Stajich, J.E., Hahn, M.W., 2005. Disentangling the effects of demography and selection in human history. *Mol. Biol. Evol.* 22 (1), 63–73.
- Tajima, F., 1989a. Statistical method for testing the neutral mutation hypothesis by DNA polymorphism. *Genetics* 123 (3), 585–595.
- Tanabe, K., Mita, T., Jombart, T., Eriksson, A., Horibe, S., Palacpac, N., Ranford-Cartwright, L., Sawai, H., Sakihama, N., Ohmae, H., Nakamura, M., Ferreira, M.U., Escalante, A.A., Prugnolle, F., Björkman, A., Färnert, A., Kaneko, A., Horii, T., Manica, A., Kishino, H., Balloux, F., 2010. *Plasmodium falciparum* accompanied the human expansion out of Africa. *Curr. Biol.* 20 (14), 1283–1289.
- Voight, B.F., Kudaravalli, S., Wen, X., Pritchard, J.K., 2006. A map of recent positive selection in the human genome. *PLoS Biol.* 4 (3), e72.
- Zeng, K., Fu, Y.-X., Shi, S., Wu, C.-I., 2006. Statistical tests for detecting positive selection by utilizing high-frequency variants. *Genetics* 174, 1431–1439.
- Zivkovic, D., Wiehe, T., 2008. Second-order moments of segregating sites under variable population size. *Genetics* 180 (1), 341–357.



Contents lists available at ScienceDirect

Theoretical Population Biology

journal homepage: www.elsevier.com/locate/tpb

The neutral frequency spectrum of linked sites

Luca Ferretti^{a,b,*}, Alexander Klassmann^{c,1}, Emanuele Raineri^d,
Sebastián E. Ramos-Onsins^e, Thomas Wiehe^c, Guillaume Achaz^b^a The Pirbright Institute, Woking, United Kingdom^b Institut de Systématique, Evolution, Biodiversité, UMR 7205, MNHN and Centre Interdisciplinaire de Recherche en Biologie, UMR 7241, Collège de France, Paris, France^c Institut für Genetik, Universität zu Köln, Köln, Germany^d CNAG-CRG, Centre for Genomic Regulation (CRG) and UPF, Barcelona, Spain^e Centre for Research in Agricultural Genomics (CRAG), Bellaterra, Barcelona, Spain

HIGHLIGHTS

- The evolution of a non-recombining locus affects the patterns of pairs of polymorphisms.
- Coalescent theory describes the neutral frequency spectrum of pairs of sites.
- The authors introduce a new frequency spectrum of sites linked to a focal mutation.
- Neutral expressions are provided for a sample of individuals and for a whole population.
- This provides a null model for loci containing neutral genetic markers.

ARTICLE INFO

Article history:

Received 29 November 2017

Available online xxxx

Keywords:

Site frequency spectrum

Linkage disequilibrium

Neutral evolution

Kingman coalescent

ABSTRACT

We introduce the conditional Site Frequency Spectrum (SFS) for a genomic region linked to a focal mutation of known frequency. An exact expression for its expected value is provided for the neutral model without recombination. Its relation with the expected SFS for two sites, 2-SFS, is discussed. These spectra derive from the coalescent approach of Fu (1995) for finite samples, which is reviewed. Remarkably simple expressions are obtained for the linked SFS of a large population, which are also solutions of the multi-allelic Kolmogorov equations. These formulae are the immediate extensions of the well known single site θ/f neutral SFS. Besides the general interest in these spectra, they relate to relevant biological cases, such as structural variants and introgressions. As an application, a recipe to adapt Tajima's D and other SFS-based neutrality tests to a non-recombining region containing a neutral marker is presented.

© 2018 Elsevier Inc. All rights reserved.

1. Introduction

One of the basic features that characterizes nucleotide polymorphisms is the Site Frequency Spectrum (SFS), that is the distribution of the mutation frequencies at each site. The SFS can be computed either for the whole (large) population, assuming that the frequency f is a continuous value in $(0, 1)$ or for a sample of n individuals, for which the frequency is a discrete variable $f = k/n$, where $k \in [1, n - 1]$. Sites with alleles at frequency 0 or 1 are not included in the SFS.

According to the standard neutral model of molecular evolution (Kimura, 1983), polymorphisms segregating in a population eventually reach a mutation–drift equilibrium. In this model, the

expected neutral spectrum is proportional to the inverse of the frequency (Wright, 1938; Ewens, 2012). Using coalescent theory, Fu (1995) derived the mean and covariance matrix for each component of the sample SFS by averaging coalescent tree realizations across the whole tree space. For a single realization of the coalescent tree, results are different and depend on the realization; for example, mutations of high frequencies can be present only for highly unbalanced genealogies (Ferretti et al., 2017). The SFS was also studied in scenarios including selection (Fay and Wu, 2000; Kim and Stephan, 2002), demography (Griffiths and Tavaré, 1994; Živković and Wiehe, 2008) or population structure (Alcala et al., 2016).

Besides its general interest, the SFS has been used to devise goodness-of-fit statistical tests to estimate the relevance of the standard neutral model for an observed dataset. SFS-based neutrality tests contrast estimations of the nucleotide variability from different bins of the sample SFS (Tajima, 1989; Fu and Li,

* Corresponding author at: The Pirbright Institute, Woking, United Kingdom.

E-mail address: luca.ferretti@gmail.com (L. Ferretti).¹ These authors contributed equally.<https://doi.org/10.1016/j.tpb.2018.06.001>

0040-5809/© 2018 Elsevier Inc. All rights reserved.

1993; Achaz, 2009). It was shown that, once the SFS under an alternative scenario (e.g. selection, demography or structure) is known, the optimal test to reject the standard neutral model is based on the difference between the standard neutral SFS and the alternative scenario SFS (Ferretti et al., 2010). All these tests assume complete linkage among variants in their null model.

Assuming independence between the sites, the observed SFS can also be used to estimate model parameters. An interesting recent approach is the estimation of piece-wise constant demography from genomewide SFS (e.g. Liu and Fu, 2015). More sophisticated methods based on the expected SFS, such as Poisson Random Field (Sawyer and Hartl, 1992; Bustamante et al., 2001, 2002) and Composite Likelihood approaches (e.g., Kim and Stephan, 2002; Li and Stephan, 2005; Kim and Nielsen, 2004; Nielsen et al., 2005), have also played an important role in the detection of events of selection across regions of the genome. However, the assumption of linkage equilibrium is often violated in genetic data. In fact, while the average spectrum is insensitive to recombination, the knowledge on linked variants affects the distribution of summary statistics, therefore the spread (and possibly the mean) of the estimated parameters (Hudson et al., 1990; Thornton, 2005). For this reason, simulations of the evolution of linked sequences are required for an accurate estimation of the statistical support for different models (Gutenkunst et al., 2009).

The joint SFS for multiple sites has been the subject of long-standing investigations. The simplest spectrum for multiple sites is the “two-locus frequency spectrum” (Hudson, 2001), which we name the “two-Sites Frequency Spectrum” or 2-SFS. Assuming independence between the sites (*i.e.* free recombination), it simply reduces to the random association between two single-sites spectra (1-SFS). For intermediate recombination, a recursion solvable for small sample size has been provided (Golding, 1984; Ethier and Griffiths, 1990) as well as a numerical solution relying on simulations (Hudson, 2001). Even without recombination, finding an analytical expression for the spectrum has proven to be difficult.

There is a close relation between the m -SFS (the joint SFS of m sites) and the multi-allelic spectrum of a single locus (defined as a sequence with one or more sites). Under the *infinite-sites* model, sites are assumed to have at most two alleles as new mutations occur exclusively at non-polymorphic sites. At the locus scale, each haplotype (the specific combination of the alleles carried at each locus) can be interpreted as a single allele at a multi-allelic locus. In the absence of recombination, each point mutation either leaves the number of different haplotypes unchanged or generates one new haplotype. Therefore, at least conceptually, the SFS for m non-recombinant biallelic sites at low mutation rate is closely related to the spectrum of $m + 1$ alleles in a multi-allelic locus. Indeed, it is possible to retrieve the latter from the former by considering the $m + 1$ alleles that result from the m polymorphic sites. However, the m -SFS contains extra-information on the different couplings between sites that is not available in the multi-allelic spectrum.

For an infinite population, the multi-alleles single-locus spectrum is the solution of a multiallelic diffusion equation (Ewens, 2012, section 5.10). Polynomial expansions were proposed to solve the diffusion equations for the SFS of an infinite population (Kimura, 1956; Littler and Fackerell, 1975; Griffiths, 1979), as well as moment-based approaches (Hobolth and Siren, 2016). Finally, a polynomial expansion of the 2-SFS has been found for two sites without recombination and with general selection coefficients (Xie, 2011). However, the reported solution is an infinite series and is in sharp contrast with the simplicity of the solution for a single neutral site: $E[\xi(f)] = \theta/f$. Furthermore, no closed form was provided for the 2-SFS of a sample.

Using a coalescent framework, the probability and size of two nested mutations were expressed by Hobolth and Wiuf (2009) as

sums of binomial coefficients. Their formulae can be rewritten as an expected SFS in terms of a finite series. However their conditioning on exactly two nested mutations skews the spectrum and simulations show that their result is valid only for $L\theta \ll 1$. Interesting analytical results on the spectrum of tri-allelic loci and recurrent mutations were obtained by Jenkins, Song and collaborators (Jenkins and Song, 2011; Jenkins et al., 2014) for the Kingman coalescent and general allelic transition matrices. More recently, Sargsyan (2015) generalized the result of Hobolth and Wiuf (2009) by conditioning on any two mutations (nested or not) and extending it to populations of variable size. Moreover, he clarified the notion and classification of the 2-SFS.

In this work, we review and present in its simplest possible form the exact solution for the expectation of the neutral sample 2-SFS without recombination, then we extend it to a closed-form solution for the continuous population 2-SFS. The solution for a finite sample was derived previously in many disguises in a coalescent framework (Fu, 1995; Jenkins and Song, 2011; Ferretti et al., 2012; Sargsyan, 2015) and its extrapolation to the limit of infinite sample sizes yields the continuous spectrum, which is a solution of the multi-allelic Kolmogorov equations. Furthermore, we derive the expected 1-SFS of sites that are completely linked to a focal mutation of known frequency. This spectrum has several potential applications. In section S1 of the Supplementary Material we extend the formulae for the continuous 2-SFS to closed expressions for the multi-allelic spectrum of a locus with three alleles.

Finally, as an application, we present a recipe to build a class of SFS-based neutrality tests for sequences containing a known neutral marker of given frequency. This is a typical scenario when the marker (and the region around it) has been detected independently as an outlier in genome-wide association studies or population differentiation studies with SNP arrays. As far as we know, this is the first proper adaptation of Tajima’s D and similar statistical tests to this kind of sequence data.

Model definition and notation

We consider a population of N haploid individuals without recombination. All subsequent results can be applied to diploids, provided that $2N$ is used instead of N , and to other cases by substituting the appropriate effective population size. We denote by μ the mutation rate per site and by $\theta = 2N\mu$ the population-scaled mutation rate per site. We work in the infinite-sites approximation, that is valid in the limit of small mutation rates $\theta \ll 1$. More precisely, our results are derived in the limit $\theta \rightarrow 0$ with fixed non-zero θL , where L is the length of the sequence. The expected value $E[\cdot]$ denotes the expectation with respect to the realizations of the evolutionary process for the sequences in the sample or in the whole population. We use *mutation* as a synonym for derived allele.

Connection between sample and population SFS

We denote by $\xi(f)$ the *density* of mutations at frequency f in the whole population and by ξ_k the *number* of mutations at frequency k/n in a sample of size n . Importantly, in both cases f or k refer to the frequency of the mutation, *i.e.* of the *derived* allele, and thus ξ corresponds to the *unfolded* SFS.

The two spectra (sample and population) are related. Assuming that a mutation has frequency f in the population, the probability of having k mutant alleles in a random sample of size n is simply given by the Binomial $\binom{n}{k} f^k (1-f)^{n-k}$. As the expected density of mutations at fixed frequency f in the population is given by $E[\xi(f)]$, one can easily derive the sample frequency from the population

frequency using the following sampling formula:

$$E[\xi_k] = \int_{\frac{1}{N}}^{1-\frac{1}{N}} \binom{n}{k} f^k (1-f)^{n-k} E[\xi(f)] df \quad (1)$$

assuming that $n \ll N$.

Conversely, the population SFS can be derived from the sample SFS using the limit of large sample size $n \rightarrow \infty$. For a sample of n individuals, the interval between the frequency bins is $1/n$ and therefore the density of mutations at the continuous frequency $f = k/n$ can be approximated² by $E[\xi(\frac{k}{n})] \approx \frac{E[\xi_k]}{1/n} = nE[\xi_k]$. The expected population spectrum can then be constructed from the limit:

$$E[\xi(f)] = \lim_{n \rightarrow \infty} nE[\xi_{\lfloor nf \rfloor}] \quad (2)$$

for frequencies not too close to $\frac{1}{N}$ or $1 - \frac{1}{N}$.

For a sample of size n , the expected neutral spectrum for constant population size is $E[\xi_k] = \theta L/k$ and consequently, we have $E[\xi(f)] = \theta L/f$ (Wright, 1938; Ewens, 2012). These results are exact for the Kingman coalescent and the diffusion equations respectively, and they are approximately valid for neutral models for frequencies $f \gg \frac{1}{N}$. For frequencies of order $\frac{1}{N}$, model-dependent corrections are needed and Eq. (2) is not valid anymore.

In the rest of this section we will deal with sample and population spectra together. We will slightly abuse the notation and switch between number and density of mutations, or probability and probability density.

Conditional 1-SFS and joint 2-SFS

In the following, we will use two related but different kinds of spectra.

The first kind is the joint 2-SFS of two bi-allelic sites. It is denoted $\xi(f_1, f_2)$ for the population and $\xi_{k,l}$ for the sample. It is defined as the density of pairs of sites with mutation frequencies at f_1 and f_2 for the population (resp. k/n and l/n for the sample). This is a natural generalization of the classical SFS for a single site. The expected spectrum $E[\xi(f_1, f_2)]$ has multiple equivalent interpretations in the small θ limit: (a) for a sequence, it is the expected density of pairs of sites that harbor mutations with frequencies f_1 and f_2 ; (b) for two randomly chosen linked polymorphic sites, it is the probability density that they contain mutations with frequencies f_1 and f_2 . Here we always consider unordered pairs of sites (the ordered case is discussed in section S2).

The second kind of spectrum is a conditional 1-SFS, a frequency spectrum of sites that are linked to a focal mutation of frequency f_0 . It is denoted $\xi(f|f_0)$ for the population and $\xi_{k|l}$ for the sample. Again, this spectrum represents both (a) the expected density of single-site mutations of frequency f in a locus linked to a focal neutral mutation of frequency f_0 and (b) the probability density that a randomly chosen site (linked to the focal site) hosts a mutation at frequency f .

Note that despite the similarity in notation, the two spectra $\xi(f, f_0)$ and $\xi(f|f_0)$ are different. The difference is the same as the one between the joint probability $p(f, f_0)$ that two sites x and x_0 have mutations of frequency f and f_0 respectively, and the conditional probability $p(f|f_0)$ that a mutation at site x has frequency f given that there is a mutation of frequency f_0 at a focal linked site x_0 . Furthermore, the joint spectrum $\xi(f, f_0)$ refers to pairs of sites – i.e. it is a 2-SFS – while the spectrum of linked sites $\xi(f|f_0)$ is a single-site SFS.

² More formally, Eq. (2) can be obtained from Eq. (1) under the assumptions that $\frac{1}{N} \ll f, 1-f$ and that the population SFS is smooth over a range of frequencies $\Delta f \sim \frac{1}{N}$.

The relation between both types of spectra can be understood from the relation between the probabilities. The expected spectrum $E[\xi(f)]$ is given by the probability to find a mutation of frequency f at a specific site, multiplied by the length of the sequence: $E[\xi(f)] = p(f)L$. As noted above, when $L = 1$ (i.e. a locus with a single site is considered), $E[\xi(f)]$ corresponds to a proper probability $p(f)$. Assuming the presence of a mutation of frequency f_0 at a focal site, we have $E[\xi(f|f_0)] = p(f|f_0)(L - 1)$. For pairs of sites, the expected number of mutations at frequencies (f, f_0) is $E[\xi(f, f_0)] = p(f, f_0)L(L - 1)$ when $f \neq f_0$ or $p(f_0, f_0)L(L - 1)/2$ when $f = f_0$. The additional factor $\frac{1}{2}$ accounts for the symmetrical case of equal frequencies $f = f_0$. The equality $p(f, f_0) = p(f|f_0)p(f_0)$ applied to sample and population spectra, results in the following relations:

$$E[\xi_{k,l}] = \frac{E[\xi_{k|l}] \cdot E[\xi_l]}{1 + \delta_{k,l}} = \begin{cases} E[\xi_{k|l}] \cdot E[\xi_l] & \text{for } k \neq l \\ \frac{1}{2} \cdot E[\xi_{k|l}] \cdot E[\xi_l] & \text{for } k = l \end{cases} \quad (3)$$

$$E[\xi(f, f_0)] = \frac{E[\xi(f|f_0)] \cdot E[\xi(f_0)]}{1 + \delta_{f,f_0}} = \begin{cases} E[\xi(f|f_0)] \cdot E[\xi(f_0)] & \text{for } f \neq f_0 \\ \frac{1}{2} \cdot E[\xi(f|f_0)] \cdot E[\xi(f_0)] & \text{for } f = f_0 \end{cases} \quad (4)$$

where $\delta_{x,y}$ is 1 if $x = y$, and 0 otherwise. Note that x and y can be either discrete or continuous variables.

By definition, the 2-SFS includes only pairs of sites that are both polymorphic. The probability that a pair of sites contains a single polymorphism of frequency k/n depends only on the 1-SFS and it is approximately equal to $2E[\xi_k]$ for $\theta \ll 1$. Consequently, on a sequence of size L hosting S polymorphic sites, the number of pairs of sites for which only one of the two is polymorphic of frequency k/n is $E[(L - S)\xi_k] = L \cdot E[\xi_k] - E[S\xi_k] \approx L \cdot E[\xi_k]$ for small θ .

2. Results

2.1. Decomposition of the 2-SFS

We follow (Sargsyan, 2015) and divide the 2-SFS $\xi(f_1, f_2)$ without recombination into two different components: one nested component $\xi^{(n)}(f_1, f_2)$ for cases where there are individuals carrying the two mutations (one is “nested” in the other), and a disjoint component $\xi^{(d)}(f_1, f_2)$ that includes disjoint mutations that are only present in different individuals. The overall spectrum is given by:

$$\xi(f_1, f_2) = \xi^{(n)}(f_1, f_2) + \xi^{(d)}(f_1, f_2) \quad (5)$$

$$\xi_{k,l} = \xi_{k,l}^{(n)} + \xi_{k,l}^{(d)} \quad (6)$$

It is noteworthy to mention that the overall spectrum is not sufficient to provide a full description of the genetic state of the two sites, while the two components $\xi^{(n)}(f_1, f_2)$, $\xi^{(d)}(f_1, f_2)$ are enough to reconstruct the genetic content of the two sites up to permutations of all the haplotypes, as it happens with the usual SFS for one site. For example, the following two sets of haplotypes (derived alleles marked in bold)

CT		CA
CA	and	CA
GA		GT

are identical from the point of view of the overall two-loci spectrum: in both samples there is just a pair of mutations with allele count 1 and 2 respectively, therefore the only (symmetrical) nonzero value of the spectrum is $\xi_{1,2} = \xi_{2,1} = 1$. However the samples can be distinguished by the two components, since in the first one the mutations are nested ($\xi_{1,2}^{(n)} = \xi_{2,1}^{(n)} = 1$), while in the

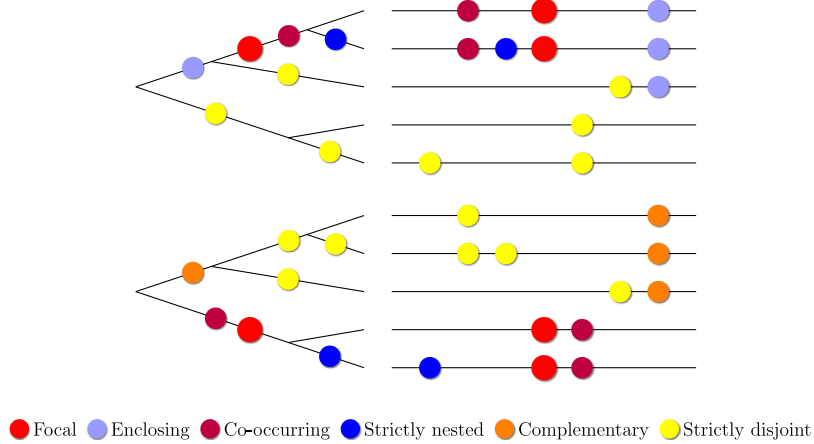


Fig. 1. A schema of two non-recombining genomic regions and their corresponding genealogical trees. The black lines on the right represent sequences and the colored circles derived alleles. This figure illustrates the classification of all possible types of mutations with respect to the focal mutation (in red) and their occurrence on the sequence tree. If the focal mutation is not on a root branch (upper panel), it is clear that mutations can be on the same branch as the focal mutation (*co-occurring*), on the subtree below (*strictly nested*), between the focal mutation and the root (*enclosing*), or on other branches (*strictly disjoint*). If the mutation is on a root branch (lower panel), there cannot be enclosing mutations, but there can be mutations on the other root branch (*complementary*). (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

second one they are disjoint ($\xi_{1,2}^{(d)} = \xi_{2,1}^{(d)} = 1$). For this reason, these two components constitute the core of the two-loci SFS.

Without recombination, the conditional 1-SFS $\xi(f|f_0)$ can be also decomposed further³ into different subspectra. They are illustrated in Fig. 1:

- $\xi^{(sn)}(f|f_0)$: *strictly nested* mutations, where the mutation is carried only by a subset of individuals with the focal mutation;
- $\xi^{(co)}(f|f_0)$: *co-occurring* mutations, where both mutations are carried by the same individuals;
- $\xi^{(en)}(f|f_0)$: *enclosing* mutations, where only a subset of individuals with the mutation also carry the focal one;
- $\xi^{(cm)}(f|f_0)$: *complementary* mutations, where each individual has exactly one of the two mutations;
- $\xi^{(sd)}(f|f_0)$: *strictly disjoint* mutations, where the mutation is carried by a subset of the individuals without the focal one.

Importantly, without recombination, enclosing and complementary mutations cannot be present together in the same sequence, as both types of branches are exclusive in a single tree.

With the above definition and using the rules of conditional probabilities $p(f, f_0) = p(f|f_0)p(f_0)$ and the interpretations discussed in the previous section, the relations between the two sets of population subspectra are:

$$E[\xi^{(n)}(f, f_0)] = \left(E[\xi^{(sn)}(f|f_0)] + E[\xi^{(co)}(f|f_0)] + E[\xi^{(en)}(f|f_0)] \right) \cdot \frac{E[\xi(f_0)]}{1 + \delta_{f,f_0}} \quad (7)$$

$$E[\xi^{(d)}(f, f_0)] = \left(E[\xi^{(cm)}(f|f_0)] + E[\xi^{(sd)}(f|f_0)] \right) \cdot \frac{E[\xi(f_0)]}{1 + \delta_{f,f_0}} \quad (8)$$

Similarly, for sample spectra, we have

$$E[\xi_{k,l}^{(n)}] = \left(E[\xi_{k,l}^{(sn)}] + E[\xi_{k,l}^{(co)}] + E[\xi_{k,l}^{(en)}] \right) \cdot \frac{E[\xi_l]}{1 + \delta_{k,l}} \quad (9)$$

$$E[\xi_{k,l}^{(d)}] = \left(E[\xi_{k,l}^{(cm)}] + E[\xi_{k,l}^{(sd)}] \right) \cdot \frac{E[\xi_l]}{1 + \delta_{k,l}} \quad (10)$$

³ We subdivide the “strictly nested” mutations of Sargsyan (2015) into *strictly nested* and *enclosing* mutations while we refer to “identical” mutations as *co-occurring*.

2.2. The joint and conditional SFS

In this section, we present the conditional and joint spectra for the sample and the population. The derivations and proofs of all equations in this section are given in Methods and sections S3 and S4 of the Supplementary Material. The folded version of the 2-SFS is provided in Appendix A.

2.2.1. The sample joint 2-SFS

The 2-loci spectrum appeared in the literature under many guises (Fu, 1995; Jenkins and Song, 2011; Ferretti et al., 2012; Sargsyan, 2015). In the infinite-sites neutral model without recombination, its expected value has a simpler form⁴:

$$E[\xi_{k,l}^{(n)}] = \begin{cases} \frac{\theta^2 L^2 \beta_n(k) - \beta_n(k+1)}{2} & \text{for } k < l \\ \frac{\theta^2 L^2 \beta_n(k)}{2} & \text{for } k = l \\ \frac{\theta^2 L^2 \beta_n(l) - \beta_n(l+1)}{2} & \text{for } k > l \end{cases}$$

$$E[\xi_{k,l}^{(d)}] = \begin{cases} \theta^2 L^2 \left(\frac{1}{kl} - \frac{\beta_n(k) - \beta_n(k+1) + \beta_n(l) - \beta_n(l+1)}{2} \right) \frac{2 - \delta_{k,l}}{2} & \text{for } k+l < n \\ \theta^2 L^2 \left(\frac{a_n - a_k}{n-k} + \frac{a_n - a_l}{n-l} - \frac{\beta_n(k) + \beta_n(l)}{2} \right) \frac{2 - \delta_{k,l}}{2} & \text{for } k+l = n \\ 0 & \text{for } k+l > n \end{cases} \quad (11)$$

with

$$a_n = \sum_{i=1}^{n-1} \frac{1}{i}, \quad \beta_n(i) = \frac{2n}{(n-i+1)(n-i)}(a_{n+1} - a_i) - \frac{2}{n-i}$$

As shown by Eq. (6), the full spectrum is simply the sum of the two above equations.

⁴ Note that the related formula (14) in the paper by Ferretti et al. (2012) has a sign error. It should be identical to the second equation in (11) up to a multiplicative factor.

2.2.2. The population joint 2-SFS

Similarly, the 2-SFS for the whole population is given by the combination of the two following equations:

$$\begin{aligned}
 E[\xi^{(n)}(f, f_0)] &= \theta^2 L^2 \\
 &\cdot \left[\frac{1}{(1 - \min(f, f_0))^2} \left(1 + \frac{1}{\min(f, f_0)} + \frac{2 \ln(\min(f, f_0))}{1 - \min(f, f_0)} \right) \right. \\
 &+ \left. \delta(f - f_0) \frac{f_0}{1 - f_0} \left(-\frac{\ln(f_0)}{1 - f_0} - 1 \right) \right] \\
 E[\xi^{(d)}(f, f_0)] &= \theta^2 L^2 \cdot \left[\frac{1}{ff_0} - \frac{1}{(1 - f)^2} \left(1 + \frac{1}{f} + \frac{2 \ln(f)}{1 - f} \right) \right. \\
 &- \frac{1}{(1 - f_0)^2} \left(1 + \frac{1}{f_0} + \frac{2 \ln(f_0)}{1 - f_0} \right) \\
 &+ \delta(f - 1 + f_0) \left(\frac{1 - f_0}{f_0^2} \ln(1 - f_0) \right. \\
 &+ \left. \left. \frac{f_0}{(1 - f_0)^2} \ln(f_0) + \frac{1}{f_0(1 - f_0)} \right) \right] \quad (12)
 \end{aligned}$$

with $E[\xi^{(n)}(f, f_0)] = 0$ for $f > f_0$ and $E[\xi^{(d)}(f, f_0)] = 0$ for $f + f_0 > 1$.

Here, we denote by $\delta(f - f_0)$ the density of the Dirac delta distribution concentrated in f_0 (i.e. $\delta(f - f_0) = 0$ for $f \neq f_0$, normalized such that $\int_{-\infty}^{\infty} \delta(f - f_0) df = 1$).

2.2.3. The sample conditional 1-SFS

The conditional 1-SFS for sites that are linked to a focal mutation of count l is simply the sum of all its components, given by the following equations:

$$\begin{aligned}
 E[\xi_{kl}^{(sn)}] &= \theta L \cdot l \frac{\beta_n(k) - \beta_n(k + 1)}{2} \quad \text{for } k < l \\
 E[\xi_{kl}^{(co)}] &= \theta L \cdot l \beta_n(k) \delta_{kl} \\
 E[\xi_{kl}^{(en)}] &= \theta L \cdot l \frac{\beta_n(l) - \beta_n(l + 1)}{2} \quad \text{for } k > l \\
 E[\xi_{kl}^{(cm)}] &= \theta L \cdot l \left(\frac{a_n - a_k}{n - k} + \frac{a_n - a_l}{n - l} - \frac{\beta_n(k) + \beta_n(l)}{2} \right) \delta_{k, n-l} \\
 E[\xi_{kl}^{(sd)}] &= \theta L \cdot \left(\frac{1}{k} - l \frac{\beta_n(k) - \beta_n(k + 1) + \beta_n(l) - \beta_n(l + 1)}{2} \right) \\
 &\quad \text{for } k + l < n
 \end{aligned} \quad (13)$$

Please note that hereafter unmet conditions imply 0 otherwise.

The strictly nested component of the conditional 1-SFS and its applications have been discussed by Griffiths and Tavaré (2003).

2.2.4. The population conditional 1-SFS

For the whole population, the expected linked SFS becomes:

$$\begin{aligned}
 E[\xi^{(sn)}(f|f_0)] &= \theta L \cdot \frac{f_0}{(1 - f)^2} \left(1 + \frac{1}{f} + \frac{2 \ln(f)}{1 - f} \right), \quad f < f_0 \\
 E[\xi^{(co)}(f|f_0)] &= \theta L \cdot \delta(f - f_0) \frac{2f_0}{1 - f_0} \left(-\frac{\ln(f_0)}{1 - f_0} - 1 \right) \\
 E[\xi^{(en)}(f|f_0)] &= \theta L \cdot \frac{f_0}{(1 - f_0)^2} \left(1 + \frac{1}{f_0} + \frac{2 \ln(f_0)}{1 - f_0} \right), \quad f > f_0 \\
 E[\xi^{(cm)}(f|f_0)] &= \theta L \cdot \delta(f - 1 + f_0) \\
 &\quad \times \left[\frac{1 - f_0}{f_0} \ln(1 - f_0) + \left(\frac{f_0}{1 - f_0} \right)^2 \ln(f_0) + \frac{1}{1 - f_0} \right] \\
 E[\xi^{(sd)}(f|f_0)] &= \theta L \cdot \left[\frac{1}{f} - \frac{f_0}{(1 - f)^2} \left(1 + \frac{1}{f} + \frac{2 \ln(f)}{1 - f} \right) \right. \\
 &\quad \left. - \frac{f_0}{(1 - f_0)^2} \left(1 + \frac{1}{f_0} + \frac{2 \ln(f_0)}{1 - f_0} \right) \right], \quad f < 1 - f_0
 \end{aligned} \quad (14)$$

2.3. Shape of the SFS

We report the full joint 2-SFS as well as the nested and disjoint components (Fig. 2). Nested mutations have preferentially a rare mutation in either site – so that the mutation at lower frequency is easily nested into the other – or are co-occurring mutations. Disjoint mutations are dominated by cases where both mutations are rare, or by complementary mutations. The large contribution of co-occurring (nested component) and complementary mutations (disjoint component) is a direct consequence of the two long branches that coalesce at the root node of a Kingman tree.

The conditional 1-SFS of linked sites and the relative contributions of each component to each frequency are shown in Fig. 3. Co-occurring and complementary mutations also account for a considerable fraction of the spectrum, especially when the focal mutation (f_0) is at high frequency. The rest of the spectrum is biased towards mutations with a lower frequency than the focal one. Strictly nested mutations are important only when the frequency of the focal mutation is intermediate or high. Enclosing mutations are rare and their frequencies are uniformly distributed, as previously noted (Hobolth and Wiuf, 2009).

Finally, in Fig. 4 we show the impact of a focal mutation of given frequency on two estimators of θ . The Watterson’s estimator $\hat{\theta}_S$ (Watterson, 1975) depends on the total number of polymorphic sites, which increases with the frequency of the focal mutation, as they inflate increasingly upper sections of the trees that contribute more to the total tree length. On the other hand, Tajima’s estimator, $\hat{\theta}_\pi$ (Tajima, 1983) is more sensitive to mutations of intermediate frequency. The difference between the two illustrates how the spectrum is skewed towards common or rare mutations. As Tajima’s D (Tajima, 1989) is proportional to the difference $\hat{\theta}_\pi - \hat{\theta}_S$, positive values for this test statistic suggest an excess of common mutations while negative values point to an excess of rare mutations. Fig. 4 shows that the spectrum has a slight excess of rare mutations at low frequencies of the focal mutation and an excess of common mutations for intermediate frequencies, while it is dominated again by rare mutations if the focal mutation is at high frequencies.

2.4. Neutrality tests for regions linked to a polymorphic neutral marker

Biallelic putative neutral markers are often used to find regions of interest in a genome. For example, genotype data from SNP arrays can be used together with phenotype measurements to find regions associated with a specific phenotype. Alternatively, if data from multiple populations are available, markers with highly differentiated frequency between populations – i.e. high F_{st} – can be used to infer potential targets of local selection. It is then natural to use sequence data to test for neutral evolution in a window around the neutral focal marker of known frequency.

Up to now, such tests did not take into account the information given by the frequency of the marker itself. However, since typical markers have biased frequencies towards intermediate values (Lachance and Tishkoff, 2013), the expected neutral frequency spectrum will likely be dramatically altered. An example of the dependence of the distribution of Tajima’s D on the frequency of the marker is shown in Figure S2.

The results of the previous section show precisely how Tajima’s D test values for a fully linked locus is biased as a function of the frequency of the marker. These biases can be computed analytically for all Tajima’s D -like SFS-based tests (Achaz, 2009) in a similar way, using Eq. (14) and simple approximations (e.g. replacing S by its conditional expected value in their denominator).

Furthermore, it is possible to develop versions of Tajima’s D and other frequency-spectrum based neutrality tests that take into

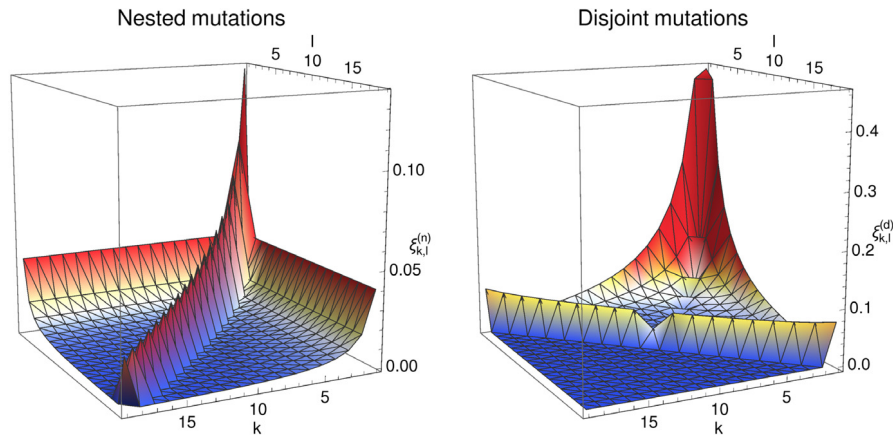


Fig. 2. Plots of nested and disjoint contributions to the two-locus frequency spectrum for $\theta L = 1, n = 20$. Note the different scales of the two plots.

Expected frequencies of mutations linked to a mutation of size k in a sample of size $n = 20$

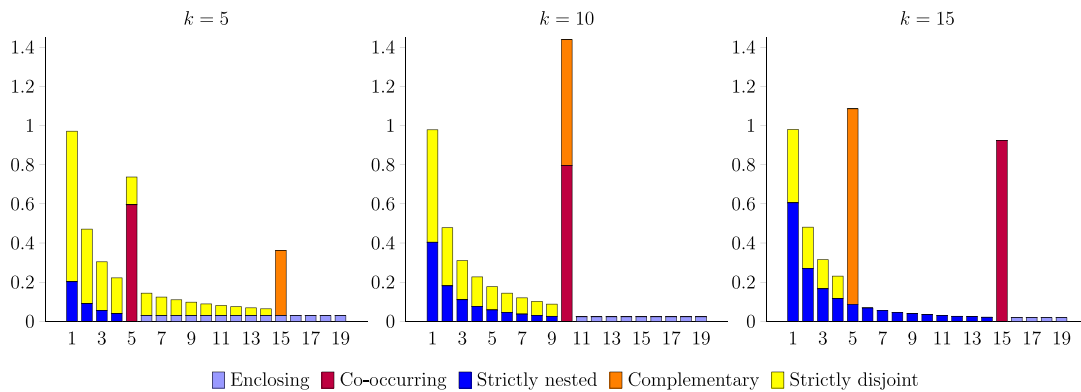


Fig. 3. Barplot of the spectrum of linked sites for $\theta L = 1, n = 20$, each column colored according to the different contributions. The focal mutation has frequency 0.25 (left), 0.5 (middle) and 0.75 (right) respectively.

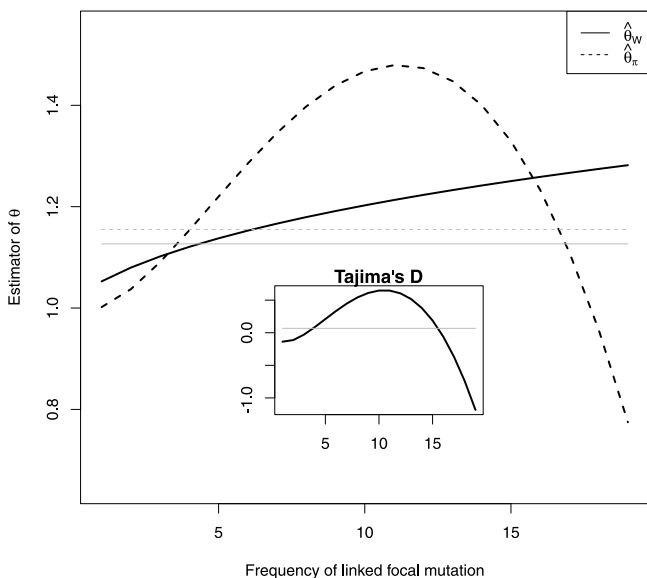


Fig. 4. Mean values of the Watterson estimator ($\hat{\theta}_w$) and Tajima estimator ($\hat{\theta}_\pi$) of θ conditioned on the presence of a linked mutation, for $\theta = 1, n = 20$. In the inset, approximate mean value of Tajima's D (computed substituting S with its mean value in the denominator). The gray lines represent the expected values conditioned on the presence of a linked polymorphism of any frequency.

account the presence of the neutral marker. The simplest approach follows Rafajlović et al. (2014) and consists in replacing the neutral spectrum $\theta L/i$ by the conditional spectrum $\xi_{i|m}$, where m is the count of the marker in the sample. The usual covariance of the spectrum from Fu (1995), which appears in the normalization of the tests, can be replaced by the one derived by Klassmann and Ferretti (2018) for the conditional spectrum. We denote the neutral marker by ϕ and the allele count of its derived allele by m . We use the observed spectrum $\hat{\xi}_{i|\phi}$ for a window of size L containing the marker to build a test of the general form (Achaz, 2009; Ferretti et al., 2010):

$$T_\Omega = \frac{\sum_{i=1}^{n-1} \Omega_i \hat{\xi}_{i|\phi} / \xi_i^0}{\sqrt{\text{Var} \left[\sum_{i=1}^{n-1} \Omega_i \hat{\xi}_{i|\phi} / \xi_i^0 \right]}} \quad (15)$$

where both the null spectrum $\xi_i^0 = E[\xi_{i|\phi}] / \theta$ and the variance in the denominator are computed under the standard neutral model, i.e. Kingman's coalescent. The real vector of parameters Ω can be chosen in any possible way, as long as it satisfies $\sum_{i=1}^{n-1} \Omega_i = 0$. For example, in the absence of a neutral marker, Tajima's D corresponds to $\Omega_i = \frac{2(n-i)}{n(n-1)}$.

The definition of the test requires that the null spectrum and the variance are conditioned on the presence of the marker ϕ . For the neutral spectrum, it is simply the sum of the expected nested and disjoint spectra presented in Eqs. (14):

$$\xi_i^0 = E[\xi_{i|m}] / \theta = \frac{E[\xi_{i|m}^{(n)}] + E[\xi_{i|m}^{(d)}]}{\theta} \quad (16)$$

On the other hand, the variance can be decomposed as

$$\text{Var} \left[\sum_{i=1}^{n-1} \Omega_i \hat{\xi}_{i|\phi} / \xi_i^0 \right] = \sum_{i=1}^{n-1} \Omega_i^2 \theta / \xi_i^0 + \sum_{i,j=1}^{n-1} \Omega_i \Omega_j \text{Cov}[\hat{\xi}_{i|\phi}, \hat{\xi}_{j|\phi}]_{\theta^2} / \xi_i^0 \xi_j^0 \quad (17)$$

where the θ contribution corresponds to the Poisson noise of the mutational process, while the θ^2 contribution to the covariance $\text{Cov}[\hat{\xi}_{i|\phi}, \hat{\xi}_{j|\phi}]_{\theta^2}$ can be easily obtained from the third moments of the spectrum derived by [Klassmann and Ferretti \(2018\)](#):

$$\begin{aligned} \text{Cov}[\hat{\xi}_{i|\phi}, \hat{\xi}_{j|\phi}]_{\theta^2} &= E[\hat{\xi}_{i|\phi}, \hat{\xi}_{j|\phi}]_{\theta^2} - \theta^2 \xi_i^0 \xi_j^0 = \\ &= E[\hat{\xi}_{i|\phi}^{(n)}, \hat{\xi}_{j|\phi}^{(n)}]_{\theta^2} + E[\hat{\xi}_{i|\phi}^{(n)}, \hat{\xi}_{j|\phi}^{(d)}]_{\theta^2} \\ &\quad + E[\hat{\xi}_{i|\phi}^{(d)}, \hat{\xi}_{j|\phi}^{(n)}]_{\theta^2} + E[\hat{\xi}_{i|\phi}^{(d)}, \hat{\xi}_{j|\phi}^{(d)}]_{\theta^2} - \theta^2 \xi_i^0 \xi_j^0 \end{aligned} \quad (18)$$

The test can then be built by putting together the results from the last section and an estimation of θ and θ^2 . The Maximum Composite Likelihood estimate can be used:

$$\hat{\theta} = S / \sum_{i=1}^{n-1} \xi_i^0, \quad \hat{\theta}^2 = (\hat{\theta})^2 \quad (19)$$

or the Method-of-Moments estimates as in the classical Tajima's D :

$$\hat{\theta} = S / \sum_{i=1}^{n-1} \xi_i^0, \quad \hat{\theta}^2 = S(S-1) / \sum_{i,j=1}^{n-1} (\xi_i^0 \xi_j^0 + \text{Cov}[\hat{\xi}_{i|\phi}, \hat{\xi}_{j|\phi}]_{\theta^2} / \theta^2) \quad (20)$$

The choice of weights for new tests of this form is somewhat arbitrary. For example, a modified version of Tajima's D could use the old weights, i.e. $\Omega_i = \frac{2(n-i)}{n(n-1)}$, or the old linear coefficients, i.e. $\Omega_i = \frac{2i(n-i)}{n(n-1)} \xi_i^0$, depending if the test should focus on the relative or absolute differences between the null and observed spectrum. Principles and formulae for a meaningful choice of the new weights are discussed in detail by [Ferretti et al. \(2010\)](#). On the other hand, once the weights are chosen, the normalization does not suffer from any degree of arbitrariness and its form depends only on the third moments computed here.

This straightforward modification of neutrality tests is a promising direction for future dedicated neutrality tests that aim at correcting multiple artefacts such as demography, knowledge of the frequency of the marker, etc.

3. Methods

3.1. The sample joint 2-SFS

To obtain the sample spectrum for pairs of mutations, we notice that this spectrum can be defined in terms of the expected value of crossproducts of the usual SFS. In detail, we have

$$E[\hat{\xi}_{k,l}] = E[\hat{\xi}_k \hat{\xi}_l], \quad \text{if } k \neq l \quad (21)$$

and

$$E[\hat{\xi}_{k,k}] = E[\hat{\xi}_k(\hat{\xi}_k - 1)]/2. \quad (22)$$

These expected values have been derived by [Fu \(1995\)](#) by coalescent methods. However his results do not distinguish the different contributions from nested and disjoint mutations to the spectrum. Tracking the origin of each term in the derivation, it is easy to show that Eqs. (24) and (28) of [Fu \(1995\)](#) contribute to nested pairs of mutations, while Eqs. (25), (29) and (30) contribute to disjoint pairs of mutations. All these terms combine linearly and

do not interfere, therefore we can decompose the resulting $E[\hat{\xi}_k \hat{\xi}_l]$ into contributions coming from Eqs. (24), (28) and (25), (29) and (30) of [Fu \(1995\)](#). This can be obtained directly by Fu's expression for the covariance matrix σ_{kl} , since $E[\hat{\xi}_k \hat{\xi}_l] = \delta_{k,l} E[\hat{\xi}_k] + E[\hat{\xi}_k] E[\hat{\xi}_l] + \theta^2 L^2 \sigma_{kl}$ and $E[\hat{\xi}_k] = \theta L/k$. A detailed review of the calculations of [Fu \(1995\)](#), tracking the parts that lead to our mutation classes, is provided in section S3 of the Supplementary Material.

The same results could also be obtained from Theorem 5.1 in [Jenkins and Song \(2011\)](#). In fact, for a special choice of allele transition matrices (in the triallelic case, a strictly lower triangular matrix with all non-zero entries equal to 1), their results for recurrent mutations for small θL (θ in their article) are mathematically equivalent to the results for mutations in an infinite-sites model. Their classification is based on the location of the mutations on the tree: their "nested mutations" correspond to strictly nested and enclosing mutations here, "mutations on the same branch" correspond to co-occurring mutations, "mutations on basal branches" correspond to complementary mutations, and "non-nested mutations" correspond to strictly disjoint mutations.

3.2. The sample conditional 1-SFS

The spectrum for sites linked to a focal mutation of count l (Eq. (13)) can be obtained from the previous spectrum (11). The first step is simply to condition on the frequency l/n of the focal mutation, i.e. dividing the 2-SFS $E[\hat{\xi}_{k,l}]$ by $E[\hat{\xi}_l]^{-\frac{1+\delta_{k,l}}{2}}$ following Eqs. (9) and (10). In fact, $E[\hat{\xi}_{k,l}] = (L-1)P[c(x) = k | c(y) = l] = L(L-1)P[c(x) = k, c(y) = l] / LP[c(y) = l] = \frac{2}{1+\delta_{k,l}} E[\hat{\xi}_{k,l}] / E[\hat{\xi}_l]$ where $c(x)$ is the derived allele count at site x . The second step is to break further the two contributions of the resulting conditional spectrum into the different components. Strictly nested, co-occurring and enclosing mutations are derived from the nested contribution and are distinguished by site frequencies only: strictly nested ones correspond to $k < l$, co-occurring ones to $k = l$ and enclosing ones to $k > l$. Similarly, from the disjoint contribution, mutations belonging to the strictly disjoint component can be obtained by selecting the frequency range $k + l < n$ while complementary ones correspond to $k + l = n$.

3.3. Population spectra

In the limit of large samples, the frequency spectra converge to the continuous SFS for infinite populations. However, the limit $n \rightarrow \infty$ should be taken with care. The easiest derivation proceeds as follows: since the conditional 1-SFS (Eq. (14)) is a single-locus spectrum, its population components can be obtained from the corresponding ones for finite samples (Eq. (13)) by direct application of Eq. (2). Then the population 2-SFS (Eq. (12)) can be reconstructed from Eqs. (7) and (8), by multiplying by the neutral spectrum $E[\xi(f_0)] = \theta L/f_0$ and by $\frac{1}{1+\delta_{f,f_0}}$ and combining the result into nested and disjoint contributions. The derivation makes use of the following functional limit of the Kronecker delta as a Dirac delta function: $n\delta_{[nf], [nf_0]} \rightarrow \delta(f - f_0)$ for $n \rightarrow \infty$. More details are given in section S4.

4. Discussion

In this article, we have provided exact closed formulae for the joint 2-SFS as well as the first expressions for the conditional 1-SFS, both for sample and population. The 2-SFS was already derived in different forms ([Jenkins and Song, 2011](#); [Ferretti et al., 2012](#); [Sargsyan, 2015](#)), but the expression presented here for the infinite-sites model is embedded in the framework of [Fu \(1995\)](#). Sample spectra were then used to derive the population spectra by letting $n \rightarrow \infty$. Importantly, our results only hold when there is no recombination, and are averaged across the tree space.

The analytical expressions provided in this paper can be intuitively understood in terms of the evolution of linked mutations. Consider a new mutation increasing in frequency by neutral drift and reaching low/intermediate frequency. We expect to find a large number of strictly disjoint and a low number of strictly nested linked mutations, since at the time of appearance of the focal mutation most existing mutations were “strictly disjoint”. The spectrum of strictly nested mutations is more skewed towards rare alleles than predicted by the neutral spectrum $1/f$, since strictly nested mutations evolve inside an expanding subpopulation. On the other hand, the spectrum of strictly disjoint mutations resembles the neutral one but with a slight bias against rare mutations, since they evolved in a slightly contracting subpopulation.

Note that for sequences linked to a mutation close to fixation, co-occurring and complementary mutations dominate. The contrast between the haplotypes produces a strong “haplotype structure”.

Interestingly, conditioning on the presence of a mutation of frequency f impacts the length and balance of the coalescent, as apparent from Fig. 4. This can be understood as follows. Rare mutations are common in any realization of the coalescent tree but especially common in the lower branches, therefore they just increase slightly the tree length and the length of the lower branches compared to the unconditioned case. Instead, mutations of intermediate frequency appear mostly in the upper branches of the tree, therefore the presence of such mutations implies higher, more balanced trees. The effect is even stronger for high frequency mutations, which reside only in the uppermost branches, implying highly unbalanced trees.

There are several potential applications of these results. Here we discuss approaches to correct SFS-based neutrality tests taking into account the presence of a neutral marker strongly linked to the genomic region. These corrections are useful in cases when the region has been selected on the basis of evidence from genome-wide association studies or studies of differentiation based on SNP arrays or other (putatively) neutral markers. This is just an example of possible extensions of neutrality tests based on these results. Other applications include the improvement of population genetic inference techniques based on the SFS, such as composite likelihood (e.g. Kim and Stephan, 2002; Li and Stephan, 2005; Kim and Nielsen, 2004; Nielsen et al., 2005) and Poisson Random Field methods (Sawyer and Hartl, 1992). These methods use analytical expressions for the SFS for a single site together with approximations of independence between different sites. For sequences with low recombination, methods could be made more rigorous by assuming independence between different pairs of sites, while taking pairwise dependence between sites into account through the two-locus SFS developed here.

The spectrum could also be useful for new neutrality tests based on linkage between mutations. Our results lead to a better understanding of the linkage disequilibrium (LD) structure among neutral loci, therefore they can be immediately applied to LD-related statistics, for example to compute average LD across non-recombining neutral loci. As an example, it can be checked numerically that the expected value of D between fully linked derived mutations is 0 according to our equations, as expected from LD theory. Furthermore, they can be used to build neutrality tests optimized to detect positive or balancing selection through its effect on the frequency spectrum of linked sites. The spectra presented here could also provide a neutral model for other scenarios, including structural variants or introgressions from different species or populations. Introgressed sequences from close species can be detected as divergent haplotypes in the locus considered, and if introgressions are rare, then the genetic variability within these haplotypes is described by the nested spectrum linked to the introgressed haplotypes.

The SFS presented here is the simplest two-locus spectrum for neutral, non-recombining mutations in a population of constant size. These results could be extended to variable population size using the approach of Živković and Wiehe (2008); Jenkins and Song (2011) and to mutations in rapidly adapting populations using the Λ -coalescent approximation and the results of Birkner et al. (2013). However, the most interesting extensions would be to consider (a) non-neutral mutations and (b) recombination.

Adding selection to the two-locus SFS would significantly enhance its potential for most of the applications discussed above. The SFS for pairs of selected mutations has been obtained by Xie (2011) as a polynomial expansion, but the numerical computation of this expansion is still cumbersome. Given the simplicity of the expression for the single-locus SFS $\xi(f) = \theta(1 - e^{-2N_e s(1-f)})/f(1 - f)(1 - e^{-2N_e s})$ (Wright, 1938; Sawyer and Hartl, 1992), we expect that closed expressions could be found for pairs of mutations with different selective coefficients. This would be a promising development for future investigations.

The classical correspondence between the Kingman model in the large n limit and the diffusion approximation suggests that the 2-SFS spectrum presented here is a solution of the diffusion equations for three alleles (Ewens, 2012, section 5.10). In fact, the nested component of the 2-SFS for $f \neq f_0$ is a stationary solution of the diffusion equation of three alleles of frequency f , $f_0 - f$ and $1 - f_0$:

$$\frac{\partial \xi}{\partial t} = \frac{1}{2N_e} \left(\frac{\partial^2}{\partial f^2} [f(1-f)\xi] + 2 \frac{\partial^2}{\partial f \partial f_0} [f(1-f_0)\xi] + \frac{\partial^2}{\partial f_0^2} [f_0(1-f_0)\xi] \right) \quad (23)$$

while the disjoint component for $f \neq 1 - f_0$ is a stationary solution of the diffusion equation of three alleles of frequency f , f_0 and $1 - f_0 - f$:

$$\frac{\partial \xi}{\partial t} = \frac{1}{2N_e} \left(\frac{\partial^2}{\partial f^2} [f(1-f)\xi] - 2 \frac{\partial^2}{\partial f \partial f_0} [ff_0\xi] + \frac{\partial^2}{\partial f_0^2} [f_0(1-f_0)\xi] \right) \quad (24)$$

The correspondence implies that the solution (12) is actually the stationary solution of the full set of diffusion equations for the system, including boundary equations for $f = f_0$ and $1 - f_0$ and boundary conditions. A direct proof of this result using methods from the theory of partial differential equations could lead to interesting developments towards new solutions for selective equations as well. Our results could also be used to test the accuracy of existing tools based on a numerical solution of the diffusion equations (Ragsdale and Gutenkunst, 2017).

On the other hand, finding the exact two-locus SFS with recombination appears to be a difficult problem. Recombination is intrinsically related to the two-locus SFS via the same definition of linkage disequilibrium. Obtaining the full two-locus spectrum with selection and recombination could open new avenues for model inference and analysis of genomic data. For this reason, many approximations and partial results have been developed since Hudson (2001), like expansions in the limit of strong recombination (Jenkins and Song, 2012). The SFS of linked loci presented in this paper could be useful as a starting point for different approaches to the effect of recombination events, for example for perturbation expansions at low recombination rates.

An immediate application of our results to recombination events is the following: since in the Ancestral Recombination Graph (Griffiths and Marjoram, 1997) the recombination events follow a Poisson process similar to mutation events, although with a different rate, the spectrum $\xi_{k|l}$ could also be reinterpreted (up to a constant) as the probability that a single recombination event affects k extant lineages in a sequence linked to a specific mutation

of frequency l , i.e. it is equivalent to the spectrum of mutation–recombination events. This approach could be applied to higher moments of the frequency spectrum and lead to new results in recombination theory.

We offer tools for computing the analytical spectra as well as performing simulations by manipulating output of the program *ms* (Hudson, 2002). The corresponding C++ code is contained in the package *coatli* developed by one of the authors and available on <http://sourceforge.net/projects/coatli/>.

Acknowledgments

We thank Wolfgang Stephan for insightful discussions. GA and LF were supported by grant ANR-12-JSV7-0007 TempoMut from Agence Nationale de la Recherche. GA was also supported by grant ANR-12-BSV7-0012 Demochips, AK and TW by grants of the German Science Foundation (DFG-SFB680 and DFG-SPP 1590).

Appendix A. The folded spectra

When no reliable outgroup sequence is available, one cannot assess if the allele is derived or ancestral. In that case, alleles can only be classified as minor (less frequent) and major (most frequent). The distribution of minor allele frequencies, known as the folded SFS, will be noted $\eta(f^*)$, where f^* denotes the minor allele frequency that ranges from 0 to 0.5. Importantly, the folded SFS can be retrieved from the full SFS by simply summing alleles at complementary frequencies:

$$\eta(f^*) = [\xi(f^*) + \xi(1 - f^*)] / (1 + \delta_{f^*,(1-f^*)}) \quad (A.1)$$

As a consequence, the single site SFS under the standard neutral model then becomes $E[\eta(f^*)] = \theta / [f^*(1 - f^*)(1 + \delta_{f^*,(1-f^*)})]$ and $E[\eta_{k^*}] = \theta n / [k^*(n - k^*)(1 + \delta_{k^*,n-k^*})]$, where k^* denotes the count of the minor allele.

Following the same idea, we define a conditional folded 1-SFS and a joint folded 2-SFS using the minor allele frequencies. Minor alleles can also be classified as “nested” or “disjoint” depending on the presence or absence of individuals enclosing both minor alleles. As for the unfolded case, this classification gives a complete description of the linkage between pairs of mutations. However, in contrast to the unfolded case, the classification has no strict evolutionary meaning. For example, “disjoint” minor alleles do not necessarily correspond to pairs of alleles born in different backgrounds. Moreover, alleles of frequency $f^* = 0.5$ (or allele count $k^* = n/2$) suffer from an ambiguity in the choice of the minor allele and therefore should be treated separately. Note also that with the exception of alleles with frequency 0.5, folded spectra do not contain complementary alleles, since the frequency of one of the two complementary alleles will exceed 0.5.

Pairs of mutations with f, f_0 both larger or smaller than 0.5 will be classified identically (as nested or disjoint) in the folded case. However, pairs of mutations with $f < 0.5$ and $f_0 > 0.5$ (or vice-versa) will swap their classification. As a consequence, the two components of the 2-SFS are:

$$\begin{aligned} E[\eta^{(n)}(f^*, f_0^*)] &= E[\xi^{(n)}(f^*, f_0^*)] + E[\xi^{(n)}(1 - f^*, 1 - f_0^*)] \\ &\quad + E[\xi^{(d)}(f^*, 1 - f_0^*)] \\ &\quad + E[\xi^{(d)}(1 - f^*, f_0^*)] \\ E[\eta^{(d)}(f^*, f_0^*)] &= E[\xi^{(d)}(f^*, f_0^*)] + E[\xi^{(n)}(f^*, 1 - f_0^*)] \\ &\quad + E[\xi^{(n)}(1 - f^*, f_0^*)] \end{aligned} \quad (A.2)$$

To obtain the conditional 1-SFS, we proceed similarly to the unfolded case. First we separate the 2-SFS above into components based on frequency. The strictly nested component corresponds to frequencies $f^* < f_0^*$ of the nested part, while the co-occurring

and enclosing components correspond to $f^* = f_0^*$ and $f^* > f_0^*$ respectively. The strictly disjoint component corresponds to the disjoint part, since there cannot be any complementary component. Then we divide each component by the expected 1-SFS $E[\eta(f_0^*)]$ to obtain

$$\begin{aligned} E[\eta^{(sn)}(f^* | f_0^*)] &= \frac{f_0^*(1 - f_0^*)}{\theta} E[\eta^{(n)}(f^*, f_0^*)] \quad \text{for } f^* < f_0^* \\ E[\eta^{(co)}(f^* | f_0^*)] &= 2 \cdot \frac{f_0^*(1 - f_0^*)}{\theta} E[\eta^{(n)}(f^*, f_0^*)] \quad \text{for } f^* = f_0^* \\ E[\eta^{(en)}(f^* | f_0^*)] &= \frac{f_0^*(1 - f_0^*)}{\theta} E[\eta^{(n)}(f^*, f_0^*)] \quad \text{for } f^* > f_0^* \\ E[\eta^{(cm)}(f^* | f_0^*)] &= 0 \\ E[\eta^{(sd)}(f^* | f_0^*)] &= (1 + \delta_{f^*,f_0^*}) \cdot \frac{f_0^*(1 - f_0^*)}{\theta} E[\eta^{(d)}(f^*, f_0^*)] \end{aligned} \quad (A.3)$$

While the classification of the pairs with frequencies $f^* = 0.5$ and/or $f_0^* = 0.5$ is ambiguous, these pairs are usually irrelevant for the population spectrum.

The sample spectra are similar. For n even, there are ambiguous pairs with k or $l = n/2$ that can be easily retrieved from Eqs. (11),(13) and treated separately. Considering only $k, l < n/2$, the sample 2-SFS is:

$$\begin{aligned} E[\eta_{k^*,l^*}^{(n)}] &= E[\xi_{k^*,l^*}^{(n)}] + E[\xi_{n-k^*,n-l^*}^{(n)}] + E[\xi_{k^*,n-l^*}^{(d)}] + E[\xi_{n-k^*,l^*}^{(d)}] \\ E[\eta_{k^*,l^*}^{(d)}] &= E[\xi_{k^*,l^*}^{(d)}] + E[\xi_{k^*,n-l^*}^{(n)}] + E[\xi_{n-k^*,l^*}^{(n)}] \end{aligned} \quad (A.4)$$

and the conditional 1-SFS is:

$$\begin{aligned} E[\eta_{k^*|l^*}^{(sn)}] &= \frac{l^*(n - l^*)}{\theta n} E[\eta_{k^*,l^*}^{(n)}] \quad \text{for } k^* < l^* \\ E[\eta_{k^*|l^*}^{(co)}] &= 2 \cdot \frac{l^*(n - l^*)}{\theta n} E[\eta_{k^*,l^*}^{(n)}] \quad \text{for } k^* = l^* \\ E[\eta_{k^*|l^*}^{(en)}] &= \frac{l^*(n - l^*)}{\theta n} E[\eta_{k^*,l^*}^{(n)}] \quad \text{for } k^* > l^* \\ E[\eta_{k^*|l^*}^{(cm)}] &= 0 \\ E[\eta_{k^*|l^*}^{(sd)}] &= (1 + \delta_{k^*,l^*}) \cdot \frac{l^*(n - l^*)}{\theta n} E[\eta_{k^*,l^*}^{(d)}] \end{aligned} \quad (A.5)$$

Appendix B. Supplementary data

Supplementary material related to this article can be found online at <https://doi.org/10.1016/j.tpb.2018.06.001>.

References

Achaz, G., 2009. Frequency spectrum neutrality tests: one for all and all for one. *Genetics* 183 (1), 249–258.

Alcala, N., Jensen, J.D., Telenti, A., Vuilleumier, S., 2016. The genomic signature of population reconnection following isolation: From theory to HIV. *G3: Genes–Genomes–Genet.* 6 (1), 107–120.

Birkner, M., Blath, J., Eldon, B., 2013. Statistical properties of the site-frequency spectrum associated with lambda-coalescents. *Genetics*, genetics–113.

Bustamante, C.D., Nielsen, R., Sawyer, S.A., Olsen, K.M., Purugganan, M.D., Hartl, D.L., 2002. The cost of inbreeding in Arabidopsis. *Nature* 416 (6880), 531–534.

Bustamante, C.D., Wakeley, J., Sawyer, S., Hartl, D.L., 2001. Directional selection and the site-frequency spectrum. *Genetics* 159 (4), 1779–1788.

Ethier, S., Griffiths, R., 1990. On the two-locus sampling distribution. *J. Math. Biol.* 29 (2), 131–159.

Ewens, W.J., 2012. *Mathematical Population Genetics 1: Theoretical Introduction*, vol. 27. Springer.

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

Ferretti, L., Ledda, A., Wiehe, T., Achaz, G., Ramos-Onsins, S.E., 2017. Decomposing the Site Frequency Spectrum: the impact of tree topology on neutrality tests. *Genetics*, genetics–116.

Ferretti, L., Perez-Enciso, M., Ramos-Onsins, S., 2010. Optimal neutrality tests based on the frequency spectrum. *Genetics* 186 (1), 353–365.

Ferretti, L., Raineri, E., Ramos-Onsins, S., 2012. Neutrality tests for sequences with missing data. *Genetics* 191 (4), 1397–1401.

- Fu, Y.-X., 1995. Statistical properties of segregating sites. *Theor. Popul. Biol.* 48 (2), 172–197.
- Fu, Y.X., Li, W.H., 1993. Statistical tests of neutrality of mutations. *Genetics* 133 (3), 693–709.
- Golding, G.B., 1984. The sampling distribution of linkage disequilibrium. *Genetics* 108 (1), 257–274.
- Griffiths, R., 1979. A transition density expansion for a multi-allele diffusion model. *Adv. Appl. Probab.* 310–325.
- Griffiths, R., Tavaré, S., 2003. The genealogy of a neutral mutation. *Oxf. Stat. Sci. Ser.* 393–413.
- Griffiths, R.C., Marjoram, P., 1997. An ancestral recombination graph. *Inst. Math. Appl.* 87, 257.
- Griffiths, R.C., Tavaré, S., 1994. Sampling theory for neutral alleles in a varying environment. *Philos. Trans. R. Soc. Lond. B Biol. Sci.* 344 (1310), 403–410.
- Gutenkunst, R.N., Hernandez, R.D., Williamson, S.H., Bustamante, C.D., 2009. Inferring the joint demographic history of multiple populations from multidimensional snp frequency data. *PLoS Genet.* 5 (10), e1000695.
- Hobolth, A., Siren, J., 2016. The multivariate Wright–Fisher process with mutation: Moment-based analysis and inference using a hierarchical beta model. *Theor. Popul. Biol.* 108, 36–50.
- Hobolth, A., Wiuf, C., 2009. The genealogy, site frequency spectrum and ages of two nested mutant alleles. *Theor. Popul. Biol.* 75 (4), 260–265.
- Hudson, R.R., 2001. Two-locus sampling distributions and their application. *Genetics* 159 (4), 1805–1817.
- Hudson, R.R., 2002. Generating samples under a Wright–Fisher neutral model of genetic variation. *Bioinformatics* 18 (2), 337–338.
- Hudson, R.R., et al., 1990. Gene genealogies and the coalescent process. *Oxf. Surv. Evol. Biol.* 7 (1), 44.
- Jenkins, P.A., Mueller, J.W., Song, Y.S., 2014. General triallelic frequency spectrum under demographic models with variable population size. *Genetics* 196 (1), 295–311.
- Jenkins, P.A., Song, Y.S., 2011. The effect of recurrent mutation on the frequency spectrum of a segregating site and the age of an allele. *Theor. Popul. Biol.* 80 (2), 158–173.
- Jenkins, P.A., Song, Y.S., 2012. Padé approximants and exact two-locus sampling distributions. *Ann. Appl. Probab.* 22 (2), 576–607.
- Kim, Y., Nielsen, R., 2004. Linkage disequilibrium as a signature of selective sweeps. *Genetics* 167 (3), 1513–1524.
- Kim, Y., Stephan, W., 2002. Detecting a local signature of genetic hitchhiking along a recombining chromosome. *Genetics* 160 (2), 765–777.
- Kimura, M., 1956. Random genetic drift in a tri-allelic locus; exact solution with a continuous model. *Biometrics* 12 (1), 57–66.
- Kimura, M., 1983. *The Neutral Theory of Molecular Evolution*. Cambridge University Press, Cambridge, Great Britain.
- Klassmann, A., Ferretti, L., 2018. The third moments of the site frequency spectrum. *Theor. Popul. Biol.*
- Lachance, J., Tishkoff, S.A., 2013. SNP ascertainment bias in population genetic analyses: why it is important, and how to correct it. *Bioessays* 35 (9), 780–786.
- Li, H., Stephan, W., 2005. Maximum-likelihood methods for detecting recent positive selection and localizing the selected site in the genome. *Genetics* 171 (1), 377–384.
- Littler, R., Fackerell, E., 1975. Transition densities for neutral multi-allele diffusion models. *Biometrics* 117–123.
- Liu, X., Fu, Y.-X., 2015. Exploring population size changes using SNP frequency spectra. *Nature Genet.* 47 (5), 555–559.
- Nielsen, R., Williamson, S., Kim, Y., Hubisz, M.J., Clark, A.G., Bustamante, C., 2005. Genomic scans for selective sweeps using SNP data. *Genome Res.* 15 (11), 1566–1575.
- Rafajlović, M., Klassmann, A., Eriksson, A., Wiehe, T., Mehlig, B., 2014. Demography-adjusted tests of neutrality based on genome-wide SNP data. *Theor. Popul. Biol.* 95, 1–12.
- Ragsdale, A.P., Gutenkunst, R.N., 2017. Inferring demographic history using two-locus statistics. *Genetics* 206 (2), 1037–1048.
- Sargsyan, O., 2015. An analytical framework in the general coalescent tree setting for analyzing polymorphisms created by two mutations. *J. Math. Biol.* 70 (4), 913–956.
- Sawyer, S.A., Hartl, D.L., 1992. Population genetics of polymorphism and divergence. *Genetics* 132 (4), 1161–1176.
- Tajima, F., 1983. Evolutionary relationship of dna sequences in finite populations. *Genetics* 105 (2), 437–460.
- Tajima, F., 1989. Statistical method for testing the neutral mutation hypothesis by DNA polymorphism. *Genetics* 123 (3), 585–595.
- Thornton, K., 2005. Recombination and the properties of Tajima's D in the context of approximate-likelihood calculation. *Genetics* 171 (4), 2143–2148.
- Watterson, G., 1975. On the number of segregating sites in genetical models without recombination. *Theor. Popul. Biol.* 7 (2), 256–276.
- Wright, S., 1938. The distribution of gene frequencies under irreversible mutation. *Proc. Natl. Acad. Sci. USA* 24 (7), 253.
- Xie, X., 2011. The site-frequency spectrum of linked sites. *Bull. Math. Biol.* 73 (3), 459–494.
- Živković, D., Wiehe, T., 2008. Second-order moments of segregating sites under variable population size. *Genetics* 180 (1), 341–357.



The third moments of the site frequency spectrum

A. Klassmann^{a,*}, L. Ferretti^b

^a Institut für Genetik, Universität zu Köln, 50674 Köln, Germany

^b The Pirbright Institute, Woking, United Kingdom



ARTICLE INFO

Article history:

Received 20 February 2017

Available online 5 January 2018

Keywords:

Single nucleotide polymorphisms

Infinite-sites model

Site frequency spectrum

Coalescent approximation

Nested mutations

Skewness

ABSTRACT

The analysis of patterns of segregating (i.e. polymorphic) sites in aligned sequences is routine in population genetics. Quantities of interest include the total number of segregating sites and the number of sites with mutations of different frequencies, the so-called *site frequency spectrum*. For neutrally evolving sequences, some classical results are available, including the expected value and variance of the spectrum in the Kingman coalescent model without recombination as calculated by Fu (1995).

In this work, we use similar techniques to compute the third moments of the frequencies of three linked sites. Based on these results, we derive analytical results for the bias of Tajima's *D* and other neutrality tests.

As a corollary, we obtain the second moments of the frequencies of two linked mutations conditional on the presence of a third mutation with a certain frequency. These moments can be used for the normalisation of new neutrality tests relying on these spectra.

© 2017 The Authors. Published by Elsevier Inc. This is an open access article under the CC BY license (<http://creativecommons.org/licenses/by/4.0/>).

1. Introduction

Statistics based on polymorphic loci are key to estimate relevant quantities in population genetics, such as the rescaled mutation rate θ . One common approach is to group together variants that appear with the same frequency in a sample and count the elements of each such group. The resulting summary statistic is called the *site frequency spectrum*.

The frequency spectrum is one of the most relevant statistics for population genetics. It can be used to infer evolutionary parameters such as mutation and recombination rate, past population history, demography and selection (Hudson, 1983; Nielsen et al., 2005; Hein et al., 2004). Often, the variants are biallelic SNPs that can be “polarized”, i.e. it is possible to say which allele is ancestral and which one is derived. This is the case for sequences with low mutation rate per base and for which an outgroup sequence is available. In what follows, we will consider exclusively this situation and assume that the evolution of these sequences can be modelled by a standard neutral Wright–Fisher model of constant population size.

Watterson (1975) credits Fisher (1930) with the first derivation (for a special case) of the first moments of the frequency spectrum. The derivation for the continuous analogue can be found in Ewens (1979), where it follows from results of diffusion theory (Kimura, 1964). Watterson (1975) himself derived the first and second moments for the sum over all classes of the frequency

spectrum, i.e. the number of segregating sites, using the technique of “moment estimators”. The full distribution of this quantity was shown by Tavaré (1984, Eq. (9.5)). The first and second moments for combinations of some components of the spectrum were later computed by Tajima (1989) using coalescent theory (Kingman, 1982) and combinatorics, while Fu (1995) completed this approach for the full frequency spectrum. A major application of his formulae is the normalisation of a class of neutrality tests such as Tajima's *D* (Tajima, 1989), as described by Achaz (2009). Recently, Hudson (2015) has given another proof of the first moments. As far as we know, higher moments of the spectrum have never been computed.

Asymptotic results for the distribution of the spectrum have been obtained by Dahmer and Kersting (2015). However, their method applies only to mutations of size less than or equal to a fixed number k in the limit of $n \rightarrow \infty$, i.e. to mutations of infinitesimal frequency $f \leq k/n \rightarrow 0$. Hence, their approach does not provide information on the full frequency spectrum in finite samples.

In this article we derive exact expressions for the third moments of the frequency spectrum. We use notation and approach of Fu (1995), with some technical modifications in order to keep the number of different cases manageable. As a by-product we state the third moment of the number of segregating sites. An immediate corollary of the third moments is the expected frequency spectrum for three linked segregating sites, which fully characterises the expected haplotype structure for triplets of sites.

We discuss the consequences of these results for the distribution of several neutrality tests that are constructed similarly to Tajima's *D* (Tajima, 1989). These tests have been designed to

* Corresponding author.

E-mail address: alexander.klassmann@uni-koeln.de (A. Klassmann).

yield under neutrality an expected value of approximately zero, but since they do not exactly so, they are biased (Tajima, 1989; Simonson et al., 1995). For the first time, we obtain general expressions for bias and skewness of these tests as a function of mutation rate and sample size.

Finally, we derive the variance of the frequency spectra of two nested or disjoint mutations linked to a third mutation of a certain size. These spectra can be used to describe neutrally evolving structural variants such as chromosomal inversions (Ferretti et al., 2017). With our results, it is possible to obtain the proper normalisation for new Tajima’s D -like tests relying on such spectra.

In the next section we state our main result and several implications. The corresponding proofs are presented largely in the subsequent section, while the combinatorial parts are deferred to the supplement.

2. Results

As is common practise in coalescent theory, we define θ as the population-scaled mutation rate per sequence, i.e. $\theta = 2pN_e\mu L$ where p is the ploidy, N_e is the effective population size, μ is the mutation rate per generation per bp and L is the length of the sequence in base pairs. We consider a sample of n sequences with $n \ll N_e$. We assume that we can distinguish between ancestral and derived alleles. A mutation (alias derived allele) is said to have size i , if i sequences of the sample carry it. The number of mutations of size i within the sample is referred to as ξ_i . The tuple ξ_1, \dots, ξ_{n-1} forms the frequency spectrum.

The model that we consider is the Kingman coalescent, with an infinite-sites model of mutations. We assume no recombination, i.e. complete linkage among sites.

2.1. The third moments of the frequency spectrum

Our main result is an analytical expression for the third moments of the frequency spectrum.

Theorem 2.1. *In the infinite sites approximation for biallelic sequences without recombination, the third moments of the frequency spectrum can be expressed as*

$$E[\xi_h \xi_i \xi_j] = \delta_{h=i=j} \tau_i \theta + (\delta_{h=i} \tau_{ij} + \delta_{i=j} \tau_{hi} + \delta_{j=h} \tau_{ij}) \theta^2 + \tau_{hij} \theta^3 \quad (1)$$

for $1 \leq h, i, j < n$. The functions τ are:

$$\tau_i = \frac{1}{i}, \quad (2)$$

$$\tau_{ij} = t_a(i, j) + t_a(j, i) + t_b(i, j) + t_b(j, i) \quad (3)$$

with

$$t_a(i, j) = \begin{cases} \frac{1}{2} (\beta_n(j) - \beta_n(j + 1)) & \text{if } j < i \\ \frac{1}{2} \beta_n(j) & \text{if } j = i \end{cases} \quad (4)$$

$$t_b(i, j) = \begin{cases} \frac{1}{ij} - \frac{1}{i(i+j)} - \frac{1}{2} (\beta_n(j) - \beta_n(j + 1)) & \text{if } i + j < n \\ \alpha_n(j) - \frac{1}{2} \beta_n(j) & \text{if } i + j = n, \end{cases}$$

and¹

$$\tau_{hij} = \sum_{\text{Permutations}(h,i,j)} t_{aa}(h, i, j) + t_{ab}(h, i, j) + t_{ba}(h, i, j) + t_{bb}(h, i, j) \quad (5)$$

¹ $\sum_{\text{Perm.}(h,i,j)} f(h, i, j) = f(h, i, j) + f(i, j, h) + f(j, h, i) + f(h, j, i) + f(i, h, j) + f(j, i, h)$.

with Eqs. (6) given in Box I using the following auxiliary functions:

$$\alpha_n(i) = \frac{1}{\binom{n-1}{i}} \sum_{k=2}^n \frac{\binom{n-k}{i-1}}{k-1}$$

$$\beta_n(i) = \frac{2}{\binom{n-1}{i}} \sum_{k=2}^n \frac{\binom{n-k}{i-1}}{k}$$

$$\alpha_n^{(2)}(i, j) = \sum_{k=2}^n \sum_{t=1}^{k-1} \frac{\binom{i-1}{t-1} \binom{n-i-j}{k-t-1}}{\binom{n-1}{k-1}} \frac{1}{k(k-1)} \alpha_k(t)$$

$$\beta_n^{(2)}(i, j) = \sum_{k=2}^n \sum_{t=1}^{k-1} \frac{\binom{i-1}{t-1} \binom{n-i-j}{k-t-1}}{\binom{n-1}{k-1}} \frac{1}{k(k-1)} \frac{\beta_k(t)}{2}$$

$$\alpha_n^{(3)}(h, i, j) = (h + 1) \alpha_n^{(2)}(i, j) - 2h \alpha_n^{(2)}(i, j + 1) + (h - 1) \alpha_n^{(2)}(i, j + 2)$$

$$\beta_n^{(3)}(h, i, j) = (h + 1) \beta_n^{(2)}(i, j) - 2h \beta_n^{(2)}(i, j + 1) + (h - 1) \beta_n^{(2)}(i, j + 2)$$

$$\alpha_n^{(4)}(h, i, j) = (h + 1) \alpha_n^{(2)}(i + 1, j) - 2h \alpha_n^{(2)}(i, j + 1) + (h - 1) \alpha_n^{(2)}(i - 1, j + 2)$$

$$\beta_n^{(4)}(h, i, j) = (h + 1) \beta_n^{(2)}(i + 1, j) - 2h \beta_n^{(2)}(i, j + 1) + (h - 1) \beta_n^{(2)}(i - 1, j + 2).$$

Remark 1. The coefficient for θ is the well known result for the expectation of the frequency spectrum

$$E[\xi_i] = \tau_i \theta = \frac{\theta}{i}. \quad (8)$$

The terms τ_{ij} are identical to the quadratic part of the second moments,

$$E[\xi_i \xi_j] = \delta_{i=j} \tau_i \theta + \tau_{ij} \theta^2, \quad (9)$$

computed by Fu (1995): $\tau_{ij} = \sigma_{ij} + \frac{1}{ij}$, with σ_{ij} defined in Eqs. (2) and (3) therein.

Remark 2. Fu (1995) showed in his Eq. (34) that $\alpha_n(i)$ and $\beta_n(i)$ can be written in a more compact form, namely

$$\alpha_n(i) = \frac{H_{n-1} - H_{i-1}}{n - i}$$

$$\beta_n(i) = \frac{2n}{(n - i + 1)(n - i)} (H_n - H_{i-1}) - \frac{2}{n - i},$$

with $H_n = \sum_{i=1}^n \frac{1}{i}$. We do not have a corresponding form for $\alpha_n^{(2)}(i, j)$ and $\beta_n^{(2)}(i, j)$. We only note that in the case of “singletons” they yield (with $H_{n,2} = \sum_{k=1}^n \frac{1}{k^2}$)

$$\alpha_n^{(2)}(1, 1) = \frac{H_{n-1,2} - \frac{1}{n} H_{n-1}}{n - 1}$$

$$\beta_n^{(2)}(1, 1) = \frac{1 - \frac{1}{n} H_{n-1}}{n - 1}.$$

Remark 3. The sum over permutations simplifies the fractions in t_b resp. t_{bb} :

$$\sum_{\text{Permutations}(i,j)} \left(\frac{1}{ij} - \frac{1}{i(i+j)} \right) = \frac{1}{ij} \quad (10)$$

$$\begin{aligned}
t_{aa}(h, i, j) &= \begin{cases} \beta_n^{(4)}(i-j, i-j, j) - \beta_n^{(4)}(i-j, i-j+1, j) & \text{if } j < i \text{ and } i < h \\ \beta_n^{(4)}(i-j, i-j, j) & \text{if } j < i \text{ and } i = h \\ \beta_n^{(2)}(1, j) - \beta_n^{(2)}(2, j) & \text{if } j = i \text{ and } i < h \\ \beta_n^{(2)}(1, j) & \text{if } j = i \text{ and } i = h \end{cases} \\
t_{ab}(h, i, j) &= \begin{cases} \beta_n^{(3)}(h-i-j, i, j) - \beta_n^{(3)}(h-i-j, i+1, j) & \text{if } i+j < h \\ \beta_n^{(2)}(i, j) - \beta_n^{(2)}(i+1, j) & \text{if } i+j = h \end{cases} \\
t_{ba}(h, i, j) &= \begin{cases} \frac{1}{2h} (\beta_n(j) - \beta_n(j+1) - \beta_n(h+j) + \beta_n(h+j+1)) \\ \quad - \beta_n^{(4)}(i-j, i-j, j) + \beta_n^{(4)}(i-j, i-j+1, j) & \text{if } j < i \text{ and } h+i < n \\ \quad - \beta_n^{(3)}(i-j, h, j) + \beta_n^{(3)}(i-j, h+1, j) \\ \quad + \beta_n^{(3)}(n-h-i, j, h) - \beta_n^{(3)}(n-h-i, j+1, h) \\ \alpha_n^{(4)}(n-h-j, n-h-j, j) - \beta_n^{(4)}(n-h-j, n-h-j, j) & \text{if } j < i \text{ and } h+i = n \\ \quad + \alpha_n^{(3)}(n-h-j, h, j) - \beta_n^{(3)}(n-h-j, h, j) \\ \quad + \beta_n^{(2)}(j, h) - \beta_n^{(2)}(j+1, h) \\ \frac{1}{2h} (\beta_n(j) - \beta_n(h+j)) + \beta_n^{(3)}(n-h-j, j, h) & \text{if } j = i \text{ and } h+i < n \\ -\beta_n^{(2)}(h, j) + \beta_n^{(2)}(h+1, j) - \beta_n^{(2)}(1, j) + \beta_n^{(2)}(2, j) \\ \frac{1}{2} (\alpha_n^{(2)}(n-j, j) + \alpha_n^{(2)}(j, n-j)) & \text{if } j = i \text{ and } h+i = n \\ \quad + \alpha_n^{(2)}(1, j) - \beta_n^{(2)}(1, j) \\ t_{bb}(h, i, j) &= \begin{cases} -\frac{1}{2i} (\beta_n(j) - \beta_n(j+1) - \beta_n(i+j) + \beta_n(i+j+1)) & \text{if } h+i+j < n \\ -\beta_n^{(3)}(n-h-i-j, i, j) + \beta_n^{(3)}(n-h-i-j, i+1, j) \\ \frac{1}{i} (\alpha_n(j) - \alpha_n(i+j)) - \frac{1}{2i} (\beta_n(j) - \beta_n(i+j)) & \text{if } h+i+j = n \\ -\beta_n^{(2)}(i, j) + \beta_n^{(2)}(i+1, j) \end{cases} \end{cases} \quad (6)
\end{aligned}$$

Box 1.

$$\begin{aligned}
&\sum_{\text{Permutations}(h, i, j)} \left(\frac{1}{(h+i+j)(h+i)h} + \frac{1}{ij(h+i)} - \frac{1}{ih(h+j)} \right) \\
&= \frac{1}{hij}. \quad (11)
\end{aligned}$$

Remark 4. The central third moments can be obtained by

$$\begin{aligned}
\mu_3[\xi_h, \xi_i, \xi_j] &= E[(\xi_h - E[\xi_h])(\xi_i - E[\xi_i])(\xi_j - E[\xi_j])] \\
&= E[\xi_h \xi_i \xi_j] - E[\xi_h]E[\xi_i \xi_j] - E[\xi_i]E[\xi_h \xi_j] \\
&\quad - E[\xi_j]E[\xi_h \xi_i] + 2E[\xi_h]E[\xi_i]E[\xi_j]. \quad (12)
\end{aligned}$$

Remark 5. If mutations cannot be classified as either ancestral or derived, usually the minor frequency of the two alleles is taken into account to form the *folded frequency spectrum*

$$\eta_i = \frac{\xi_i + \xi_{n-i}}{1 + \delta_{i=n-i}}.$$

The corresponding third moments can be computed analogously to the second moments (Eq. (9) in Fu (1995)):

$$\begin{aligned}
E[\eta_h \eta_i \eta_j] &= (E[\xi_h \xi_i \xi_j] + E[\xi_h \xi_i \xi_{n-j}] + E[\xi_h \xi_{n-i} \xi_j] \\
&\quad + E[\xi_h \xi_{n-i} \xi_{n-j}] \\
&\quad + E[\xi_{n-h} \xi_i \xi_j] + E[\xi_{n-h} \xi_i \xi_{n-j}] + E[\xi_{n-h} \xi_{n-i} \xi_j] \\
&\quad + E[\xi_{n-h} \xi_{n-i} \xi_{n-j}]) \\
&\quad \cdot \frac{1}{(1 + \delta_{h=n-h})(1 + \delta_{i=n-i})(1 + \delta_{j=n-j})}. \quad (13)
\end{aligned}$$

2.2. The frequency spectrum of three linked sites

The components t_{aa} , t_{ab} , t_{ba} and t_{bb} correspond to different linkage patterns of three mutations (without recombination). We call a derived mutation a to be *nested within* or *nested inside* a derived mutation b , if a is present only in sequences that contain b . We call pairs of mutations simply *nested*, if one is nested within the other. If two mutations are present in non-overlapping sets of sequences, we refer to them as *disjoint*. Linked mutations are either nested or disjoint. Hence, three derived mutations can have four possible relations (see also Section 3.1.2 “Averaging over topologies”):

- *fully nested*: one mutation (of size j) is nested inside another mutation (of size i) which itself is nested inside the third mutation (of size h). This relation corresponds to $t_{aa}(h, i, j)$.
- *disjoint within nested*: two disjoint mutations (of sizes i and j , resp.) are nested within the third mutation (of size h). This relation corresponds to $t_{ab}(h, i, j) + t_{ab}(h, j, i)$.
- *nested within disjoint*: two mutations (of sizes h and i , resp.) are mutually disjoint and the third mutation (of size j) is nested inside the second. (Consequently the first and third are disjoint, too). This relation corresponds to $t_{ba}(h, i, j)$.
- *fully disjoint*: all three mutations (of sizes h , i and j , resp.) are mutually disjoint. This relation corresponds to $\sum_{\text{Permutations}(h, i, j)} t_{bb}(h, i, j)$.

Therefore, the spectrum of three sites can be easily decomposed by separating the components t_{aa} , t_{ab} , t_{ba} and t_{bb} .

The nested and disjoint components of the frequency spectrum for pairs of sites give a complete description of the haplotype

structure of two sites (up to permutations of individuals and sites) (Ferretti et al., 2017). Analogously the frequency spectrum for triplets of segregating sites is given by

$$E[\xi_{h,i,j}] = \begin{cases} E[\xi_h \xi_i \xi_j] & \text{for } h \neq i, \\ & h \neq j, i \neq j \\ E[\xi_h \xi_i (\xi_i - 1)]/2 = (E[\xi_h \xi_i^2] - E[\xi_h \xi_i])/2 & \text{for } i = j \neq h \\ E[\xi_h \xi_i (\xi_h - 1)]/2 = (E[\xi_h^2 \xi_i] - E[\xi_h \xi_i])/2 & \text{for } h = j \neq i \\ E[\xi_h \xi_j (\xi_h - 1)]/2 = (E[\xi_h^2 \xi_j] - E[\xi_h \xi_j])/2 & \text{for } h = i \neq j \\ E[\xi_h (\xi_h - 1)(\xi_h - 2)]/6 = (E[\xi_h^3] - 3E[\xi_h^2] + 2E[\xi_h])/6 & \text{for } h = i = j. \end{cases} \quad (14)$$

This spectrum is equivalent to a complete characterisation of the haplotype spectrum of three sites.

2.3. A recursion equation for nested mutations of identical size

Recursion equations are generally a useful tool to investigate branching processes (of which the coalescent is a special case), since they allow to derive properties by induction (Kimmel and Axelrod, 2015).

For a function $f(n, i)$, $n \geq 1$ and $1 < i \leq n$ we consider the following recursion equation:

$$f(n + 1, i) = (1 - \frac{i}{n})f(n, i) + \frac{i-1}{n}f(n, i-1). \quad (15)$$

This equation has been used to prove that the number of leaves descending from one of the two root branches of a coalescent tree is uniformly distributed, an important property that helps to understand some patterns of variation in molecular sequences (Tajima, 1983, Eq. (2)), (Wakeley, 2008, Eq. (3.36)). Likewise, Hudson (2015) exploited the fact that τ_i fulfils (15) to show that the first moment of the frequency spectrum can be obtained by an inductive proof over the sample size n .

We show that the recursion equation holds, too, for the second and third moments of nested mutations of the same size:

Proposition 2.1. $\alpha_n(i)$, $\beta_n(i)$, $\alpha_n^{(2)}(1, i)$, $\beta_n^{(2)}(1, i)$ and, consequently, $t_a(i, i)$ and $t_{aa}(i, i)$, regarded as functions of n and i , fulfil Eq. (15).

2.4. The third moments of the number of segregating sites

The number of segregating sites is given by $S = \sum_{i=1}^{n-1} \xi_i$. Although various expressions for its complete distribution are known (Wakeley, 2008, Eqs. (3.32)–(3.34), (4.3)), it is not obvious how to derive individual moments from them. We will prove the following theorem in the same way as Theorem 2.1.

Theorem 2.2. Writing $H_{n,m} = \sum_{i=1}^n \frac{1}{i^m}$ for the n th harmonic number of order m , the third moment (resp. central moment) of the number of segregating sites S for a sample of size n yields:

$$E[S^3] = H_{n-1,1}\theta + 3(H_{n-1,1}^2\theta + H_{n-1,2})\theta^2 + (H_{n-1,1}^3 + 3H_{n-1,1}H_{n-1,2} + 2H_{n-1,3})\theta^3 \quad (16)$$

$$\mu_3[S] = E[(S - E[S])^3] = H_{n-1,1}\theta + 3H_{n-1,2}\theta^2 + 2H_{n-1,3}\theta^3.$$

Since

$$\sum_{h=1}^{n-1} \sum_{i=1}^{n-1} \sum_{j=1}^{n-1} E[\xi_h \xi_i \xi_j] = E[S^3]$$

and

$$\sum_{h=1}^{n-1} \sum_{i=1}^{n-1} \sum_{j=1}^{n-1} \mu_3[\xi_h, \xi_i, \xi_j] = \mu_3[S],$$

Table 1
Weights and references of the analysed neutrality tests.

Test	Weights Ω_i	Reference
$D_{(Tajima)}$	$(n-i)/\binom{n}{2} - 1/ia_n$	Tajima (1989)
$D_{(Fu\&Li)}$	$1/ia_n - \delta_{i,1}$	Fu and Li (1993)
$F_{(Fu\&Li)}$	$(n-i) - \delta_{i,1}$	Fu and Li (1993)
$H_{(Fay\&Wu)}$	$(n-2i)/\binom{n}{2}$	Fay and Wu (2000)
$E_{(Zeng)}$	$1/(n-1) - 1/ia_n$	Zeng et al. (2006)

the coefficients for θ , θ^2 and θ^3 derived from Theorems 2.1 and 2.2 have to be the same, yielding

Corollary 2.1. The following identities hold for the functions τ_i , τ_{ij} and τ_{hij} defined in Theorem 2.1:

$$\sum_{i=1}^{n-1} \tau_i = H_{n-1} \quad (17)$$

$$\sum_{i=1}^{n-1} \sum_{j=1}^{n-1} \tau_{ij} = H_{n-1,1}^2 + H_{n-1,2} \quad (18)$$

$$\sum_{h=1}^{n-1} \sum_{i=1}^{n-1} \sum_{j=1}^{n-1} \tau_{hij} = H_{n-1,1}^3 + 3H_{n-1,1}H_{n-1,2} + 2H_{n-1,3}. \quad (19)$$

While Eq. (17) holds trivially, we give in the supplement explicit proofs of Eqs. (18) and (19) as a consistency check for Theorem 2.1.

2.5. Skewness and bias of Tajima's D and similar neutrality tests

One of the applications of the frequency spectrum is to test if the observed patterns in sequences are compatible with neutral evolutionary models. Several neutrality tests like e.g. Tajima's D (Tajima, 1989), fall into a general class that relies on normalised linear combinations of the frequency spectrum (Achaz, 2009; Ferretti et al., 2010). Their general form is

$$T_\Omega = \frac{\sum_{i=1}^{n-1} i\Omega_i \xi_i}{\sqrt{\text{Var}[\sum_{i=1}^{n-1} i\Omega_i \xi_i]}} \quad , \quad \sum_{i=1}^{n-1} \Omega_i = 0 \quad (20)$$

where the variance in the denominator

$$\text{Var} \left[\sum_{i=1}^{n-1} i\Omega_i \xi_i \right] = \theta \sum_{i=1}^{n-1} i^2 \Omega_i^2 \tau_i + \theta^2 \sum_{i,j=1}^{n-1} ij\Omega_i \Omega_j \left(\tau_{ij} - \frac{1}{ij} \right)$$

is a linear combination of θ and θ^2 . These two quantities, if unknown, are usually estimated from S and S^2 by the method of moments: $\hat{\theta} = S/H_{n-1,1}$ and $\hat{\theta}^2 = S(S-1)/(H_{n-1,1}^2 + H_{n-1,2})$. The weights Ω_i for some commonly used neutrality tests are given in Table 1.

In this section, we explore the additional information that the third moments of the spectrum reveal about the distribution of neutrality tests with respect to their skewness and bias.

It is well known that the distributions of these tests tend to be biased and skewed (Tajima, 1989; Hudson, 1991; Simonsen et al., 1995; Rafajlović et al., 2014). First, let us assume that θ is known. In this case these tests are normalised to mean 0 and variance 1 under the neutral coalescent with constant population size: $E[T_\Omega] = 0$ and $\text{Var}[T_\Omega] = 1$. Consequently they are not biased and the skewness $\gamma = \mu_3/\sigma^3$ equals the third moment of the test statistic:

$$\gamma(T_\Omega) = E[T_\Omega^3] = \frac{\sum_{i=1}^{n-1} \sum_{j=1}^{n-1} \sum_{k=1}^{n-1} ijk\Omega_i \Omega_j \Omega_k \cdot E[\xi_i \xi_j \xi_k]}{\text{Var}[\sum_{i=1}^{n-1} i\Omega_i \xi_i]^{3/2}}. \quad (21)$$

In Fig. 1 we compare Eq. (21) with values obtained by standard neutral coalescent simulations with 'ms' (Hudson, 2002) for two

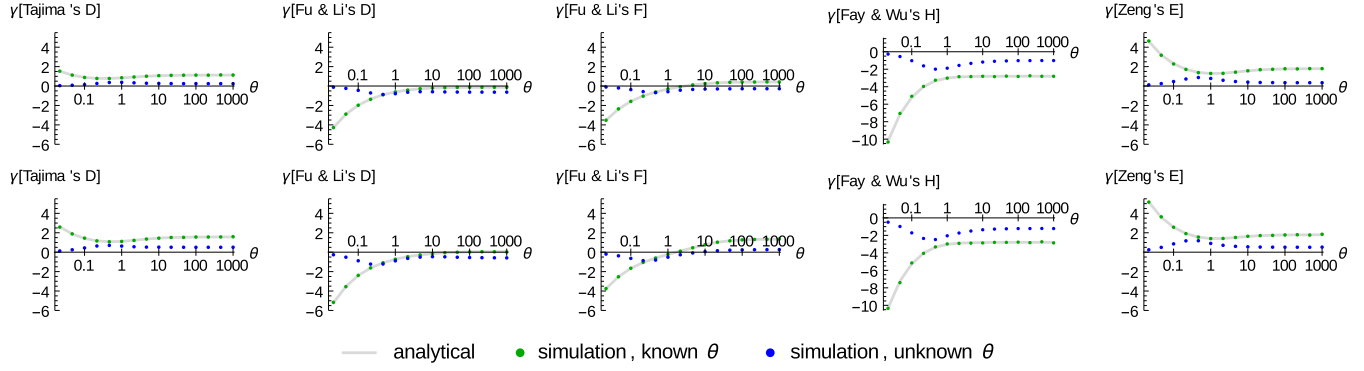


Fig. 1. Skewness of neutrality tests given in Table 1 for sample size $n = 50$ (top) and $n = 500$ (bottom). The analytical skewness was obtained by Eq. (21). For simulations, the skewness was estimated by $\hat{\gamma} = \frac{\frac{1}{2} \sum_i (x_i - \bar{x})^3}{\left(\frac{1}{n-1} \sum_i (x_i - \bar{x})^2\right)^{3/2}}$ over 10^6 genealogies (x_i is the value of the test statistic for a single genealogy). The test statistics were calculated using the true θ (green points) and Wattersons estimator $\hat{\theta} = \frac{S}{H_{n-1}}$ (blue points), respectively. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

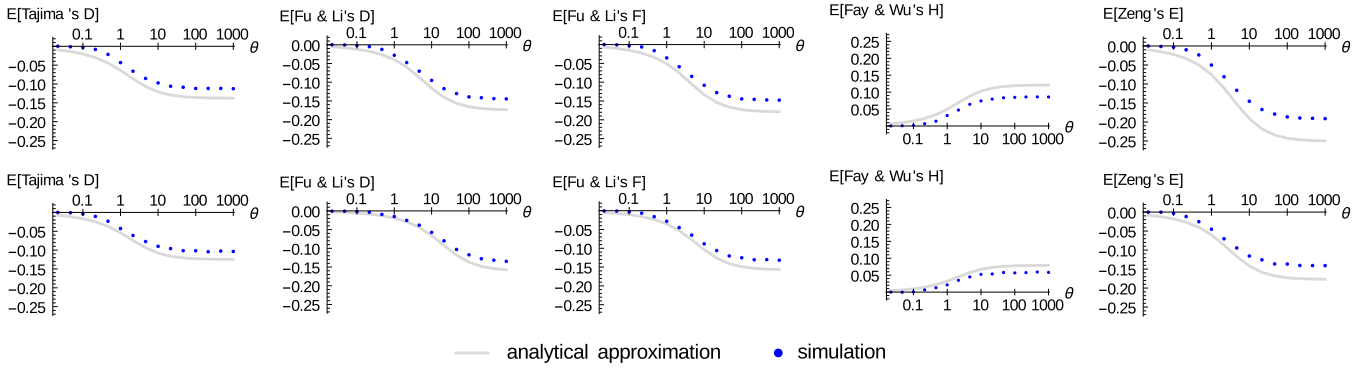


Fig. 2. The expected value of the test statistics given in Table 1 with sample size $n = 50$ (top) and $n = 500$ (bottom). The deviation from zero is the bias of the tests. Shown are our analytical approximation Eq. (22) and values, obtained by simulation with 'ms', averaged over 10^6 genealogies.

$$E[T_{\Omega}] \approx - \frac{\sum_{k=1}^{n-1} k \Omega_k \left[\frac{E[\xi_k S]}{H_{n-1,1}} \sum_{i=1}^{n-1} i^2 \Omega_i^2 \tau_i + \frac{E[\xi_k S(S-1)]}{H_{n-1,1}^2 + H_{n-1,2}} \sum_{i,j=1}^{n-1} ij \Omega_i \Omega_j \left(\tau_{ij} - \frac{1}{ij} \right) \right]}{2 \left[\theta \sum_{i=1}^{n-1} i^2 \Omega_i^2 \tau_i + \theta^2 \sum_{i,j=1}^{n-1} ij \Omega_i \Omega_j \left(\tau_{ij} - \frac{1}{ij} \right) \right]^{3/2}}, \quad (22)$$

Box II.

sample sizes and a broad range of θ -values. For known θ they agree perfectly. All test statistics show the biggest skew for very small θ , while they approach a constant value for $\theta > 10$. In practise, however, the parameter θ usually has to be estimated from the data and the denominator in Eq. (21) being a function of the estimator, contributes to the skewness. The figure shows that this has a relatively large effect, but surprisingly for most considered values of θ it reduces the skewness.

For θ unknown and estimated from S , we can still make use of the third moments. In this case, we can compute an approximation for the bias of the test statistic. We apply the following formula for the Taylor expansion of moments of random variables² X, Y with $E[X] = 0$ and $Y > 0$ almost surely

² From the general expansion (e.g. Van Erp and Van Gelder (2007))

$$E \left[\frac{X}{\sqrt{Y}} \right] \approx \frac{E[X]}{\sqrt{E[Y]}} - \frac{E[XY] - E[X]E[Y]}{2E[Y]^{3/2}} + \frac{3E[X]\text{Var}[Y]}{8E[Y]^{5/2}}.$$

$$E \left[\frac{X}{\sqrt{Y}} \right] \approx - \frac{E[XY]}{2E[Y]^{3/2}}$$

and the fact that $E[\sum_{k=1}^{n-1} k \Omega_k \xi_k] = 0$ to obtain the bias given by Eq. (22) in Box II with $E[\xi_k S] = \sum_{i=1}^{n-1} E[\xi_k \xi_i]$ resp. $E[\xi_k S^2] = \sum_{i,j=1}^{n-1} E[\xi_k \xi_i \xi_j]$.

In Fig. 2 we depict our approximative analytical result (22) together with estimations of the bias from standard coalescent simulations (Hudson, 2002), using the same two sample sizes and range of θ -values as for Fig. 1. Eq. (22) gives a reasonably good approximation of the bias of the test statistics, taking into account that it represents only the first term of a bivariate Taylor expansion.

2.6. The variance of the frequency spectrum of linked sites

We will use the nomenclature introduced by Sargsyan (2015) and expanded in Ferretti et al. (2017). We call a certain mutation of interest *focal* and we refer to it as ϕ . As in Section 2.2, further mutations that appear in at least one individual together with it,

are called *nested* while all others are called *disjoint*. More specifically, we refer to the number of mutations of size i that are nested with the focal mutation by $\xi_{i,\phi}^{(n)}$ and to those that are disjoint by $\xi_{i,\phi}^{(d)}$. Evidently, the number of overall occurrences of mutations of size i , given ϕ , is $\xi_{i,\phi} = \xi_{i,\phi}^{(n)} + \xi_{i,\phi}^{(d)}$. We now condition on the focal mutation ϕ being a mutation of size h and write $\xi_{i|h}^{(n)}$ for the number of mutations of size i nested with a mutation of size h (still excluding the focal mutation itself) and $\xi_{i|h}^{(d)}$ correspondingly for disjoint mutations. The expectation value of the number of mutations of size i conditional on the presence of a different mutation of size h is given by

$$E[\xi_{i|h}] = E[\xi_{i|h}^{(n)}] + E[\xi_{i|h}^{(d)}] = E[\xi_{i,h}^{(n)}]/E[\xi_h] + E[\xi_{i,h}^{(d)}]/E[\xi_h]. \quad (23)$$

The summands on the right side of Eq. (23) can be calculated directly from the results of Fu (1995) and are given in Ferretti et al. (2017). From Theorem 2.1, allowing for the interpretation of the terms t_{xx} given in Section 2.2, follows

Corollary 2.2. *Conditional on a mutation of size h , the second moments of two further mutations of sizes i and j are given by*

$$E[\xi_{i|h}\xi_{j|h}] = E[\xi_{i|h}^{(n)}\xi_{j|h}^{(n)}] + E[\xi_{i|h}^{(n)}\xi_{j|h}^{(d)}] + E[\xi_{i|h}^{(d)}\xi_{j|h}^{(n)}] + E[\xi_{i|h}^{(d)}\xi_{j|h}^{(d)}], \quad (24)$$

where the summands of Eq. (24) correspond to the conditional second moments of mutations of sizes i and j both nested in a mutation of size h , one nested and one disjoint and both disjoint, respectively, given by

$$E[\xi_{i|h}^{(n)}\xi_{j|h}^{(n)}] = h \left(\delta_{i=j}t_a(h, i)\theta + \left(t_{ab}(h, i, j) + t_{ba}(h, j, i) + \sum_{\text{Permutations}(h,i,j)} t_{aa}(h, i, j) \right) \theta^2 \right)$$

$$E[\xi_{i|h}^{(n)}\xi_{j|h}^{(d)}] = h (t_{ab}(i, j, h) + t_{ba}(i, h, j) + t_{ba}(j, h, i) + t_{ba}(j, i, h)) \theta^2$$

$$E[\xi_{i|h}^{(d)}\xi_{j|h}^{(n)}] = h (t_{ab}(j, i, h) + t_{ab}(j, h, i) + t_{ba}(i, h, j) + t_{ba}(i, j, h)) \theta^2$$

$$E[\xi_{i|h}^{(d)}\xi_{j|h}^{(d)}] = h \left(\delta_{i=j}t_b(h, i)\theta + \left(t_{ba}(h, i, j) + t_{ba}(h, j, i) + \sum_{\text{Permutations}(h,i,j)} t_{bb}(h, i, j) \right) \theta^2 \right).$$

2.7. Numerical results

In Fig. 3 we compare the analytical results of our Theorem 2.1 with numerical results obtained from coalescent simulations. We use “ms” (Hudson, 2002) to generate sample sequences and from their frequency spectra we calculate estimates of the third moments. For increasing sample size n the “off-diagonal” elements of the three-dimensional array of third moments, i.e. the values $E[\xi_h\xi_i\xi_j]$ for $h \neq i \neq j$, get increasingly small; as a consequence the relative difference between analytical and simulated values is largest for these elements, causing the maximum relative difference over all elements to increase with n . The graphs clearly

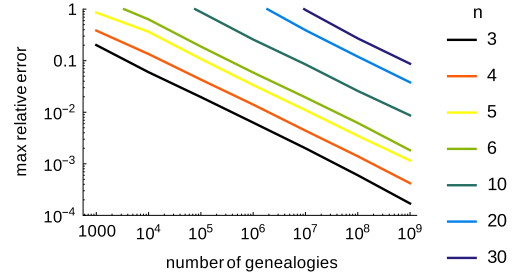


Fig. 3. The relative error between the third moments of the frequency spectrum obtained by Eq. (1) and values resulting from standard coalescent simulations with “ms” (Hudson, 2002). We computed relative errors $e = \max_{h,i,j} \frac{|E[\xi_h\xi_i\xi_j] - \bar{\xi}_h\bar{\xi}_i\bar{\xi}_j|}{E[\xi_h\xi_i\xi_j]}$ where each $\bar{\xi}_h\bar{\xi}_i\bar{\xi}_j$ represents the average of $\xi_h\xi_i\xi_j$ over 10^3 til 10^9 simulated genealogies. The figure shows the average over 100 of these relative errors e . The colours indicate different sample sizes n . $\theta = 1$. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

show that with increasing number of simulated genealogies, their average values converge to our analytical results.

Fig. 4 shows all third moments of the frequency spectrum as given in Theorem 2.1 for a sample of size $n = 5$. As in the two-dimensional case, the values of the diagonals (where two or more indices are either equal or sum to n) dominate.

In Fig. 5 we compare the covariances $Cov[\xi_i\xi_j]$ of the unconditional frequency spectrum for a sample of size $n = 10$ with the covariances between nested and disjoint mutations in a sample of size $n = 20$, conditioned on the presence of a mutation of size $k = 10$. The spectra of nested, resp. disjoint, sites are still dominated by the variances, while the correlation of “mirror sites” (ξ_i and ξ_{n-i} in the unconditional spectrum), is lost. There is almost no correlation between nested and disjoint sites.

2.8. Comparison with asymptotic analytical results

Dahmer and Kersting (2015) showed the convergence of the distribution of the components of the spectrum to centred and rescaled i.i.d. Gaussian variables in the large n limit. More precisely, they state that for fixed k , the total lengths l_i of branches with i descendants are asymptotically independent and normally distributed:

$$\sqrt{\frac{n}{\ln(n)}} \left(l_1 - 1, l_2 - \frac{1}{2}, \dots, l_k - \frac{1}{k} \right) \xrightarrow[n \rightarrow \infty]{} N(0, \mathbf{1}_{k \times k}). \quad (26)$$

For given lengths l_i , each component of the spectrum ξ_i has an independent Poisson distribution with parameter θl_i . For large θ , i.e. ignoring the Poisson noise, we have that $\xi_i = \theta l_i + O(\sqrt{\theta})$, hence the above equation becomes for $\theta \rightarrow +\infty$

$$\sqrt{\frac{n}{\ln(n)}} \lim_{\theta \rightarrow +\infty} \left(\frac{\xi_1}{\theta} - 1, \frac{\xi_2}{\theta} - \frac{1}{2}, \dots, \frac{\xi_k}{\theta} - \frac{1}{k} \right) \xrightarrow[n \rightarrow \infty]{} N(0, \mathbf{1}_{k \times k}). \quad (27)$$

In the limit of large n , the l_j can be roughly treated as independent Gaussian random variables with mean $1/j$ and variance $\ln(n)/n$, and similarly the ξ_j can be treated as independent random variables with mean θ/j and variance $\theta/j + \theta^2 \ln(n)/n$. If all moments would be uniformly bounded, this would yield the approximation

$$E[\xi_h\xi_i\xi_j]_{\theta^3} = \tau_{hij}\theta^3 = E[\xi_h]E[\xi_i]E[\xi_j] + \frac{\ln(n)}{n}(\delta_{h=i}E[\xi_j] + \delta_{h=j}E[\xi_i] + \delta_{i=j}E[\xi_h])\theta^2 + o\left(\frac{\ln(n)}{n}\right).$$

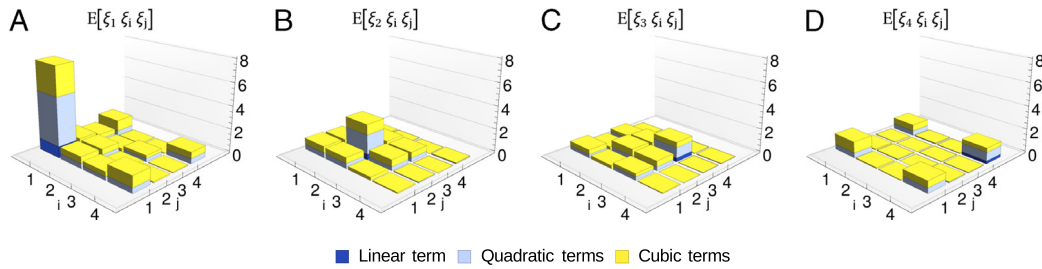


Fig. 4. The analytical expected values (obtained by Eq. (1)) of all third moments for $n = 5, \theta = 1$ and the respective contributions of the linear, quadratic and cubic terms.

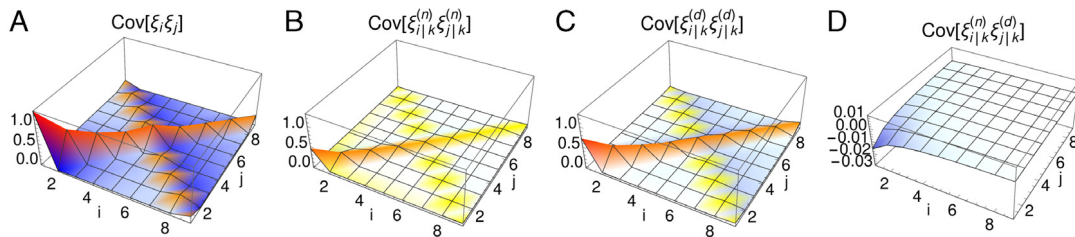


Fig. 5. Comparison between unconditional and conditional covariances. Panel A: unconditional covariances $Cov[\xi_i, \xi_j]$ for sample size $n = 10$, calculated with the formulae of Fu (1995). The remainder graphs show the covariances between mutations conditional on a mutation of size $k = 10$ in a sample of size $n = 20$, obtained by Eqs. (25). Panel B shows the covariances between mutations nested within the focal mutation, panel C the covariances of mutations both disjoint and panel D the covariance between nested and disjoint mutations.

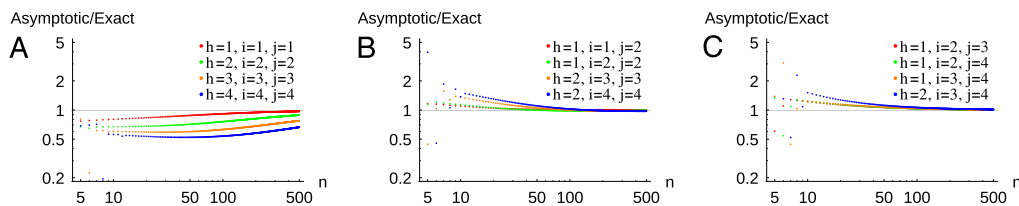


Fig. 6. From Dahmer and Kersting (2015) follow that the random variables ξ_h, ξ_i, ξ_j are approximately independent in samples of size n , as long as $h, i, j \ll n$. Shown is the ratio of the asymptotic approximation (28) to the exact expression (6) for small fixed indices $1 \leq h, i, j \leq 4$ and varying sample size n . Panel A: all indices are the same; panel B: two indices differ; panel C: all indices differ.

However the distribution of each component of the spectrum ξ_k shows excesses of outliers and heavy tails (Janson and Kersting, 2011), hence the convergence in distribution proved by Dahmer and Kersting does not imply the convergence of the moments, and therefore does not imply the scaling (28). This reduces the usefulness of the asymptotic results, in particular for the case of mutations of identical size ($h = i = j$). Fig. 6 shows that the asymptotic expansion is reasonably good for moments ξ_h, ξ_i, ξ_j with $h, i, j \ll n$ and at least two indices differing. If any of the indices h, i, j is greater than $\frac{n}{2}$, the asymptotic results seem to be of little help.

3. Methods

3.1. Proof of Theorem 2.1

3.1.1. Separation of estimation

A coalescent tree is constructed by two independent stochastic processes, namely its branching pattern (the topology) and the lengths of its branches (coalescent times). The idea of Fu (1995) is to decompose the tree into small parts, called *lines*, by cutting each branch along *states* which are delineated by coalescent events (cf. Fig. 7). He first calculates the probabilities of all hierarchical relationships between those lines by transforming the probabilistic problem into a combinatorial one. Second, he computes the first and second moments of the number of mutations on each line, deriving from a third random process, which depends only on the lengths of the lines. Although mutations on each line arise independently, their numbers on lines of the same state are indirectly correlated because of shared line lengths. The moments of the

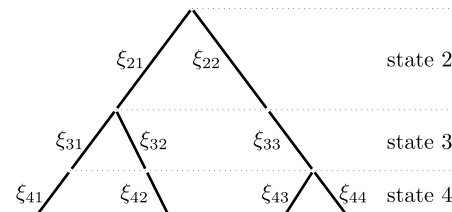


Fig. 7. This coalescent tree represents a genealogy of four sequences sampled from a large population. It is decomposed into lines kl , on which ξ_{kl} mutations occur, depending on the length of the line. Let us focus on the number of mutations of size two, i.e. which are present on two sequences: For a tree with a topology \mathcal{T} as depicted, only the lines 21, 22 and 33 have two leaves and hence contribute to this number. Hence for the “indicator variables” we have: $\epsilon_{21}(2) = \epsilon_{22}(2) = \epsilon_{33}(2) = 1$ and $\epsilon_{kl}(2) = 0$ for all other lines. It follows that $E[\xi_2]_{Topology=\mathcal{T}} = E[\xi_{21}] + E[\xi_{22}] + E[\xi_{33}]$. Averaging over all topologies yields $E[\xi_2]$.

individual lines, averaged over the topologies, yield the desired moments of the frequency spectrum.

We re-use method and notation of Fu (1995) with appropriate extensions. A thorough explanation of the main ingredients of his proof, albeit with somewhat different notation, has been given in Durrett (2008). An extended “reprint” of the more technical parts can be found in the supplement of our companion paper (Ferretti et al., 2017).

We define ξ_{kl} as the number of mutations occurring on line l of state k . Furthermore the index variables $\epsilon_{kl}(i)$ indicate whether the corresponding line has i descendants at state n , (i.e. they take the values 1 resp. 0). It follows that (cf. Fig. 7)

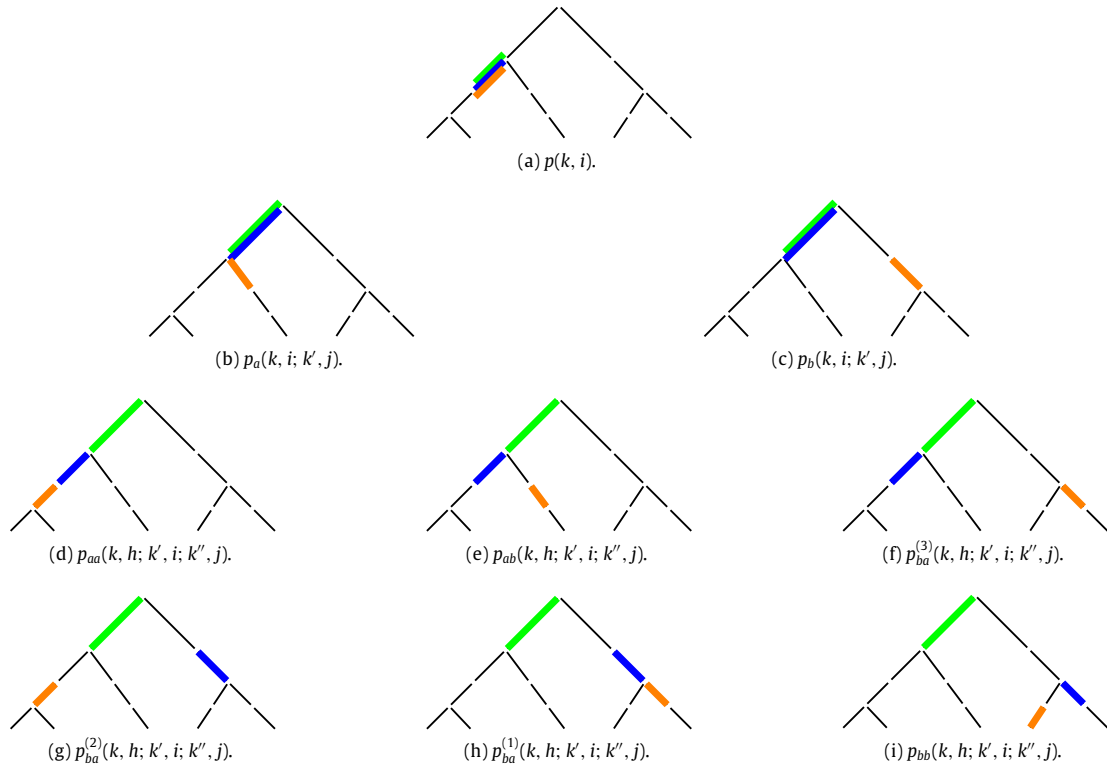


Fig. 8. The hierarchical relationships between three lines of a coalescent tree and their corresponding probabilities.

$$\xi_i = \sum_{k=2}^n \sum_{l=1}^k \epsilon_{kl}(i) \xi_{kl}. \tag{29}$$

In the following we use the fact that the index l serves only to distinguish lines of the same state, but otherwise has no meaning, since all lines of the same state are equivalent. The indicator variables are idempotent ($\epsilon_{kl}(i)^2 = \epsilon_{kl}(i)$) and independent of the number of mutations ξ_{kl} . The expectation values of the indicator variables correspond to probabilities, which we will define in the following subsection.

3.1.2. Averaging over topologies

The statistical properties of the lines in a coalescent tree are related to *Pólya urn theory*, originally aimed at modelling the spread of infectious diseases, while other applications in theoretical biology have been found later (Mahmoud, 2008). A Pólya urn process starts with an urn containing balls of various colours, from which repeatedly a ball is drawn and put back together with an additional ball of the same colour. In the coalescent, the balls correspond to lines and the addition of a ball is equivalent to a split of a line into two. The corresponding probabilities are reviewed in Griffiths and Tavaré (2003) and our Eqs. (30)–(32) follow from Eq. (2.1) therein. We introduce the following notation: $p_{k \rightarrow n}(t \rightarrow i)$ is the probability that t lines at state k have i descendants at state n . This probability is

$$p_{k \rightarrow n}(t \rightarrow i) = \frac{\binom{i-1}{t-1} \binom{n-i-1}{k-t-1}}{\binom{n-1}{k-1}}. \tag{30}$$

At this point it is helpful to define $\binom{-1}{-1} = 1$, while binomial coefficients containing any other combination of one or two negative numbers are set to zero (cf. Durrett (2008)). This makes it possible to subsume in the Eqs. (30)–(32) the case that $t = k$ lines of state k yield $i = n$ lines at state n (which is true with probability 1). Later on, these special cases will be resolved separately and none of the expressions in the section *Results* relies on this definition.

The probability that t and u (different) lines at state k have respectively i and j descendants at state n is

$$p_{k \rightarrow n}(t \rightarrow i, u \rightarrow j) = \frac{\binom{i-1}{t-1} \binom{j-1}{u-1} \binom{n-i-j-1}{k-t-u-1}}{\binom{n-1}{k-1}}. \tag{31}$$

And for three such (non-overlapping) sets of lines the probability yields

$$p_{k \rightarrow n}(s \rightarrow h, t \rightarrow i, u \rightarrow j) = \frac{\binom{h-1}{s-1} \binom{i-1}{t-1} \binom{j-1}{u-1} \binom{n-h-i-j-1}{k-s-t-u-1}}{\binom{n-1}{k-1}}. \tag{32}$$

We split the computation of the expectation values of the indicator variables (which define the topology) into several cases, pictured in Fig. 8. Using the above notation we can now state the probabilities for each case. We start with those derived by Fu: The probability that one line at state k has i descendants at state n is (Fu, 1995, Eq. (14))

$$\begin{aligned} p(k, i) &= p_{k \rightarrow n}(1 \rightarrow i) \\ &= \frac{\binom{n-i-1}{k-2}}{\binom{n-1}{k-1}}. \end{aligned} \tag{33}$$

The joint probability that one line at state k and one nested line at state $k' \geq k$ have i respective j descendants at state n is (Fu, 1995, Eq. (18))

$$\begin{aligned} p_a(k, i; k', j) &= \sum_{t=1}^{k'-1} p_{k \rightarrow k'}(1 \rightarrow t) \frac{t}{k'} p_{k' \rightarrow n}(t-1 \rightarrow i-j, 1 \rightarrow j) \\ &= \sum_{t=1}^{k'-1} \frac{\binom{k'-t-1}{k-2}}{\binom{k'-1}{k-1}} \frac{t}{k'} \frac{\binom{i-j-1}{t-2} \binom{n-i-1}{k'-t-1}}{\binom{n-1}{k'-1}}. \end{aligned} \tag{34}$$

The joint probability that one line at state k and one disjoint (not nested) line at state $k' \geq k$ have i resp. j descendants at state n is

(Fu, 1995, Eqs. (19) and (20))

$$\begin{aligned}
 p_b(k, i; k', j) &= \sum_{t=1}^{k'-1} p_{k \rightarrow k'}(1 \rightarrow t) \frac{k' - t}{k'} p_{k' \rightarrow n}(t \rightarrow i, 1 \rightarrow j) \\
 &= \sum_{t=1}^{k'-1} \frac{\binom{k'-t-1}{k-2}}{\binom{k'-1}{k-1}} \frac{k' - t}{k'} \frac{\binom{i-1}{t-1} \binom{n-i-j-1}{k'-t-2}}{\binom{n-1}{k'-1}}.
 \end{aligned} \tag{35}$$

In Eqs. (34) and (35) the summation index t runs over the possible numbers of descendants that the line of state k may have at state k' . Since no single line can be ancestor of all k' lines, this number has an upper limit of $k' - 1$. There are more constraints on t as detailed by Fu (1995) (e.g. a line from state k can have at most $k' - k + 1$ descendants at state k' , hence only values $t \leq k' - k + 1$ contribute to the sum), however these are already implicit in the binomial coefficients.

Note, that Fu defined Eqs. (34)–(35) only for the case $k < k'$. Using the special definition for the binomial coefficient, they include the case $k = k'$ (Durrett, 2008): if the lines are from the same state, then $t = 1$ and we have $p_a(k, i; k, j) = \delta_{i=j} \frac{1}{k} p(k, i)$ and $p_b(k, i; k, j) = \frac{k-1}{k} \frac{\binom{n-i-j-1}{k-3}}{\binom{n-1}{k-1}}$. These two equations correspond to Eqs. (14) and (15) of Fu (1995).

Hence the probability that a line at k and a line at state k' have i resp. j descendants at state n yields for $2 \leq k \leq k' \leq n$:

$$p(k, i; k', j) = p_a(k, i; k', j) + p_b(k, i; k', j). \tag{36}$$

Now we derive the probabilities involving three lines. These may be all of the same state, of two different states or of three different states. We assume $k \leq k' \leq k''$. We take a single line at each state k, k' and k'' respectively and subdivide along their possible relationships. We denote the lines l, l' and l'' respectively. The six cases are (compare Fig. 8):

- aa : l' is a descendant of l and l'' is a descendant of l'
- ab : l' and l'' are both descendants of l , but l'' is not a descendant of l'
- $ba^{(3)}$: l' is a descendant of l , but l'' is not
- $ba^{(2)}$: l'' is a descendant of l , but l' is not
- $ba^{(1)}$: l'' is a descendant of l' , but both are not descendants of l
- bb : no line is a descendant of any of the other two lines.

The probability of the first case yields

$$\begin{aligned}
 p_{aa}(k, h; k', i; k'', j) &= \sum_{t=1}^{k'-1} \sum_{t_1=0}^{k''-2} \sum_{t_2=1}^{k''-t_1-1} p_{k \rightarrow k'}(1 \rightarrow t) \frac{t}{k'} p_{k' \rightarrow k''}(t-1 \rightarrow t_1, 1 \rightarrow t_2) \\
 &\quad \times \frac{t_2}{k''} p_{k'' \rightarrow n}(t_1 \rightarrow h-i, t_2-1 \rightarrow i-j, 1 \rightarrow j) \\
 &= \sum_{t=1}^{k'-1} \sum_{t_1=0}^{k''-2} \sum_{t_2=1}^{k''-t_1-1} \frac{\binom{k'-t-1}{k-2}}{\binom{k'-1}{k-1}} \frac{t}{k'} \frac{\binom{t_1-1}{t-2} \binom{k''-t_1-t_2-1}{k'-t-1}}{\binom{k''-1}{k'-1}} \frac{t_2}{k''} \\
 &\quad \times \frac{\binom{h-i-1}{t_1-1} \binom{i-j-1}{t_2-2} \binom{n-h-1}{k''-t_1-t_2-1}}{\binom{n-1}{k''-1}}.
 \end{aligned} \tag{37}$$

In words, the summation goes over the probability that a random line of state k has t descendants at state k' , times the probability that another randomly chosen line at that state is one of these, times the probability that the second chosen line has t_2 and the other $t - 1$ lines have t_1 descendants at state k'' , times the probability that a third randomly chosen line of that state belongs to the t_2 lines, and finally that the $t_1, t_2 - 1$ and 1 line of state k'' have $h - i, i - j$ and j descendants at state n , respectively.

The remaining probabilities yield:

$$\begin{aligned}
 p_{ab}(k, h; k', i; k'', j) &= \sum_{t=2}^{k'-1} \sum_{t_1=1}^{k''-2} \sum_{t_2=1}^{k''-t_1} p_{k \rightarrow k'}(1 \rightarrow t) \frac{t}{k'} p_{k' \rightarrow k''}(t-1 \rightarrow t_1, 1 \rightarrow t_2) \frac{t_1}{k''} \\
 &\quad \times p_{k'' \rightarrow n}(t_1-1 \rightarrow h-i-j, t_2 \rightarrow i, 1 \rightarrow j) \\
 &= \sum_{t=2}^{k'-1} \sum_{t_1=1}^{k''-2} \sum_{t_2=1}^{k''-t_1} \frac{\binom{k'-t-1}{k-2}}{\binom{k'-1}{k-1}} \frac{t}{k'} \frac{\binom{t_1-1}{t-2} \binom{k''-t_1-t_2-1}{k'-t-1}}{\binom{k''-1}{k'-1}} \frac{t_1}{k''} \\
 &\quad \times \frac{\binom{h-i-j-1}{t_1-2} \binom{i-1}{t_2-1} \binom{n-h-1}{k''-t_1-t_2-1}}{\binom{n-1}{k''-1}},
 \end{aligned} \tag{38}$$

$$\begin{aligned}
 p_{ba}^{(3)}(k, h; k', i; k'', j) &= \sum_{t=1}^{k'-1} \sum_{t_1=0}^{k''-2} \sum_{t_2=1}^{k''-t_1-1} p_{k \rightarrow k'}(1 \rightarrow t) \frac{t}{k'} p_{k' \rightarrow k''}(t-1 \rightarrow t_1, 1 \rightarrow t_2) \\
 &\quad \times \frac{k'' - t_1 - t_2}{k''} p_{k'' \rightarrow n}(t_1 \rightarrow h-i, t_2 \rightarrow i, 1 \rightarrow j) \\
 &= \sum_{t=1}^{k'-1} \sum_{t_1=0}^{k''-2} \sum_{t_2=1}^{k''-t_1-1} \frac{\binom{k'-t-1}{k-2}}{\binom{k'-1}{k-1}} \frac{t}{k'} \frac{\binom{t_1-1}{t-2} \binom{k''-t_1-t_2-1}{k'-t-1}}{\binom{k''-1}{k'-1}} \\
 &\quad \times \frac{k'' - t_1 - t_2}{k''} \frac{\binom{h-i-1}{t_1-1} \binom{i-1}{t_2-1} \binom{n-h-j-1}{k''-t_1-t_2-2}}{\binom{n-1}{k''-1}},
 \end{aligned} \tag{39}$$

$$\begin{aligned}
 p_{ba}^{(2)}(k, h; k', i; k'', j) &= \sum_{t=1}^{k'-1} \sum_{t_1=1}^{k''-2} \sum_{t_2=1}^{k''-t_1} p_{k \rightarrow k'}(1 \rightarrow t) \frac{k' - t}{k'} p_{k' \rightarrow k''}(t \rightarrow t_1, 1 \rightarrow t_2) \\
 &\quad \times \frac{t_1}{k''} p_{k'' \rightarrow n}(t_1-1 \rightarrow h-j, t_2 \rightarrow i, 1 \rightarrow j) \\
 &= \sum_{t=1}^{k'-1} \sum_{t_1=1}^{k''-2} \sum_{t_2=1}^{k''-t_1} \frac{\binom{k'-t-1}{k-2}}{\binom{k'-1}{k-1}} \frac{k' - t}{k'} \frac{\binom{t_1-1}{t-1} \binom{k''-t_1-t_2-1}{k'-t-2}}{\binom{k''-1}{k'-1}} \frac{t_1}{k''} \\
 &\quad \times \frac{\binom{h-j-1}{t_1-2} \binom{i-1}{t_2-1} \binom{n-h-i-1}{k''-t_1-t_2-1}}{\binom{n-1}{k''-1}},
 \end{aligned} \tag{40}$$

$$\begin{aligned}
 p_{ba}^{(1)}(k, h; k', i; k'', j) &= \sum_{t=1}^{k'-1} \sum_{t_1=1}^{k''-2} \sum_{t_2=1}^{k''-t_1} p_{k \rightarrow k'}(1 \rightarrow t) \frac{k' - t}{k'} p_{k' \rightarrow k''}(t \rightarrow t_1, 1 \rightarrow t_2) \frac{t_2}{k''} \\
 &\quad \times p_{k'' \rightarrow n}(t_1 \rightarrow h, t_2-1 \rightarrow i-j, 1 \rightarrow j) \\
 &= \sum_{t=1}^{k'-1} \sum_{t_1=1}^{k''-2} \sum_{t_2=1}^{k''-t_1} \frac{\binom{k'-t-1}{k-2}}{\binom{k'-1}{k-1}} \frac{k' - t}{k'} \frac{\binom{t_1-1}{t-1} \binom{k''-t_1-t_2-1}{k'-t-2}}{\binom{k''-1}{k'-1}} \frac{t_2}{k''} \\
 &\quad \times \frac{\binom{h-1}{t_1-1} \binom{i-j-1}{t_2-2} \binom{n-h-i-1}{k''-t_1-t_2-1}}{\binom{n-1}{k''-1}},
 \end{aligned} \tag{41}$$

$$\begin{aligned}
 p_{bb}(k, h; k', i; k'', j) &= \sum_{t=1}^{k'-1} \sum_{t_1=1}^{k''-2} \sum_{t_2=1}^{k''-t_1} p_{k \rightarrow k'}(1 \rightarrow t) \frac{k' - t}{k'} p_{k' \rightarrow k''}(t \rightarrow t_1, 1 \rightarrow t_2) \\
 &\quad \times \frac{k'' - t_1 - t_2}{k''} p_{k'' \rightarrow n}(t_1 \rightarrow h, t_2 \rightarrow i, 1 \rightarrow j) \\
 &= \sum_{t=1}^{k'-1} \sum_{t_1=1}^{k''-2} \sum_{t_2=1}^{k''-t_1} \frac{\binom{k'-t-1}{k-2}}{\binom{k'-1}{k-1}} \frac{k' - t}{k'} \frac{\binom{t_1-1}{t-1} \binom{k''-t_1-t_2-1}{k'-t-2}}{\binom{k''-1}{k'-1}} \\
 &\quad \times \frac{k'' - t_1 - t_2}{k''} \frac{\binom{h-1}{t_1-1} \binom{i-1}{t_2-1} \binom{n-h-i-j-1}{k''-t_1-t_2-2}}{\binom{n-1}{k''-1}}.
 \end{aligned} \tag{42}$$

Since the six cases cover all possible combinations, the total probability that three lines at state k, k' and k'' resp. (with $k \leq k' \leq k''$) have h, i and j resp. descendants at state n is given by

$$\begin{aligned}
 p(k, h; k', i; k'', j) &= p_{aa}(k, h; k', i; k'', j) + p_{ab}(k, h; k', i; k'', j) \\
 &+ p_{ba}^{(3)}(k, h; k', i; k'', j) \\
 &+ p_{ba}^{(2)}(k, h; k', i; k'', j) \\
 &+ p_{ba}^{(1)}(k, h; k', i; k'', j) \\
 &+ p_{bb}(k, h; k', i; k'', j).
 \end{aligned} \tag{43}$$

We now relate the indicator variables of Eq. (29) to the probabilities (33)–(42). For one and two lines we restate the results obtained by Fu (1995, text and equations without number, before Eq. (22))

$$E[\epsilon_{kl}(i)] = p(k, i) \tag{44}$$

$$\begin{aligned}
 E[\epsilon_{kl}(i)\epsilon_{k'l'}(j)] &= \delta_{i=j}p(k, i) \quad \text{if } k = k' \text{ and } l = l' \\
 E[\epsilon_{kl}(i)\epsilon_{k'l'}(j)] &= p(k, i; k', j) \quad \text{if } k = k' \\
 E[\epsilon_{kl}(i)\epsilon_{k'l'}(j)] &= p(k, i; k', j) \quad \text{else.}
 \end{aligned} \tag{45}$$

We add the expressions for three lines (still assuming $k \leq k' \leq k''$):

$$\begin{aligned}
 E[\epsilon_{kl}(h)\epsilon_{k'l'}(i)\epsilon_{k''l''}(j)] &= \delta_{h=i=j}p(k, i) \quad \text{if } k = k' = k'' \\
 &\quad \text{and } l = l' = l'' \\
 E[\epsilon_{kl}(h)\epsilon_{k'l'}(i)\epsilon_{k''l''}(j)] &= \delta_{h=i}p(k, i; k'', j) \quad \text{if } k = k' \text{ and } l = l' \\
 E[\epsilon_{kl}(h)\epsilon_{k'l'}(i)\epsilon_{k''l''}(j)] &= \delta_{i=j}p(k, h; k', i) \quad \text{if } k' = k'' \text{ and } l' = l'' \\
 E[\epsilon_{kl}(h)\epsilon_{k'l'}(i)\epsilon_{k''l''}(j)] &= p(k, h; k', i; k'', j) \quad \text{else.}
 \end{aligned} \tag{46}$$

3.1.3. The third moments of the number of mutations on individual lines

The third moments of the number of mutations on three individual lines can be calculated with help of the “law of total expectation”, which is used repeatedly in coalescent theory, e.g. Wakeley (2008, Eq. (4.9)). We have to make case distinctions with respect to two or three lines being identical (and hence sharing the amount of mutations) or not identical, but of the same state (sharing line length). Mutations on lines of different states are not correlated. We state their third moments in terms of their first moments since this will turn out to be convenient later on.

Proposition 3.1. For any $1 \leq k, k', k'' < n, 1 \leq l \leq k, 1 \leq l' \leq k', 1 \leq l'' \leq k''$ the following equation holds:

$$\begin{aligned}
 E[\xi_{kl}\xi_{k'l'}\xi_{k''l''}] &= \delta_{k=k'=k''}\delta_{l=l'=l''}E[\xi_{k1}] \\
 &+ \delta_{k=k'=k''}(\delta_{l=l'} + \delta_{l=l''} + \delta_{l'=l''})E[\xi_{k1}]^2 \\
 &+ \delta_{k=k'}\delta_{l=l'}E[\xi_{k1}]E[\xi_{k'l'1}] \\
 &+ \delta_{k=k'}\delta_{l=l''}E[\xi_{k1}]E[\xi_{k'l''1}] \\
 &+ \delta_{k'=k''}\delta_{l'=l''}E[\xi_{k'l'1}]E[\xi_{k'l''1}] \\
 &+ (2\delta_{k=k'=k''} + \delta_{k=k'} + \delta_{k=k''} + \delta_{k'=k''} + 1)E[\xi_{k1}]E[\xi_{k'l'1}]E[\xi_{k'l''1}].
 \end{aligned} \tag{47}$$

Proof. Let X be a random variable. It can be easily shown that if X is exponentially distributed ($X \sim \text{Exp}(\lambda)$), then the first three moments of X are $E[X] = \frac{1}{\lambda}, E[X^2] = \frac{2}{\lambda^2}$ and $E[X^3] = \frac{6}{\lambda^3}$. If X is Poisson-distributed ($X \sim \text{Poisson}(\mu)$), then $E[X] = \mu, E[X^2] = \mu + \mu^2$ and $E[X^3] = \mu + 3\mu^2 + \mu^3$. In agreement with the definition of the coalescent the ξ_{kl} are distributed as $\xi_{kl} \sim \text{Poisson}(\frac{\theta}{2}T_k)$ with $T_k \sim \text{Exp}(\frac{2}{k(k-1)})$. ξ_{kl} and $\xi_{k'l'}$ are independent if $k \neq k'$ while ξ_{kl} and

$\xi_{k'l'}$ are independent conditional on T_k for $l \neq l'$.

$$\begin{aligned}
 E[\xi_{kl}^3] &= E[E[\xi_{kl}^3|T_k]] \\
 &= E[T_k\frac{\theta}{2} + 3(T_k\frac{\theta}{2})^2 + (T_k\frac{\theta}{2})^3] \\
 &= \frac{2}{k(k-1)}\frac{\theta}{2} + 3 \cdot 2\frac{4}{k^2(k-1)^2}\frac{\theta^2}{4} + 6\frac{8}{k^3(k-1)^3}\frac{\theta^3}{8}
 \end{aligned} \tag{48}$$

$$\begin{aligned}
 &= \frac{1}{k(k-1)}\theta + \frac{6}{k^2(k-1)^2}\theta^2 + \frac{6}{k^3(k-1)^3}\theta^3 \\
 &= E[\xi_{k1}] + 6E[\xi_{k1}]^2 + 6E[\xi_{k1}]^3 \\
 E[\xi_{kl}^2\xi_{k'l'}] &= E[E[\xi_{kl}^2\xi_{k'l'}|T_k]] \\
 &= E[E[\xi_{kl}^2|T_k]E[\xi_{k'l'}|T_k]] \\
 &= E[(T_k\frac{\theta}{2} + (T_k\frac{\theta}{2})^2)T_k\frac{\theta}{2}] \\
 &= \frac{2}{k^2(k-1)^2}\theta^2 + \frac{6}{k^3(k-1)^3}\theta^3 \\
 &= 2E[\xi_{k1}]^2 + 6E[\xi_{k1}]^3
 \end{aligned} \tag{49}$$

$$\begin{aligned}
 E[\xi_{kl}\xi_{k'l'}\xi_{k''l''}] &= E[E[\xi_{kl}\xi_{k'l'}\xi_{k''l''}|T_k]] \\
 &= E[E[\xi_{kl}|T_k]E[\xi_{k'l'}|T_k]E[\xi_{k''l''}|T_k]] \\
 &= E[(T_k\frac{\theta}{2})^3] \\
 &= \frac{6}{k^3(1-3)^3}\theta^3 \\
 &= 6E[\xi_{k1}]^3
 \end{aligned} \tag{50}$$

$$\begin{aligned}
 E[\xi_{kl}^2\xi_{k'l'}] &= E[\xi_{kl}^2]E[\xi_{k'l'}] \\
 &= \frac{1}{k(k-1)k'(k'-1)}\theta^2 + \frac{2}{k^2(k-1)^2k'(k'-1)}\theta^3 \\
 &= E[\xi_{k1}]E[\xi_{k'l'1}] + 2E[\xi_{k1}]^2E[\xi_{k'l'1}]
 \end{aligned} \tag{51}$$

$$\begin{aligned}
 E[\xi_{kl}\xi_{k'l'}\xi_{k''l''}] &= E[\xi_{kl}\xi_{k'l'}]E[\xi_{k''l''}] \\
 &= \frac{2}{k^2(k-1)^2k'(k'-1)}\theta^3 \\
 &= 2E[\xi_{k1}]^2E[\xi_{k'l'1}]
 \end{aligned} \tag{52}$$

$$E[\xi_{kl}\xi_{k'l'}\xi_{k''l''}] = E[\xi_{k1}]E[\xi_{k'l'1}]E[\xi_{k''l''1}] \tag{52}$$

3.1.4. Combining results

We average now the third moments of individual lines over topologies by inserting Eqs. (44)–(47) into Eq. (29):

$$\begin{aligned}
 E[\xi_h\xi_i\xi_j] &= E[(\sum_{k=2}^n \sum_{l=1}^k \epsilon_{kl}(h)\xi_{kl}) \\
 &\quad \times (\sum_{k'=2}^n \sum_{l'=1}^{k'} \epsilon_{k'l'}(i)\xi_{k'l'}) (\sum_{k''=2}^n \sum_{l''=1}^{k''} \epsilon_{k''l''}(j)\xi_{k''l''})] \\
 &= \sum_{k=2}^n \sum_{k'=2}^n \sum_{k''=2}^n \sum_{l=1}^k \sum_{l'=1}^{k'} \sum_{l''=1}^{k''} E[\epsilon_{kl}(h)\epsilon_{k'l'}(i)\epsilon_{k''l''}(j)] \\
 &\quad \times E[\xi_{kl}\xi_{k'l'}\xi_{k''l''}] \\
 &= \delta_{h=i=j} \sum_{k=2}^n kE[\epsilon_{k1}(h)]E[\xi_{k1}]
 \end{aligned} \tag{53}$$

$$\begin{aligned}
 & + \sum_{k=2}^n k^2 (\delta_{h=i} E[\epsilon_{k1}(i) \epsilon_{k2}(j)] + \delta_{i=j} E[\epsilon_{k1}(j) \epsilon_{k2}(h)]) \\
 & + \delta_{j=h} E[\epsilon_{k1}(h) \epsilon_{k2}(i)] E[\xi_{k1}]^2 \\
 & + \sum_{k=2}^n \sum_{k'=2}^n k k' (\delta_{h=i} E[\epsilon_{k1}(i) \epsilon_{k2}(j)] + \delta_{i=j} E[\epsilon_{k1}(j) \epsilon_{k2}(h)]) \\
 & + \delta_{j=h} E[\epsilon_{k1}(h) \epsilon_{k2}(i)] E[\xi_{k1}] E[\xi_{k2}] \\
 & + 2 \sum_{k=2}^n k E[\epsilon_{k1}(h)] E[\xi_{k1}]^3 \\
 & + \sum_{k=2}^n \sum_{k'=2}^n k k' (\delta_{h=i} E[\epsilon_{k1}(i) \epsilon_{k'1}(j)] + \delta_{i=j} E[\epsilon_{k1}(j) \epsilon_{k'1}(h)]) \\
 & + \delta_{j=h} E[\epsilon_{k1}(h) \epsilon_{k'1}(i)] E[\xi_{k1}]^2 E[\xi_{k'1}] \\
 & + \sum_{k=2}^n \sum_{k'=2}^n \sum_{k''=2}^n k k' k'' E[\epsilon_{k1}(h) \epsilon_{k'1}(i) \epsilon_{k''1}(j)] E[\xi_{k1}] E[\xi_{k'1}] \\
 & \times E[\xi_{k''1}] \\
 = & \delta_{h=i=j} \sum_{k=2}^n k p(k, h) E[\xi_{k1}] \\
 & + \delta_{h=i} \sum_{k=2}^n \sum_{k'=k}^n k k' (p(k, i; k', j) \\
 & + p(k, j; k', i)) E[\xi_{k1}] E[\xi_{k2}] \\
 & + \delta_{i=j} \sum_{k=2}^n \sum_{k'=k}^n k k' (p(k, j; k', h) \\
 & + p(k, h; k', j)) E[\xi_{k1}] E[\xi_{k2}] \\
 & + \delta_{j=h} \sum_{k=2}^n \sum_{k'=k}^n k k' (p(k, h; k', j) \\
 & + p(k, j; k', h)) E[\xi_{k1}] E[\xi_{k2}] \\
 & + \sum_{k=2}^n \sum_{k'=k}^n \sum_{k''=k'}^n k k' k'' (p(k, h; k', i; k'', j) \\
 & + p(k, i; k', j; k'', h) + p(k, j; k', h; k'', i) \\
 & + p(k, h; k', j; k'', i) + p(k, i; k', h; k'', j) \\
 & + p(k, j; k', i; k'', h)) E[\xi_{k1}] E[\xi_{k'1}] E[\xi_{k''1}]. \tag{54}
 \end{aligned}$$

Applying Eq. (22) of Fu (1995) to the first term of (54) yields Eq. (2):

$$\sum_{k=2}^n k p(k, i) E[\xi_{k1}] = \frac{\theta}{i} = \tau_i \theta, \tag{55}$$

and applying his Eq. (23) to the next three terms of (54) yields Eq. (4):

$$\sum_{k=2}^n \sum_{k'=k}^n k k' (p(k, i; k', j) + p(k, j; k', i)) E[\xi_{k1}] E[\xi_{k2}] = \tau_{ij} \theta^2. \tag{56}$$

We define the remaining terms of (54) as functions

$$\begin{aligned}
 t_x(h, i, j) = & \theta^{-3} \sum_{k=2}^n \sum_{k'=k}^n \sum_{k''=k'}^n k k' k'' p_x(k, h; k', i; k'', j) \\
 & \times E[\xi_{k1}] E[\xi_{k'1}] E[\xi_{k''1}] \tag{57}
 \end{aligned}$$

where x stands for $\{aa, ab, ba^{(3)}, ba^{(2)}, ba^{(1)}, bb\}$ and finally we set

$$t_{ba}(h, i, j) = t_{ba}^{(3)}(i, j, h) + t_{ba}^{(2)}(i, h, j) + t_{ba}^{(1)}(h, i, j). \tag{58}$$

In the supplement we transform these functions to yield (6).

We offer an implementation in C++ for numerical calculation of the third moments, given n and θ , using the expressions (1)–(6). As a control, we implemented the unsimplified functions (57), too. Within rounding errors ($< 10^{-12}$) they yield the same values as (6) for all third moments $E[\xi_h \xi_i \xi_j]$ and tested sample sizes $2 \leq n \leq 17$. With the algebraic computing software MATHEMATICA (Wolfram Research, Inc., 2014) we were able to prove for the same range of n that expressions (6) and (57)–(58) are exactly equivalent. The source code is contained in the package “coatli”, downloadable at <http://sourceforge.net/projects/coatli>.

3.2. Proof of Proposition 2.1

Proof. The functions $\alpha_n(i)$, $\beta_n(i)$, $\alpha_n^{(2)}(1, i)$ and $\beta_n^{(2)}(1, i)$ have the form

$$\frac{1}{i} \sum_{k=2}^n \frac{\binom{n-k}{i-1}}{\binom{n-1}{i-1}} c_k, \tag{59}$$

with coefficients c_k which do not depend neither on n nor on i . We set

$$f(n, i) = \frac{1}{i} \frac{\binom{n-k}{i-1}}{\binom{n-1}{i-1}} \tag{60}$$

and have for any $2 \leq i \leq n$ and $2 \leq k \leq n$

$$\begin{aligned}
 \left(1 - \frac{i}{n}\right) f(n, i) + \frac{i-1}{n} f(n, i-1) & = \left(1 - \frac{i}{n}\right) \frac{1}{i} \frac{\binom{n-k}{i-1}}{\binom{n-1}{i-1}} \\
 & + \frac{i-1}{n} \frac{1}{i-1} \frac{\binom{n-k}{i-2}}{\binom{n-1}{i-1}} = \frac{1}{i} \frac{n-i}{n} \frac{\binom{n-k}{i-1}}{\binom{n-1}{i-1}} + \frac{1}{n} \frac{\binom{n-k}{i-2}}{\binom{n-1}{i-1}} \\
 & = \frac{1}{i} \frac{\binom{n-k}{i-1}}{\binom{n-1}{i-1}} + \frac{1}{i} \frac{\binom{n-k}{i-2}}{\binom{n-1}{i-1}} = \frac{1}{i} \frac{\binom{n+1-k}{i-1}}{\binom{n-1}{i-1}} = f(n+1, i).
 \end{aligned}$$

Together with $\binom{n+1-k}{i-1} = 0$ for $k = n+1$ and $i > 1$ follows the proposition.

3.3. Proof of Theorem 2.2

The proof is identical to that of Theorem 2.1, except that we can ignore the relationships between lines.

Proof.

$$\begin{aligned}
 E[S^3] & = E\left[\left(\sum_{k=2}^n \sum_{l=1}^k \xi_{kl}\right) \left(\sum_{k'=2}^n \sum_{l'=1}^{k'} \xi_{k'l'}\right) \left(\sum_{k''=2}^n \sum_{l''=1}^{k''} \xi_{k''l''}\right)\right] \\
 & = \sum_{k=2}^n \sum_{l=1}^k \sum_{k'=2}^n \sum_{l'=1}^{k'} \sum_{k''=2}^n \sum_{l''=1}^{k''} E[\xi_{kl} \xi_{k'l'} \xi_{k''l''}] \\
 & \stackrel{(47)}{=} \sum_{k=2}^n k E[\xi_k] + \sum_{k=2}^n k^2 E[\xi_k^2] + 3 \sum_{k=2}^n \sum_{k'=2}^n k k' E[\xi_{k1}] E[\xi_{k'1}] \\
 & + 2 \sum_{k=2}^n k^3 E[\xi_{k1}] + 3 \sum_{k=2}^n \sum_{k'=2}^n k k' E[\xi_{k1}] E[\xi_{k'1}] \\
 & + \sum_{k=2}^n \sum_{k'=2}^n \sum_{k''=2}^n k k' k'' E[\xi_{k1}] E[\xi_{k'1}] E[\xi_{k''1}]
 \end{aligned}$$

$$\begin{aligned}
&= \sum_{k=1}^{n-1} \frac{1}{k} \theta + 3 \sum_{k=1}^{n-1} \frac{1}{k^2} \theta^2 + 3 \left(\sum_{k=1}^{n-1} \frac{1}{k} \theta \right)^2 + 2 \sum_{k=1}^{n-1} \frac{1}{k^3} \theta^3 \\
&+ 3 \sum_{k=1}^{n-1} \frac{1}{k^2} \theta^2 \sum_{k=1}^{n-1} \frac{1}{k} \theta + \left(\sum_{k=1}^{n-1} \frac{1}{k} \theta \right)^3.
\end{aligned}$$

In the supplement we give an alternative derivation using the approach of Watterson (1975), which does not rely on coalescent theory.

4. Discussion

Kingman's coalescent (Kingman, 1982) is an established model to describe the patterns of mutations in neutral populations. For this reason, coalescent methods were used to compute analytically the expectation and covariance of the frequency spectrum (Fu, 1995). Here, we derive for the first time its third moments. We hope, they add a valuable building block to coalescent theory.

Furthermore we show how to compute analytically the bias of several important neutrality tests. Moreover, we describe the joint frequency spectrum for triplets of sites (fully characterising their expected haplotype structure).

The conditional frequency spectrum can be useful to characterise structural variation such as chromosomal inversions and introgressions (Ferretti et al., 2017). Although structural variants have been studied already a long time (Corbett-Detig and Hartl, 2012), recent improvements of high-throughput sequencing technology allow their investigation on a much larger scale (Sudmant et al., 2015). When alleles are found at intermediate frequency, it is not obvious, whether they are under balancing selection, ongoing positive selection or just neutrally evolving by genetic drift (Hoffmann and Rieseberg, 2008). The standard (unconditional) frequency spectrum might not be well-suited for inferring the fate of such variants. Especially in regions with an inversion, recombination can be strongly inhibited (Kirkpatrick, 2010), which allows to partition the spectrum into nested and disjoint components with respect to the inverted sequences. Nested/disjoint spectra can hence be used to extend the class of frequency spectrum based tests on neutrality to cope with genomic features such as inversions and introgressions. The proper normalisation of such tests requires the knowledge of the corresponding variances and covariances given in Eq. (25).

Note that there is a close relation between the joint spectrum of multiple sites and the multi-allelic spectrum of a single locus (Ferretti et al., 2017). In fact, at low mutation rates, we can consider the multiple sites as a single locus with multiple alleles, and retrieve the multi-allelic spectrum for the locus by considering the frequencies of the $m + 1$ alleles that result from the m polymorphic sites. In this light, our results can be used to derive the full quadri-allelic frequency spectrum. This could be applied to several multi-allelic variants, the more relevant being nucleotide polymorphism (which have at most four alleles A,C,G,T). Related results can be found in Jenkins and Song (2011) and Bhaskar et al. (2012).

Acknowledgements

We thank Thomas Wiehe and the anonymous reviewers for comments and advice on this manuscript, and Iulia Dahmer and Götz Kersting for insightful discussions. AK was supported by a grant of the German Science Foundation (DFG-SFB680) to T. Wiehe (University of Cologne). LF was supported by funding from BBSRC grant BBS/E/I/00007039.

Appendix A. Supplementary data

Supplementary material related to this article can be found online at <https://doi.org/10.1016/j.tpb.2017.12.002>.

References

- Achaz, G., 2009. Frequency spectrum neutrality tests: One for all and all for one. *Genetics* 183, 249–258.
- Bhaskar, A., Kamm, J.A., Song, Y.S., 2012. Approximate sampling formulas for general Finite-Alleles models of mutation. *Adv. Appl. Probab.* 44, 408–428.
- Corbett-Detig, R.B., Hartl, D.L., 2012. Population genomics of inversion polymorphisms in *Drosophila melanogaster*. *PLoS Genet.* 8.
- Dahmer, I., Kersting, G., 2015. The internal branch lengths of the kingman coalescent. *Ann. Appl. Probab.* 25, 1325–1348.
- Durrett, R., 2008. *Probability Models for DNA Sequence Evolution*, second ed. Springer.
- Ewens, W., 1979. *Mathematical Population Genetics*. Springer Verlag.
- Fay, J.C., Wu, C.-I., 2000. Hitchhiking under positive Darwinian selection. *Genetics* 155, 1405–1413.
- Ferretti, L., Klassmann, A., Wiehe, T., Ramos-Onzins, S., Achaz, G., 2017. The expected neutral frequency spectrum of two linked sites. <https://doi.org/10.1101/100123>.
- Ferretti, L., Perez-Enciso, M., Ramos-Onsins, S.E., 2010. Optimal neutrality tests based on the frequency spectrum. *Genetics* 186, 353–365.
- Fisher, R.A., 1930. The distribution of gene ratios for rare mutations. *Proc. Roy. Soc. Edinburgh* 50, 205–220.
- Fu, Y.-X., 1995. Statistical properties of segregating sites. *Theor. Popul. Biol.* 48, 172–197.
- Fu, Y.-X., Li, W.-H., 1993. Statistical tests of neutrality of mutations. *Genetics* 133, 693–709.
- Griffiths, R., Tavaré, S., 2003. The genealogy of a neutral mutation. In: *Highly Structured Stochastic Systems*. Oxford university press, pp. 393–412.
- Hein, J., Szierup, M., Wiuf, C., 2004. *Gene Genealogies, Variation and Evolution: A Primer in Coalescent Theory*. Oxford university press.
- Hoffmann, A.A., Rieseberg, L.H., 2008. Revisiting the Impact of Inversions in Evolution: From Population Genetic Markers to Drivers of Adaptive Shifts and Speciation?. *Annu. Rev. Ecol. Evol. Syst.* 39, 21–42.
- Hudson, R.R., 1983. Properties of a neutral allele model with intragenic recombination. *Theor. Popul. Biol.* 23, 183–201.
- Hudson, R.R., 1991. Gene genealogies and the coalescent process. In: *Oxford Surveys in Evolutionary Biology*. Oxford university press, pp. 1–44.
- Hudson, R.R., 2002. Generating samples under a Wright-Fisher neutral model of genetic variation. *Bioinformatics* 18, 337–338.
- Hudson, R.R., 2015. A new proof of the expected frequency spectrum under the standard neutral model. *PLoS One* 10 (7), e0118087.
- Janson, S., Kersting, G., 2011. On the total external length of the evolving Kingman coalescent. *Electron. J. Probab.* 80, 2203–2218.
- Jenkins, P.A., Song, Y.S., 2011. The effect of recurrent mutation on the frequency spectrum of a segregating site and the age of an allele. *Theor. Popul. Biol.* 80, 158–173.
- Kimmel, M., Axelrod, D., 2015. *Branching Processes in Biology*. Springer.
- Kimura, M., 1964. Diffusion models in population genetics. *J. Appl. Probab.* 1, 177–232.
- Kingman, J.F.C., 1982. The coalescent. *Stochastic Process. Appl.* 13, 235–248.
- Kirkpatrick, M., 2010. How and why chromosome inversions evolve. *PLoS Biol.* 8, e1000501.
- Mahmoud, H.M., 2008. *Pólya Urn Models*. Chapman & Hall/CRC, Boca Raton, Florida, USA.
- Nielsen, R., Bustamante, C.D., Clark, A.G., Glanowski, S., Sackton, T.B., et al., 2005. A scan for positively selected genes in the genomes of humans and chimpanzees. *PLoS Biol.* 3, e170.
- Rafajlović, M., Klassmann, A., Eriksson, A., Wiehe, T., Mehlig, B., 2014. Demography-adjusted tests of neutrality based on genome-wide SNP data. *Theor. Popul. Biol.* 95, 1–12.
- Sargsyan, O., 2015. An analytical framework in the general coalescent tree setting for analyzing polymorphisms created by two mutations. *J. Math. Biol.* 70, 913–956.
- Simonsen, K.L., Churchill, G. a., Aquadro, C.F., 1995. Properties of statistical tests of neutrality for DNA polymorphism data. *Genetics* 141, 413–429.

- Sudmant, P.H., Rausch, T., Gardner, E.J., Handsaker, R.E., Abyzov, A., et al., 2015. An integrated map of structural variation in 2,504 human genomes. *Nature* 526, 75–81.
- Tajima, F., 1983. Evolutionary relationship of DNA sequences in finite populations. *Genetics* 105, 437–460.
- Tajima, F., 1989. Statistical method for testing the neutral mutation hypothesis by DNA polymorphism. *Genetics* 123, 585–595.
- Tavaré, S., 1984. Line-of-descent and genealogical processes, and their applications in population genetics models. *Theor. Popul. Biol.* 26, 119–164.
- Van Erp, N., Van Gelder, P., 2007. On the moments of functions of random variables using multivariate Taylor expansion, part I. In: 5th International Probabilistic Workshop-Taerwe & Proske (Eds), Ghent.
- Wakeley, J., 2008. *Coalescent Theory: An Introduction*. W. H. Freeman.
- Watterson, G., 1975. On the number of Segregating Sites in Genetical Models without Recombination. *Theor. Popul. Biol.* 7, 256–276.
- Wolfram Research, Inc., 2014. *Mathematica* 10.0.
- Zeng, K., Fu, Y.-X., Shi, S., Wu, C.-I., 2006. Statistical tests for detecting positive selection by utilizing high-frequency variants. *Genetics* 174, 1431–1439.

SPECIAL ISSUE: POPULATION GENOMICS WITH R
REHH 2.0: a reimplementa-tion of the R package REHH to detect positive selection from haplotype structure

MATHIEU GAUTIER*,†, ALEXANDER KLASSMANN‡ and RENAUD VITALIS*,†

*INRA, UMR CBGP, Montferrier-sur-Lez F-34988, France, †Institut de Biologie Computationnelle, Montpellier F-34095, France, ‡Universität zu Köln, Köln D-50674, Germany

Abstract

Identifying genomic regions with unusually high local haplotype homozygosity represents a powerful strategy to characterize candidate genes responding to natural or artificial positive selection. To that end, statistics measuring the extent of haplotype homozygosity within (e.g. EHH, iHS) and between (Rsb or XP-EHH) populations have been proposed in the literature. The REHH package for R was previously developed to facilitate genome-wide scans of selection, based on the analysis of long-range haplotypes. However, its performance was not sufficient to cope with the growing size of available data sets. Here, we propose a major upgrade of the REHH package, which includes an improved processing of the input files, a faster algorithm to enumerate haplotypes, as well as multithreading. As illustrated with the analysis of large human haplotype data sets, these improvements decrease the computation time by more than one order of magnitude. This new version of REHH will thus allow performing iHS-, Rsb- or XP-EHH-based scans on large data sets. The package REHH 2.0 is available from the CRAN repository (<http://cran.r-project.org/web/packages/rehh/index.html>) together with help files and a detailed manual.

Keywords: EHH, footprints of selection, iHS, Rsb, XP-EHH

Received 3 August 2016; revision received 29 October 2016; accepted 31 October 2016

Introduction

Next-generation sequencing (NGS) technologies have deeply transformed the nature of polymorphism data. While population geneticists were, until recently, limited by the amount of available data in a handful of presumably independent markers, they now have access to dense single nucleotide polymorphism (SNP) data in both model and nonmodel species (Davey *et al.* 2011). In those species where genome assemblies are available, the analysis of haplotype structure in a population has proved useful to detect recent positive selection (Sabeti *et al.* 2002). Consider neutral mutations arising in a population: if any of these has, by chance, increased in frequency after a certain period of time, then recombination should have had time to break down linkage disequilibrium (LD) around it, thereby decreasing the length of haplotypes on which this mutation is located. Common variants are therefore expected to be old and standing on short haplotypes. If a mutation is selected for, however, it should expand in the population before recombination has time to break down the haplotype on which it

occurred. A powerful strategy to characterize candidate genes responding to natural or artificial positive selection thus consists in identifying genomic regions with unusually high local haplotype homozygosity, relatively to neutral expectation (Sabeti *et al.* 2002).

For that purpose, Sabeti *et al.* (2002) introduced a new metric, referred to as the extended haplotype homozygosity (EHH), which measures the decay of haplotype homozygosity as a function of genetic distance from a focal SNP. Tests of departure of EHH from neutral expectation were proposed, based on coalescent simulations of demographic history. Voight *et al.* (2006) later introduced a test statistic (iHS) based on the standardized log ratio of the integrals of the observed decay of EHH computed for the ancestral and the derived alleles at the focal SNP. Finally, cross-population statistics were proposed, to contrast EHH profiles between populations: XP-EHH (Sabeti *et al.* 2007) and Rsb (Tang *et al.* 2007). These haplotype-based methods of detecting selection have been applied on human data (see, e.g., Pickrell *et al.* 2009; Vitti *et al.* 2013), a wide range of livestock (see, e.g., Flori *et al.* 2014; Barson *et al.* 2015; Bosse *et al.* 2015) and plant species (see, e.g., Wang *et al.* 2014; Jin *et al.* in press), and also nonmodel species (see, e.g., Roesti *et al.* 2015; Mueller *et al.* 2016).

Correspondence: Renaud Vitalis, Fax: +33 (0)4 99 62 33 45; E-mail: vitalis@supagro.inra.fr

A few years ago, we developed REHH (Gautier & Vitalis 2012), a package for the statistical software R (R Development Core Team, 2016), to detect recent positive selection from the analysis of long-range haplotypes. Since then, two alternative programs were released: SELSCAN (Szpiech & Hernandez 2014), which introduces multithreading to improve computational efficiency, and HAPBIN (Maclean *et al.* 2015), which in addition to multithreading offers considerable gain in computation time thanks to a new computational approach based on a bitwise algorithm.

Here, we propose a major upgrade of the REHH package (Gautier & Vitalis 2012), which includes an improved algorithm to enumerate haplotypes, as well as multithreading. These improvements decrease the computation time by more than an order of magnitude, as compared to the previous REHH version (1.13), which eases the analysis of big data sets. Improving computation times is also useful to evaluate the power and sensitivity of the methods by means of simulations, or for inference: see, for example, the package COALA for R (Staab & Metzler 2016), which simulates sequence data following a given model of evolution and uses REHH to compute EHH-based summary statistics.

Below we provide a brief overview of the statistics and tests available in REHH 2.0 and give a detailed worked example of the analysis of chromosome 2 in humans (HSA2), from two HapMap samples: CEU (Utah residents with Northern and Western European ancestry from the CEPH collection) and JPT+CHB (Japanese in Tokyo, Japan, and Chinese from Beijing, China). We use this example as a guideline to use REHH 2.0. We further show how REHH was improved since the previous version, and how it compares to the alternative programs SELSCAN (Szpiech & Hernandez 2014) and HAPBIN (Maclean *et al.* 2015).

Overview of the EHH-based tests

In this section, we provide an overview of the EHH-based tests. The rationale of the computations is also illustrated in Box 1.

Within population tests

The allele-specific extended haplotype homozygosity: EHH. At a focal SNP and for a given core allele (ancestral or derived), the allele-specific extended haplotype homozygosity (EHH) is defined as the probability that two randomly chosen chromosomes (carrying the core allele considered) are identical by descent (IBD) (Sabeti *et al.* 2002). IBD is assayed by computing homozygosity at all SNPs within an interval surrounding the core region (Sabeti *et al.* 2002). The EHH thus aims at measuring to which extent an extended haplotype has been transmitted without recombination. In practice, the EHH

($EHH_{a_s,t}$) of a tested core allele a_s (by convention, $a_s = 1$ for the ancestral and $a_s = 2$ for the derived allele) for a focal SNP s over the chromosome interval comprised between the core allele a_s and the SNP t is computed as:

$$EHH_{a_s,t} = \frac{1}{n_{a_s}(n_{a_s} - 1)} \sum_{k=1}^{K_{a_s,t}} n_k(n_k - 1) \quad (\text{eqn 1})$$

where $K_{a_s,t}$ represents the number of distinct haplotypes (extending from SNP s to SNP t) carrying the core allele a_s , n_k is the observed count for the k th haplotype, and $n_{a_s} = \sum_{k=1}^{K_{a_s,t}} n_k$ gives the total number of haplotypes carrying the core allele a_s .

The integrated (allele-specific) EHH: iHH. By definition, irrespective of the allele considered, EHH equals 1 at the focal SNP and decays monotonically to 0 as one moves away from the focal SNP (Voight *et al.* 2006). For a given core allele, the integrated EHH (iHH) (Voight *et al.* 2006) is defined as the area under the EHH curve with respect to map position. In REHH (Gautier & Vitalis 2012), this definite integral is computed using the trapezoidal rule. In practice, the integral is only computed for the regions of the curve above an arbitrarily defined threshold `limehh` (e.g. $EHH > 0.05$). To avoid edge effects at the chromosome boundaries, iHH integrals are not computed if the leftmost or the rightmost value of the EHH curve stand above the `limehh` threshold. Following Voight *et al.* (2006), we also introduced the `maxgap` argument in the `calc_ehh` (`()`) function, to specify the maximum tolerated gap size (in bp) in the physical map between any two consecutive SNPs. Large gaps (e.g. centromeric regions) may indeed spuriously inflate the area under the EHH curve, which may result in false positives. Note that Voight *et al.* (2006) further applied a penalty (proportional to physical distances) to the genetic distance between successive SNPs separated by more than 20 kb when computing the iHH. We did not implement this option in REHH, although it might easily be done by modifying the positions of the markers in the SNP information file.

The standardized ratio of core alleles iHH: iHS. Let UniHS represents the log ratio of the iHH for its ancestral (iHH_a) and derived (iHH_d) alleles (Voight *et al.* 2006):

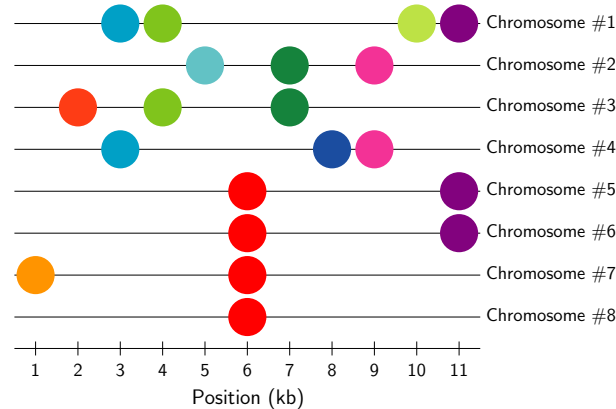
$$\text{UniHS} = \log\left(\frac{iHH_a}{iHH_d}\right) \quad (\text{eqn 2})$$

The iHS of a given focal SNP s ($iHS(s)$) is then defined following Voight *et al.* (2006) as:

$$iHS(s) = \frac{\text{UniHS}(s) - \mu_{\text{UniHS}}^s}{\sigma_{\text{UniHS}}^s} \quad (\text{eqn 3})$$

Box 1 An illustrated overview of the EHH-based tests

Box Fig. 1 shows a hypothetical pattern of variation in 8 aligned chromosomes genotyped at 11 single nucleotide polymorphisms (SNPs). We assume that we can delineate which variant at each position is ancestral and which is a new (derived) mutation: that is, the variants are ‘polarized’. Furthermore, the physical distance is set to 1000 bp (1 kb) between any two consecutive variants. The depicted pattern is meant to represent a toy example, where a single derived favourable mutation (in red, at position 6 kb) has spread into the population, sweeping genetic variation in its vicinity.



Box Fig. 1 Schematic view of 11 SNPs in 8 aligned chromosomes. Each of the eight lines symbolizes a chromosome and a filled circle represents a derived allele at the corresponding SNP position. [Colour figure can be viewed at wileyonlinelibrary.com].

All test statistics used in the `REHH` package start with the calculation of the (decay of) homozygosity around a focal SNP. The EHH takes different alleles separately (although Sabeti *et al.* 2002 used several SNPs to define ‘core SNP alleles’, it is now customary to compute the EHH for each single SNP, as in the `REHH` package). By contrast, the EHHS is calculated at a given SNP for the whole sample of chromosomes.

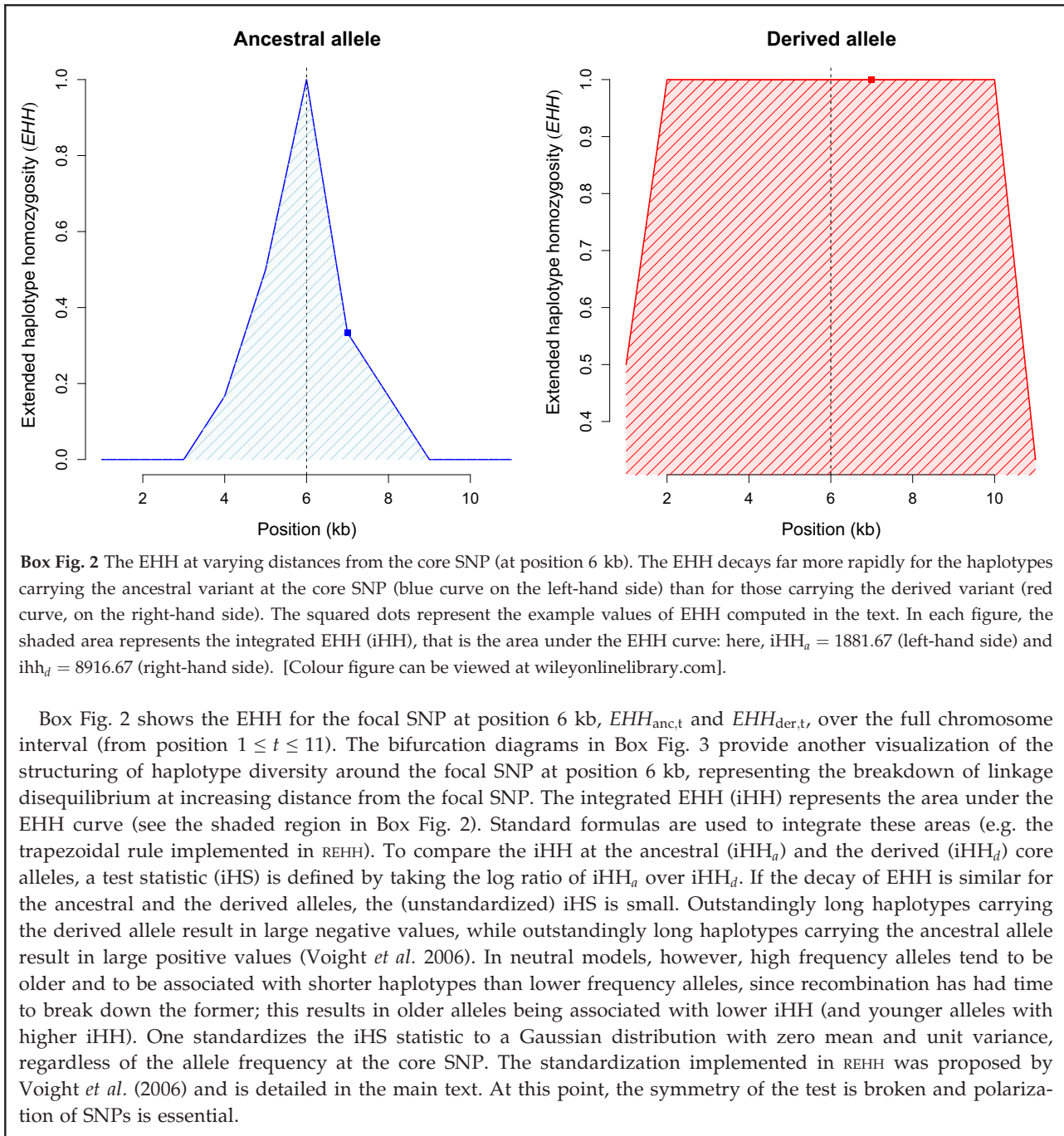
Let us outline the computation of EHH step by step. We start at the central position (6 kb), extending haplotypes to the right. We restrict our attention to the stretch extending from this variant to its right neighbour, at position 7 kb. Since all chromosomes carrying the derived allele ($a_s = \text{‘der’}$) at the focal SNP (chromosomes #5 to #8) are identical within the interval from 6 to 7 kb, we observe only one haplotype ($K_{\text{anc},7} = 1$) and therefore the extended haplotype homozygosity is $EHH_{\text{der},7} = 1$. There are two distinct chromosomes carrying the ancestral allele ($a_s = \text{‘anc’}$) at the focal SNP within the interval from 6 to 7 kb (chromosomes #1 to #4), and hence $K_{\text{anc},7} = 2$: chromosomes #1 and #4 are identical ($n_1 = 2$), as are chromosomes #2 and #3 ($n_2 = 2$), in the interval from 6 to 7 kb. Therefore, using equation (1) in the main text:

$$EHH_{\text{anc},7} = \frac{1}{n_{\text{anc}}(n_{\text{anc}} - 1)} \sum_{k=1}^{K_{\text{anc},7}} n_k(n_k - 1) = \frac{1}{4 \times 3} [2 \times (2 - 1) + 2 \times (2 - 1)] = \frac{1}{3}$$

where $\mu_{\text{UniHS}}^{p_s}$ and $\sigma_{\text{UniHS}}^{p_s}$ represent, respectively, the average and the standard deviation of the UniHS computed over all the SNPs with a derived allele frequency p_s similar to that of the core SNP s . In practice, the derived allele frequencies are generally binned so that each bin is large enough (e.g. >10 SNPs) to obtain reliable estimates of $\mu_{\text{UniHS}}^{p_s}$ and $\sigma_{\text{UniHS}}^{p_s}$. The iHS is constructed to have an approximately standard Gaussian distribution and to be comparable across SNPs regardless of their underlying allele frequencies. Hence, one may further transform iHS into p_{iHS} (Gautier & Naves 2011):

$$p_{\text{iHS}} = -\log_{10}(1 - 2|\Phi(\text{iHS}) - 0.5|) \quad (\text{eqn 4})$$

where $\Phi(x)$ represents the Gaussian cumulative distribution function. Assuming most of the genotyped SNPs behave neutrally (i.e. that the genomewide empirical iHS distribution is a fair approximation of the neutral distribution), p_{iHS} may thus be interpreted as a two-sided p -value (in a $-\log_{10}$ scale) associated with the null hypothesis of selective neutrality.

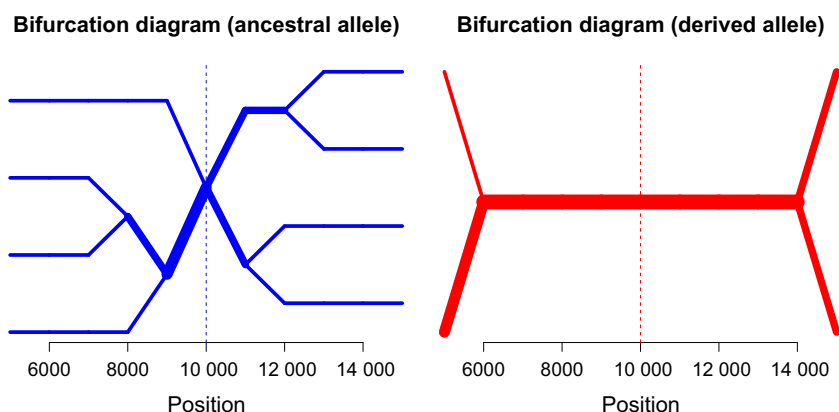


Pairwise population tests

The site-specific extended haplotype homozygosity: EHHS. At a focal SNP, the site-specific extended haplotype homozygosity (EHHS) is defined as the probability that two randomly chosen chromosomes are IBD at all SNPs within an interval surrounding the core region (Sabeti *et al.* 2007; Tang *et al.* 2007). EHHS might approximately be viewed as linear combination of the EHH's for the two alternative alleles, with some weights depending on

the corresponding allele frequencies. Two different EHHS estimators, further referred to as $EHHS_{Sabeti}^{Sabeti}$ and $EHHS_{Tang}^{Tang}$, have been proposed by Sabeti *et al.* (2007) and Tang *et al.* (2007), respectively. For a focal SNP s over a chromosome interval extending to SNP t , these are computed as (using the same notation as above):

$$EHHS_{s,t}^{Sabeti} = \frac{1}{n_s(n_s - 1)} \sum_{a_s=1}^2 \left(\sum_{k=1}^{K_{a_s,t}} n_k(n_k - 1) \right) \quad (\text{eqn 5})$$



Box Fig. 3 Haplotype bifurcation diagrams drawn for the ancestral (left-hand side) and derived (right-hand side) allele of the core SNP at position 6 kb. This diagram is bidirectional, representing the breakdown of LD at increasing distance: moving away from the core SNP, each variant is an opportunity for a node; the diagram divides if two alleles are present at this marker (hence defining two new haplotypes). The thickness of the lines corresponds to the counts of long-distance haplotype in the sample. [Colour figure can be viewed at wileyonlinelibrary.com].

Now, let us outline the computation of the site-specific EHH (EHHS), which may approximately be viewed as a linear (weighted) combination of EHH for the ancestral and the derived alleles. The EHHS yields a single value per population (and not per allele, as opposed to the EHH). There are two different definitions of EHHS in the literature (see Sabeti *et al.* 2007; Tang *et al.* 2007), which, in our example, give quite different values; there is, however, to the best of our knowledge, no demonstrated advantage of one over the other. Here, we will only compute Sabeti *et al.* (2007) statistic. Restricting our attention to the stretch extending from the focal SNP at position 6 kb to its right neighbour at position 7 kb, we get, using equation (5) in the main text:

$$\begin{aligned} \text{EHH}_{6,7}^{\text{Sabeti}} &= \frac{1}{n_6(n_6 - 1)} \left[\sum_{k=1}^{K_{\text{anc},7}} n_k(n_k - 1) + \sum_{k=1}^{K_{\text{der},7}} n_k(n_k - 1) \right] \\ &= \frac{1}{8 \times 7} [2 \times (2 - 1) + 2 \times (2 - 1) + 4 \times (4 - 1)] = \frac{2}{7} \end{aligned}$$

where $n_s = \sum_{a_s=1}^2 n_{a_s}$ and

$$\text{EHHS}_{s,t}^{\text{Tang}} = \frac{1 - h_{\text{hap}}^{(s,t)}}{1 - h_{\text{all}}^{(s)}} \quad (\text{eqn 6})$$

where

- $h_{\text{all}}^{(s)} = \frac{n_s}{n_s - 1} \left(1 - \frac{1}{n_s^2} \sum_{a_s=1}^2 n_{a_s}^2 \right)$ is an estimator of the focal SNP heterozygosity
- $h_{\text{hap}}^{(s,t)} = \frac{n_s}{n_s - 1} \left(1 - \frac{1}{n_s^2} \sum_{a_s=1}^2 \left(\sum_{k=1}^{K_{a_s,t}} n_k^2 \right) \right)$ is an estimator of haplotype heterozygosity over the chromosome interval extending from SNP s to SNP t .

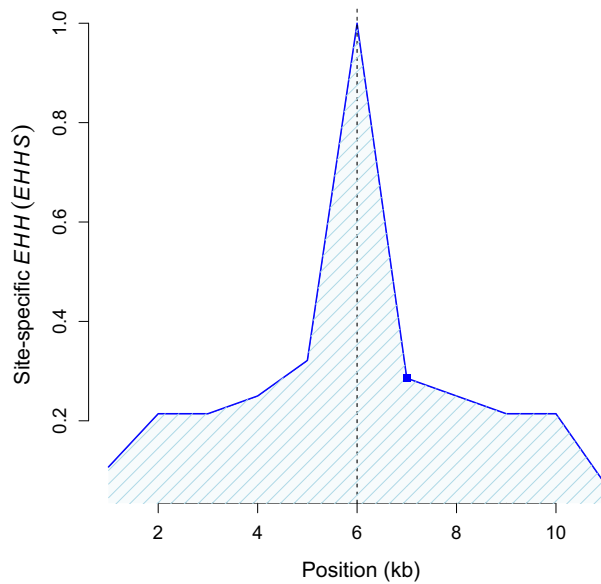
The integrated EHHS: iES

As for the EHH (see above), EHHS equals 1 at the focal SNP and decays monotonically to 0 as one moves away from the focal SNP. At the focal SNP, and in a similar

fashion as the iHH, iES is defined as the integrated EHHS (Tang *et al.* 2007). Depending on the EHHS estimator considered, $\text{EHHS}_{\text{Sabeti}}^{\text{Sabeti}}$ or $\text{EHHS}_{\text{Tang}}^{\text{Tang}}$, two different iES estimators, further referred to as $\text{iES}_{\text{Sabeti}}^{\text{Sabeti}}$ and $\text{iES}_{\text{Tang}}^{\text{Tang}}$, can be computed.

The standardized ratios of pairwise population iES: XP-EHH. For a given SNPs, let $\text{LRiES}_{\text{Sabeti}}^{\text{Sabeti}}(s)$ (respectively, $\text{LRiES}_{\text{Tang}}^{\text{Tang}}(s)$) represents the (unstandardized) log ratio of the $\text{iES}_{\text{pop1}}^{\text{Sabeti}}(s)$ and $\text{iES}_{\text{pop2}}^{\text{Sabeti}}(s)$ (respectively, $\text{iES}_{\text{pop1}}^{\text{Tang}}(s)$ and $\text{iES}_{\text{pop2}}^{\text{Tang}}(s)$) computed in two different populations (Sabeti *et al.* 2007; Tang *et al.* 2007):

$$\begin{aligned} \text{LRiES}_{\text{Sabeti}}^{\text{Sabeti}}(s) &= \log \left(\frac{\text{iES}_{\text{pop1}}^{\text{Sabeti}}(s)}{\text{iES}_{\text{pop2}}^{\text{Sabeti}}(s)} \right) \quad \text{and} \\ \text{LRiES}_{\text{Tang}}^{\text{Tang}}(s) &= \log \left(\frac{\text{iES}_{\text{pop1}}^{\text{Tang}}(s)}{\text{iES}_{\text{pop2}}^{\text{Tang}}(s)} \right) \end{aligned} \quad (\text{eqn 7})$$



Box Fig. 4 The EHHS at varying distances from the core SNP (at position 6 kb), computed following Sabeti *et al.* (2007). The squared dot represents the example value of EHHS computed in the text. The shaded area represents the integrated EHHS (iES), that is the area under the EHHS curve: here, $iES = 2553.57$. [Colour figure can be viewed at wileyonlinelibrary.com].

Box Fig. 4 shows the EHHS for the focal SNP at position 6 kb, $EHHS_{6,t}$, over the full chromosome interval (from position $1 \leq t \leq 11$). Just as the integrated EHH (iHH) represents the area under the EHH curve, the integrated EHHS (iES) represents the area under the EHHS curve (see the shaded region in Box Fig. 4). Note that, contrary to the EHH, the EHHS does not require the polarization of the SNPs. To compare the EHHS between populations, a test statistic (iES) is defined by taking the log ratio of EHHS measured in population 1 over EHHS measured in population 2. If the decay of EHHS is similar in both populations, the (unstandardized) iES is small. Strongly positive (respectively, negative) log ratios indicate outstandingly slow (respectively, fast) decay of EHHS in population 1, relatively to population 2, indicative of positive selection in population 1 (respectively, population 2). As for the iHS, the iES is standardized, yielding XP-EHH for Sabeti *et al.*'s (2007) definition of iES and R_{sb} for Tang *et al.*'s (2007).

In reality, variants are obviously not evenly spaced and their relative physical positions are important in the integration step. However, there are also tests (see, e.g., nS_L by Ferrer-Admetlla *et al.* 2014) which neglect on purpose this information in order to achieve robustness against varying recombination rate. Although such tests are not implemented in our package, we assume that setting equal distances between consecutive SNPs should lead to equivalent results.

The XP-EHH (Sabeti *et al.* 2007; Tang *et al.* 2007) for a given focal SNP are then standardized as:

$$XP-EHH(s) = \frac{LRiES^{Sabeti}(s) - \text{med}_{LRiES^{Sabeti}}}{\sigma_{LRiES^{Sabeti}}} \quad \text{and}$$

$$R_{sb}(s) = \frac{LRiES^{Tang}(s) - \text{med}_{LRiES^{Tang}}}{\sigma_{LRiES^{Tang}}} \quad (\text{eqn 8})$$

where $\text{med}_{LRiES^{Sabeti}}$ (respectively, $\text{med}_{LRiES^{Tang}}$) and $\sigma_{LRiES^{Sabeti}}$ (respectively, $\sigma_{LRiES^{Tang}}$) represent the median and standard deviation of the $LRiES^{Sabeti}(s)$ (respectively, $LRiES^{Tang}(s)$) computed over all the analysed SNPs. As

recommended by Tang *et al.* (2007), the median is used instead of the mean because it is less sensitive to extreme data points. As for the iHS (see above), XP-EHH and R_{sb} are constructed to have an approximately standard Gaussian distribution. They may further be transformed into p_{XP-EHH} or $p_{R_{sb}}$:

$$p_{XP-EHH} = -\log_{10}(1 - 2|\Phi_{(XP-EHH)} - 0.5|) \quad \text{and}$$

$$p_{R_{sb}} = -\log_{10}(1 - 2|\Phi_{(R_{sb})} - 0.5|) \quad (\text{eqn 9})$$

where $\Phi(x)$ represents the Gaussian cumulative distribution function. Assuming most of the genotyped SNPs behave neutrally (i.e. the genomewide empirical distributions of XP-EHH and R_{sb} are fair approximations of

their corresponding neutral distributions), p_{XP-EHH} and p_{Rsb} may thus be interpreted as a two-sided p -values (in a $-\log_{10}$ scale) associated with a null hypothesis of selective neutrality. Alternatively, one may also compute p'_{XP-EHH} or p'_{Rsb} as:

$$\begin{aligned} p'_{XP-EHH} &= -\log_{10}(\Phi_{(XP-EHH)}) & \text{and} & \\ p'_{Rsb} &= -\log_{10}(\Phi_{(Rsb)}) & \text{(eqn 10)} & \end{aligned}$$

(see Gautier & Naves 2011); p'_{XP-EHH} and p'_{Rsb} may then be interpreted as one-sided P -values (in a $-\log_{10}$ scale) allowing the identification of those sites displaying outstandingly high EHHs in population *pop2* (represented in the denominator of the corresponding LRiES) relatively to the reference population (*pop1*).

Materials and methods

A new efficient algorithm to explore haplotype variability

In the previous version of REHH (1.13), the distribution of haplotype counts for the entire interval from the core SNP to the distance x was computed for each x independently, entailing repeatedly the same calculations. In the new version of REHH (2.0), the distribution of haplotype counts for the interval from the core SNP to the distance $x+1$ is updated consecutively from the distribution of haplotype counts corresponding to the interval between the core SNP and x . We have at any position x an index set that records which sequences belong to the same extended haplotype (i.e. which sequences are identical in the corresponding interval). If the SNP at position $x+1$ has different alleles within a group of hitherto identical sequences, the index set is simply enlarged and the group is splitted correspondingly. Since this update does not depend upon the previous positions ($x-1, x-2, \dots$), the index set does not need to be stored for each position, which makes the algorithm memory and, therefore, time effective. The new algorithm does not affect the output. In particular, as in the previous version, haplotypes are not extended over a position where the sequence carries a missing value.

Human haplotype data

Two HSA2 haplotype data sets were downloaded from the HapMap project (phase III) (The International HapMap Consortium, 2010) website (<ftp://ftp.ncbi.nlm.nih.gov/hapmap/>). They consisted of 236 haplotypes of 116 430 SNPs from the CEU and 342 haplotypes from the JPT+CHB populations, respectively. Further details about these data (including the phasing procedure) can

be found on the HapMap website. For each SNP, the ancestral and derived alleles were determined according to the chimpanzee genome reference (using the *db-snp_chimp_B36.gff* annotation file available at ftp://ftp.ncbi.nlm.nih.gov/hapmap/gbrowse/2010-08_phaseII+III/gff/). Such ancestral information is indeed required to carry out iHS-based tests (see above). As a result, 6230 SNPs (5.35%) for which ancestral/derived states could not be unambiguously determined were discarded from further analyses leading to a total of 110 200 SNPs per analysed haplotype.

Computation

For comparison purposes, the different haplotype data sets were analysed using the software packages REHH (both the previous version 1.13 and the new version 2.0), SELSCAN (version 1.1.0b) (Szpiech & Hernandez 2014) and HAPBIN (version 1.0.0) (Maclean *et al.* 2015). Default options were generally used except for the minimal threshold on the minor allele frequency (MAF) that was set to 0.01 for all programs. In addition, for the SELSCAN program, both the window size around the core SNPs (`--ehh-win` option) and the maximum allowed gap in bp between two consecutive SNPs (`--max-gap` option) were set to 10^9 (this was made to disallow these options that are not considered in the HAPBIN program). Similarly, the `--max-extent` option was inactivated by setting `--max-extent=-1`. For the HAPBIN programs (i.e. *ihsbin* and *xpehhbin*), the EHH and EHHs cut-off values (defined to stop the calculation of unstandardized iHS and iES) were set to 0.05 (i.e. the default value in SELSCAN and REHH). For all programs, the standardization of iHS was performed with allele frequency bins of 0.01, as controlled by the `freqbin` argument in the `ihh2ihs()` function of the REHH package, and the `bins` argument for the program `norm` of the SELSCAN package and the program *ihsbin* of the HAPBIN package. The command lines used for the different programs, together with the corresponding input data files are provided in the Appendix S1 (Supporting information).

Finally, for each analysis and parameter set, the real (actual elapsed) computation time (provided by the Unix command `time`) was averaged over ten independent runs. All analyses were run on a standard computer running under Linux Debian 8.5 and equipped with an Intel® Xeon® 6-core processor W3690 (3.46 GHz, 12M cache). Note that the Unix command `taskset` was used to control the number of working threads for the analyses with the HAPBIN programs (since neither the *ihsbin* nor the *xpehhbin* programs allow to chose the number of threads to be used).

Results and discussion

Analysis of the human chromosome 2 data sets

For illustration purpose, we used REHH 2.0 to analyse two human data sets consisting of 236 and 342 haplotypes of 110 200 SNPs mapping to HSA2 that were sampled in the CEU and JPT+CHB populations, respectively. As for the performance comparisons described below, default options were used except that the maximum tolerated gap (`maxgap` argument) to report the different statistics was set to 200 kb. The chromosome-wide scans of *iHS* for the CEU and the JPT+CHB populations, respectively, are plotted in Fig. 1A. The most significant SNP maps at position 134 705 895 bp for the CEU population (*iHS* = -5.36) and at position 202 230 069 bp for the JPT+CHB population (*iHS* = -4.99). The chromosome-wide scans of XP-EHH and *Rsb*, which contrast EHHs profiles between the CEU and the JPT+CHB populations,

are plotted in Fig. 1B. The most significant SNP maps at position 136 533 558 bp for *Rsb*-based test (*Rsb* = 6.16) and at position 136 523 244 bp for the XP-EHH-based test (XP-EHH = 5.64). For this latter SNP (mapping to region #5 as defined below), the haplotype bifurcation diagrams for the ancestral and derived alleles within the CEU population are plotted in Fig. 1C and Fig. 1D, respectively, using the `bifurcation.diagram()` function from the REHH package. These diagrams, introduced by Sabeti *et al.* (2002), provide an helpful visualization of the structuring of haplotype diversity around each core alleles at increasing distance (see Box Fig. 3). Note that in this example, the extent of haplotype homozygosity associated with the derived allele (Fig. 1D), relatively to that associated with the ancestral allele (Fig. 1C), is consistent with the negative *iHS* measure at this SNP (*iHS* = -3.23).

To further identify regions displaying strong footprints of selection, we split the HSA2 chromosome

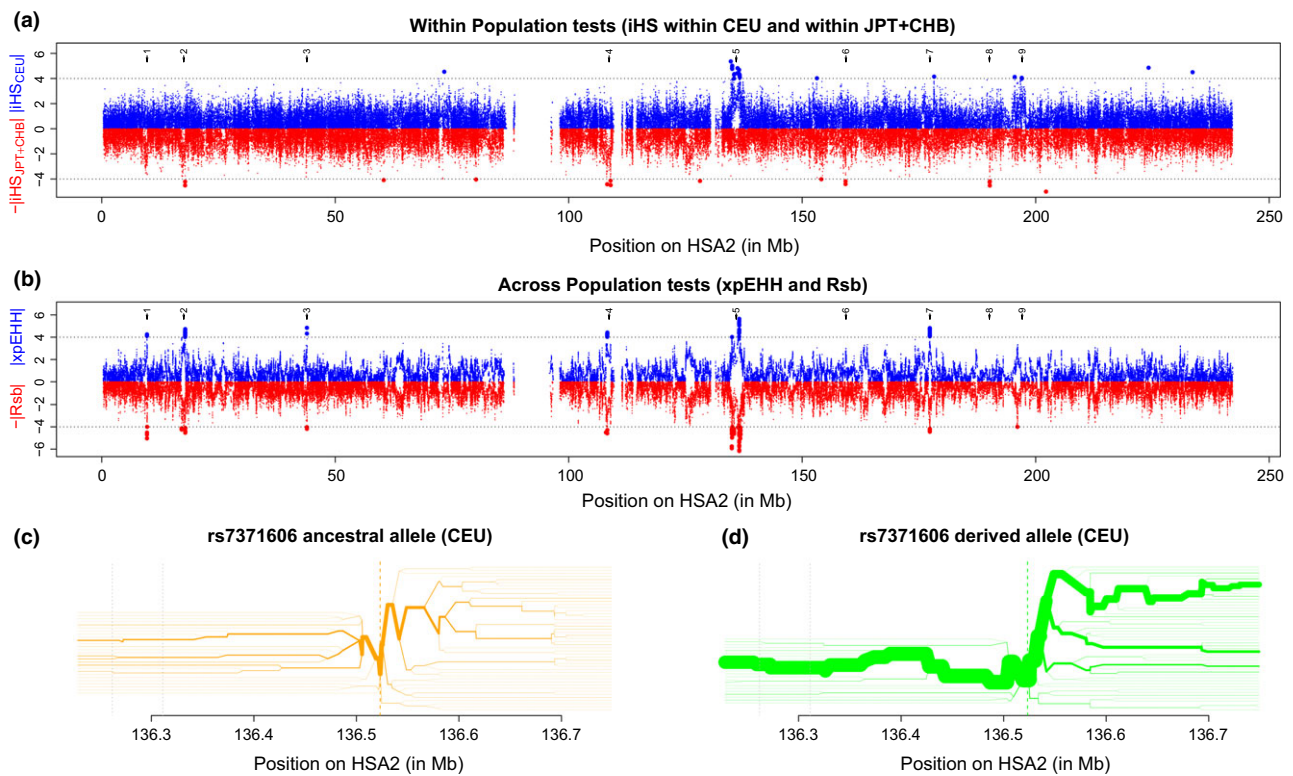


Fig. 1 Analysis of the human chromosome 2 haplotype data sets (hg18 human genome assembly) for the CEU and JPT+CHB populations with REHH 2.0. (A) Plot of *iHS* against physical distance, in the CEU ($|iHS|$ in blue) and the JPT+CHB ($-|iHS|$ in red) populations. (B) Plot of XP-EHH ($|XP-EHH|$ in blue) and *Rsb* ($-|Rsb|$ in red) between the CEU and JPT+CHB populations. In (A) and (B), the horizontal dotted lines indicate the $|iHS|$ significance threshold of 4 that was used to identify significant regions (see Table 1) and the arrows at the top of the graph indicate the mid-position of the significant regions described in Table 1. We chose to represent $|XP-EHH|$ and $-|Rsb|$ in B) for convenience. The sign of these statistics, which informs on the origin of the signal (in the CEU or the JPT+CHB population), is provided in Table 1. (C) and (D) Haplotype bifurcation diagrams drawn for the ancestral and derived allele, respectively, of the SNP rs7371606 in the CEU population (XP-EHH peak position of region #5 described in Table 1 and containing the LCT gene). In (C) and (D), the two grey vertical dotted lines delimit the LCT gene. [Colour figure can be viewed at wileyonlinelibrary.com].

Table 1 Regions of HSA2 harbouring strong signals of selection

ID	Position* (size)	Candidate gene (position)	Test	Peak position†	Selected population (overlab with other studies‡)
1	9.250–10.00 (0.75)	YWHAQ (9.641–9.688)	XP-EHH Rsb iHS _{CEU}	9.700 (–4.25; 3) 9.701 (–5.03; 4) 9.700 (2.18; 0)	JPT+CHB
2	16.75–18.25 (1.50)	MSGN1 (17.861–17.862)	iHS _{JPT+CHB} XP-EHH Rsb iHS _{CEU}	9.732 (3.64; 0) 17.871 (–4.72; 18) 17.890 (–4.52; 8) 18.150 (3.69; 0)	JPT+CHB
3	43.50–44.25 (1.75)	ABCG8 (43.919–43.959)	iHS _{JPT+CHB} XP-EHH Rsb iHS _{CEU}	17.856 (4.51; 2) 43.955 (–4.85; 2) 43.957 (–4.17; 2) 44.177 (–3.14; 0)	JPT+CHB
4	108.00–109.25 (1.25)	SULT1C2 (108.271–108.292) EDAR (108.877–108.972)	XP-EHH Rsb iHS _{CEU} iHS _{JPT+CHB}	108.273 (–4.41; 19) 108.253 (–4.58; 3) 109.016 (2.46; 0) 108.982 (4.48; 4)	JPT+CHB (Vo., Ta., Sa.)
5	134.50–137.25 (2.75)	LCT (136.262–136.311) MCM6 (136.314–136.335)	XP-EHH Rsb iHS _{CEU} iHS _{JPT+CHB}	136.523 (5.64; 17) 136.533 (6.16; 73) 134.706 (–5.36; 19) 134.727 (–3.76; 0)	CEU (Vo., Ta., Sa.)
6	159.00–159.75 (0.75)	PKP4 (159.021–159.246)	XP-EHH Rsb iHS _{CEU} iHS _{JPT+CHB}	159.381 (–2.98; 0) 159.380 (–2.88; 0) 159.745 (2.86; 0) 159.293 (4.40; 2)	JPT+CHB
7	177.00–177.75 (0.75)	n.a.	XP-EHH Rsb iHS _{CEU} iHS _{JPT+CHB}	177.338 (–4.82; 16) 177.337 (–4.43; 7) 177.336 (–2.57; 0) 177.108 (3.46; 0)	JPT+CHB (Sa.)
8	189.75–190.50 (0.75)	SLC40A1 (190.133–190.154)	XP-EHH Rsb iHS _{CEU} iHS _{JPT+CHB}	190.195 (–1.11; 0) 190.190 (–1.78; 0) 190.326 (2.94; 0) 190.177 (4.51; 3)	JPT+CHB
9	196.75–197.50 (0.75)	HECW2 (196.772–197.166)	XP-EHH Rsb iHS _{CEU} iHS _{JPT+CHB}	196.794 (2.13; 0) 196.755 (2.09; 0) 197.030 (4.05; 3) 197.332 (2.34; 0)	CEU (Ta.)

*All the position are given in Mb with respect to the hg18 human genome assembly.

†In parentheses: the value of the test statistics at the peak position; the number of SNPs in the window that have a test statistic (in absolute value) above the threshold of 4.

‡Significant tests of selection found in other studies for the same regions are indicated: Vo. stands for Voight *et al.* (2006); Ta. stands for Tang *et al.* (2007) and Sa. stands for Sabeti *et al.* (2007).

into 950 consecutive 500 kb windows (with a 250 kb overlap). Windows with at least 2 SNPs displaying an absolute value of the statistic >4 (which approximately corresponds to a two-sided P -value $<10^{-4}$, see above) for at least one of the four test statistics were deemed significant. Significant overlapping windows were then merged, leading to a total of nine regions harbouring strong signals of selection, the characteristics of which are detailed in Table 1 (see also Fig. 1). As expected, most of the regions identified here overlap with the regions identified in previously published genome scans for samples with the same origin (Voight *et al.*

2006; Sabeti *et al.* 2007; Tang *et al.* 2007) (see Table 1). For instance, regions #4 and #5 that lie, respectively, in the vicinity of the EDAR gene (under selection in Asian populations) and the LCT gene (under selection in European populations) have been extensively characterized in the literature (e.g., Peter *et al.* 2012). We detected more regions than previously reported in the aforementioned studies, most probably because our analyses are based on different assessment of significance. A more detailed description of the newly identified regions is, however, beyond the scope of the present article.

Note finally that XP-EHH- and Rsb-based scans gave consistent results, with the exception of the region in the vicinity of the LCT gene (#5 in Table 1 and Fig. 1) where a double peak was observed with Rsb (consistent with the iHS profile within CEU) and a single peak with XP-EHH. Yet, the Pearson's correlation coefficient between these statistics was equal to 0.843, which illustrates the close similarity of these two metrics.

Comparing the performances of REHH 2.0 relatively to REHH 1.13, SELSCAN and HAPBIN packages

The two CEU and JPT+CHB human data sets were further analysed with REHH 1.13 to evaluate the gain in real computation time resulting from the modifications introduced in version 2.0. Note that extensive tests were done during the development of version 2.0, to ensure that the same estimates (for the iHH and iES statistics) were obtained with both versions. Only very marginal differences were, however, sometimes observed in the estimates of iES^{Tang} . For instance, the Pearson's correlation coefficient between the resulting Rsb computed across the CEU and JPT+CHB populations with version REHH 1.13 and REHH 2.0 was found equal to 0.999992 (instead of 1.0). This is actually due to the introduction of the computation of iES^{Sabeti} in version 2.0 to estimate XP-EHH. Indeed, we chose to define the same cut-off value for both statistics during the computation of the component variable EHHS (controlled with the option `limehhs`, set to 0.05 by default).

An improved processing of the input file

The first major modification introduced in REHH version 2.0 deals with the processing of input files (haplotype and SNP informationfiles) using the function `data2haplohh()`. Our own experience with earlier versions of the package together with feedback from several users prompted us to optimize data import and to

improve allele recoding, which was inefficient in previous versions. Considering standard input haplotype file format (which is common to both versions), and with alleles encoded in the appropriate format (`{0,1,2}` for missing data, ancestral and derived alleles, respectively), the new `data2haplohh()` function is about 2.5 times faster than the previous one (see Table 2). In addition, the allele recoding option results in slightly better processing performances and corrects for some minor character conversion issues that sometimes occurred when it was used with the previous versions. Finally, the new haplotype format (with haplotypes in columns), corresponding to the output file of the SHAPEIT phasing program (O'Connell *et al.* 2014), was found to be the most efficient to process (see Table 2).

With data sets of increasing complexity and size, such improvement in the processing of input files is critical to REHH users. Processing a data set as large as the JPT+CHB one (consisting of 342 haplotype with 110 200 SNPs) now takes <12 s. Note, however, that for this file a maximum of about 1 Gb RAM was used, for a net memory size change of 240 Mb. For larger data sets, RAM requirements may therefore be limiting for some computers.

A faster and parallel algorithm to explore haplotype variability

The second major modification introduced in REHH version 2.0 concerns the core algorithm that computes the distribution of haplotype counts, which underlies the calculation of all the metrics of interest (iHS, Rsb and XP-EHH). As shown in Table 3, this new algorithm allows to decrease the computation times by more than one order of magnitude, as compared to the algorithm implemented in REHH version 1.13. Hence, for the computation of iHS in the CEU population (respectively, the JPT+CHB population) on a single thread, the real computation times were 13.7 (respectively, 21.8) times smaller on

Table 2 Comparison of the real computation times (in seconds) required to process input data files with the `data2haplohh()` function for the versions 1.13 and 2.0 of the REHH package. Two data sets consisting, respectively, of 236 and 342 haplotypes of 110 200 SNPs for the CEU and JPT+CHB populations were considered (see the main text). For each of these data sets, the table gives the average computation times \pm standard deviation) across ten independent runs, either with or without (in parentheses) allele recoding (using the option `allele.recode`)

	Haplotype format	CEU haplotypes	CHB+JPT haplotypes
REHH 1.13	Standard	>36 000* (29.97 \pm 0.29)	>36 000* (34.62 \pm 0.60)
REHH 2.0	Standard	9.858 \pm 0.39 (10.73 \pm 0.16)	14.56 \pm 0.17 (15.61 \pm 0.26)
REHH 2.0	Transposed†	7.882 \pm 0.10 (8.832 \pm 0.50)	11.80 \pm 0.20 (12.91 \pm 0.14)

*As mentioned in the manual, REHH version 1.x is quite inefficient in allele recoding. Versions 1.x are also prone to error (e.g. if some alleles are coded as "T").

†Using the new option `haplotype.in.columns=T`.

Table 3 Comparison of the real computation time (in seconds) required to compute the different EHH-based statistics for the versions 1.13 and 2.0 of the REHH, the SELSCAN and the HAPBIN packages. For each analysis, the table gives the average computation time (\pm standard deviation) across ten independent runs. For each program, analyses were run either on a single thread or on four threads (except for REHH 1.13 version, which is not parallelized)

Program	#threads	iHS _{ceu}	iHS _{chb+jpt}	XP-EHH	Rsb	Total*
REHH 1.13	1	1759 \pm 29	3045 \pm 31	n.a.	4803 \pm 58	4805 \pm 58
REHH 2.0	1	128 \pm 1.0	140 \pm 2.1	268 \pm 1.8	268 \pm 1.8	269 \pm 1.8
	4	37.8 \pm 0.3	40.2 \pm 0.3	77.1 \pm 0.5	77.1 \pm 0.5	78.5 \pm 0.5
SELSCAN	1	1237 \pm 17	1503 \pm 29	3833 \pm 100	n.a.	6573 \pm 86
	4	324 \pm 6.5	391 \pm 6.5	969 \pm 5.6	n.a.	1684 \pm 9.3
HAPBIN	1	17.6 \pm 0.2	20.0 \pm 0.1	47.4 \pm 0.2	n.a.	85.0 \pm 0.3
	4	5.68 \pm 0.7	7.42 \pm 0.1	13.2 \pm 0.0	n.a.	26.2 \pm 0.7

*In REHH, the function `scan_hh` computes iHH and iES simultaneously. It therefore needs to be run only once per haplotype data set. As a result, computing XP-EHH (and/or Rsb) requires almost no extra time, once iHS for the two populations has been computed.

average. Interestingly, the computation time for the JPT+CHB data set (which is approximately 1.34 times larger than the CEU one in terms of number of SNPs \times number of haplotype) was only 1.09 times slower than for the latter. Conversely, the real computation time was 1.73 times slower for JPT+CHB relatively to CEU with REHH version 1.13. As shown in Fig. 2, a more detailed

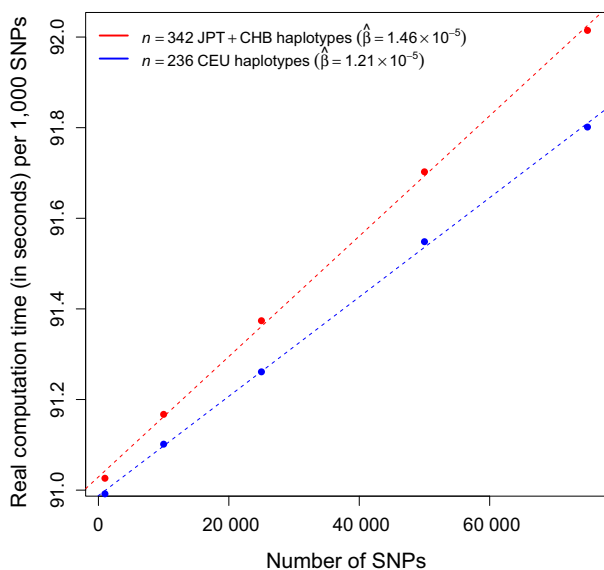


Fig. 2 Empirical profiling of the `scan_hh()` function that computes iHH and iES. Real computation times (in seconds) were estimated for 10 random samples of 1000, 10 000, 25 000, 50 000 and 75 000 SNPs, respectively, taken from the $n = 236$ CEU (blue) and $n = 342$ JPT+CHB (red) haplotypes by running the `scan_hh()` function with default options (i.e. on a single thread). For each set of SNPs, the resulting average computation time per 1000 SNPs is plotted against the corresponding numbers of SNPs. Dotted lines represent the estimated regression lines with the underlying estimated regression coefficients ($\hat{\beta}$) indicated in the legend. [Colour figure can be viewed at wileyonlinelibrary.com].

(yet empirical) profiling confirmed that the computational burden was approximately linearly related to the number of haplotypes but suggested an exponential relationship with the number of SNPs. Nevertheless, the increasing rate of the per SNP computation time remained small (<2 ms for every 100 000 additional SNPs).

To further improve computational speed, the characterization of haplotype structure is now performed using OpenMP parallelization across SNPs in genome-wide scans. Using four threads then leads to an additional decrease of about 3.5 times in computation times (see Table 3). Parallelization might alternatively be performed at a higher level, using the R package `parallel`, by analysing different chromosomes on different threads. Our motivation for a low-level OpenMP implementation was to reduce the computational burden, as well as the memory requirements: it is indeed more efficient to parallelize the computation of EHH-related statistics for a given chromosome (which requires to store haplotype data for a single chromosome in the RAM), rather than to parallelize the computation of EHH-related statistics across chromosomes (which would require to store haplotype data for several chromosomes simultaneously).

Overall, the whole analysis of the HSA2 haplotype files used in this study took about 1.5 min (including the processing of input files) with REHH 2.0 and more than 1.3 h with REHH 1.3. This corresponds to the computation of iHS within the CEU and within the JPT+CHB populations, as well as the computation of Rsb and XP-EHH.

Comparing REHH 2.0 to the SELSCAN and HAPBIN programs

Finally, we compared REHH 2.0 with SELSCAN (Szpiech & Hernandez 2014) and HAPBIN (Maclean *et al.* 2015), which were recently published. Both programs are

written in C++ language and include parallelization. Computation times for the different analyses, either on a single or four threads, are provided in Table 3. The new version of REHH outperforms SELSCAN by about one order of magnitude. Moreover, running REHH on a single thread is still more than twice as fast as running SELSCAN on four threads. It should also be noticed that running a full analysis consisting of the estimation of iHS within and XP-EHH between the CEU and JPT+CHB populations results in a significant additional burden with SELSCAN (Table 3). Conversely, HAPBIN was found to be more than five times faster than REHH 2.0, most likely as a result of its more efficient algorithm to explore haplotype variability. Yet, given the small computation times achieved by both programs, REHH 2.0 remains competitive relative to HAPBIN for most practical applications.

Correlations between the estimated iHS and XP-EHH obtained with the different programs are given in Table 4. Estimates of XP-EHH were in very good agreement among the different software packages. Similarly, estimates for iHS were almost the same between REHH 2.0 and SELSCAN but slightly depart from those obtained with HAPBIN. Although we did not further investigate the origin of these discrepancies, this might probably be related to a different definition of haplotype homozygosity in HAPBIN, as compared to Sabeti *et al.* (2007) (see the definition of EHH in the Supplementary Material of Maclean *et al.* 2015).

Conclusion

Although the R package REHH (Gautier & Vitalis 2012) has been widely used since its first release, the increasing dimension of haplotype data sets typically available in most species led to serious limitations. This stimulated the development of alternative R-free solutions (Szpiech & Hernandez 2014; Maclean *et al.* 2015). In this study, we introduced substantial changes in the REHH package to

improve its computational efficiency by one to several orders of magnitude. This was achieved by modifying the processing of the input files and, most importantly, by improving and parallelizing the core algorithm that computes the distribution of haplotype counts. As a result, REHH 2.0 clearly outperforms the SELSCAN package (Szpiech & Hernandez 2014) and competes with HAPBIN (Maclean *et al.* 2015), the fastest program to date. A decisive advantage of REHH 2.0 over these programs is that it allows working within the multiplatform R environment. As such, it benefits from several graphical tools that facilitate visual interpretation of the results.

Acknowledgements

We are grateful to all users of the previous version for their feedback that helped to improve the package. We wish to thank Emmanuel Paradis for his invitation to contribute to this special issue and three anonymous reviewers for their constructive comments on an earlier draft of this manuscript. This work was supported in part by a grant of the German Science Foundation (DFG-SFB680) to AK.

References

- Barson NJ, Aykanat T, Hindar K *et al.* (2015) Sex-dependent dominance at a single locus maintains variation in age at maturity in salmon. *Nature*, **528**, 405–408.
- Bosse M, Megens HJ, Madsen O *et al.* (2015) Using genome-wide measures of coancestry to maintain diversity and fitness in endangered and domestic pig populations. *Genome Research*, **25**, 970–981.
- Davey JW, Hohenlohe PA, Etter PD, Boone JQ, Catchen JM, Blaxter ML (2011) Genome-wide genetic marker discovery and genotyping using next-generation sequencing. *Nature Reviews Genetics*, **12**, 499–510.
- Ferrer-Admetlla A, Liang M, Korneliusen T, Nielsen R (2014) On detecting incomplete soft or hard selective sweeps using haplotype structure. *Molecular Biology and Evolution*, **5**, 1275–1291.
- Flori L, Thevenon S, Dayo GK *et al.* (2014) Adaptive admixture in the west african bovine hybrid zone: insight from the borgou population. *Molecular Ecology*, **23**, 3241–3257.
- Gautier M, Naves M (2011) Footprints of selection in the ancestral admixture of a new world creole cattle breed. *Molecular Ecology*, **20**, 3128–3143.
- Gautier M, Vitalis R (2012) rehh: an r package to detect footprints of selection in genome-wide snp data from haplotype structure. *Bioinformatics*, **28**, 1176–1177.
- Jin J, Lee M, Bai B *et al.* (in press) Draft genome sequence of an elite dura palm and whole-genome patterns of dna variation in oil palm. *DNA Research*, doi: 10.1093/dnares/dsw036.
- Maclean CA, Hong NPC, Prendergast JGD (2015) hapbin: an efficient program for performing haplotype-based scans for positive selection in large genomic datasets. *Molecular Biology and Evolution*, **32**, 3027–3029.
- Mueller JC, Kuhl H, Timmermann B, Kempnaers B (2016) Characterization of the genome and transcriptome of the blue tit cyanistes caeruleus: polymorphisms, sex-biased expression and selection signals. *Molecular Ecology Resources*, **16**, 549–561.
- O'Connell J, Gurdasani D, Delaneau O *et al.* (2014) A general approach for haplotype phasing across the full spectrum of relatedness. *PLoS Genetics*, **10**, e1004234.

Table 4 Correlation between the estimated iHS and XP-EHH statistics across the programs REHH (version 2.0), SELSCAN and HAPBIN. The pairwise Pearson's correlation coefficients for the iHS computed in the CEU and the JPT+CHB (in parenthesis) populations are given in the upper diagonal. The pairwise Pearson's correlation coefficients for the XP-EHH computed across the CEU and JPT+CHB populations are given in the lower diagonal

	REHH	SELSCAN	HAPBIN
REHH	n.a.	0.991 (0.993)	0.907 (0.945)
SELSCAN	0.985	n.a.	0.907 (0.945)
HAPBIN	0.986	0.994	n.a.

- Peter BM, Huerta-Sanchez E, Nielsen R (2012) Distinguishing between selective sweeps from standing variation and from a *de novo* mutation. *PLoS Genetics*, **8**, e1003011.
- Pickrell JK, Coop G, Novembre J *et al.* (2009) Signals of recent positive selection in a worldwide sample of human populations. *Genome Research*, **19**, 826–837.
- R Development Core Team (2016) *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria. Available from: <http://www.R-project.org/>
- Roesti M, Kueng B, Moser D, Berner D (2015) The genomics of ecological vicariance in threespine stickleback fish. *Nature Communications*, **6**, 8767.
- Sabeti PC, Reich DE, Higgins JM *et al.* (2002) Detecting recent positive selection in the human genome from haplotype structure. *Nature*, **419**, 832–837.
- Sabeti PC, Varilly P, Fry B, Lohmueller J, Hostetter E *et al.* (2007) Genome-wide detection and characterization of positive selection in human populations. *Nature*, **449**, 913–918.
- Staab PR, Metzler D (2016) Coala: an r framework for coalescent simulation. *Bioinformatics*, **32**, 1903–1904.
- Szpiech ZA, Hernandez RD (2014) selscan: an efficient multithreaded program to perform ehh-based scans for positive selection. *Molecular Biology and Evolution*, **31**, 2824–2827.
- Tang K, Thornton KR, Stoneking M (2007) A new approach for using genome scans to detect recent positive selection in the human genome. *PLoS Biology*, **5**, e171.
- The International HapMap3 Consortium (2010) Integrating common and rare genetic variation in diverse human populations. *Nature*, **467**, 52–58.
- Vitti JJ, Grossman SR, Sabeti PC (2013) Detecting natural selection in genomic data. *Annual Review of Genetics*, **47**, 97–120.
- Voight BF, Kudaravalli S, Wen X, Pritchard JK (2006) A map of recent positive selection in the human genome. *PLoS Biology*, **4**, e72.
- Wang M, Yu Y, Haberer G *et al.* (2014) The genome sequence of african rice (*Oryza glaberrima*) and evidence for independent domestication. *Nature Genetics*, **46**, 982–988.

R.V. wrote, reviewed and revised the current manuscript and vignette.

Data accessibility

REHH 2.0 is available from the CRAN repository (<http://cran.r-project.org/web/packages/rehh/index.html>). A help file together with a detailed vignette manual (the current version is provided as a Appendix S2, Supporting information) is included in the package.

Example data set: The input haplotype data and SNP information files (in REHH, SELSCAN and HAPBIN format) are provided in the Appendix S1 (Supporting information), as a compressed archive named `FileS1.tar.gz`. This archive also contains command lines that were used to run the different programs.

Supporting Information

Additional Supporting Information may be found in the online version of this article:

Appendix S1 Compressed archive named `FileS1.tar.gz` containing example input haplotype data and SNP information files in the REHH, SELSCAN and HAPBIN format.

Appendix S2 Detailed user manual (vignette) for the REHH 2.0.

M.G. and R.V. conceived the package. A.K. reimplemented the haplotype counting algorithm. M.G., A.K., and

3 Discussion

The four articles reprinted in the last chapter will be commented on each separately, followed by general conclusions.

3.1 Demography-adjusted tests of neutrality

The study on demography-adjusted tests can be regarded as an amalgamation of two previous achievements: the adaptation of TAJIMA's D and a few similar tests to allow for non-trivial demographies, carried out by Živković and Wiehe [2008], and the observation that these tests, in their standard version, can be regarded as instances of a whole family by [Achaz, 2009]. We extended the framework of the latter to include the demography-adapted versions of the former. This placement into a unifying scheme substantially facilitates the computational implementation of the tests as well as the development of further tests of the same family.

Additionally, we were able to estimate the parameters of a simple model of stepwise population size changes for the non-admixed populations of phase 1 of the 1000 genomes project. The integration of these parameters into tests like TAJIMA's D and others yielded the desired result of approximately normalized distributions of the test statistics. However, it became clear during the course of the study, that since the adaptation of the tests is done once and then applied to the whole genome, it does not resolve the essential problem that the effects of demography may vary across the genome [Jensen et al., 2005]. Instead, our simulations led us to underline the conclusion of Živković and Wiehe [2008] that severe population size changes, unlike those we observed for humans, but like the very strong bottleneck estimated for a population of European fruit flies, essentially flatten out the distribution of the test statistics and cannot be remediated by the incorporation of demography (cf. Figure 8 therein). Nevertheless, the endeavour clarified somewhat the potential and the limits of adaptability of frequency spectrum based tests.

Another aspect was the comparison of a whole-genome scan for selection using TAJIMA's D with a study performed roughly ten years earlier [Carlson et al., 2005]. Although we tried to re-apply faithfully their methods, the regions found, called "contiguous regions of TAJIMA's D reduction (CRTR)" showed meagre overlap between corresponding populations of both studies. Without doubt this partly owes to different sources of data: our sample sizes were roughly four times larger and only the European samples represented the same population; and we used unbiased SNPs, deriving from whole-genome sequencing, yielding an approximately 20 times higher number of SNPs. Furthermore, we found that the definition of CRTR is vulnerable to slight variations within the data. In order to estimate the sample noise, we computed for each population the CRTRs in four random sub-samples of similar size as those in Carlson et al. [2005] and found that only half of the CRTRs of each sub-sample were shared with the remaining three sub-samples.

The TAJIMA's D values calculated by Carlson et al. [2005] are currently (march 2018) still available as tracks within the UCSC Genome Browser [Kent et al., 2002] for the human genome assemblies hg17 and hg16, while for newer assemblies no corresponding tracks are offered by UCSC and only for three populations by others [Pybus et al., 2014]. To fill this gap, the test statistics calculated by us have been reformatted as tracks for assembly hg19. They are available via the homepage of the Bioinformatics group at the Institute of Genetics, University of Cologne, and can be added either by upload or a simple URL

copy&paste to the UCSC Genome Browser track panel. Figure 3.1 presents screen-shots of those tracks for three populations around the gene *EDAR*, which in both scans formed part of a CRTR. The values of Carlson et al. [2005] are skewed towards higher values, reflecting the ascertainment bias caused by the over-representation of common variants in the SNP array used for genotyping. The influence of population demography, which is supposed to affect the whole genome, can be seen in the difference between the tracks in the middle showing the standard test values and those at the bottom which derive from the demography-adjusted tests. Although the effect is weak, it is tempting to see in the slightly negative values of the original TAJIMA's D in the African population a weak signal for selection, although, in fact, demography included, they oscillate around zero. The tracks of Carlson et al. [2005] and ours look very similar, implying that their information content is comparable. Thus, a disproof of the approach of Carlson et al. [2005] on the ground of SNP ascertainment bias as “largely meaningless” [Wang et al., 2006] seems not warranted.

3.2 The neutral frequency spectrum of linked sites

The joined spectrum of two linked loci within a non-recombining genomic region together with a classification of the possible relations between the loci allows to define one-dimensional spectra conditioned on the existence of a certain focal mutation. These spectra differ markedly from the unconditional spectra. With other words, an average frequency spectrum containing a mutation of a certain size does not conform to the $\frac{1}{x}$ rule of the unconditional spectrum. This causes a distortion of test statistics like TAJIMA's D with the effect being largest for an almost fixed focal mutation (Figure 4 of the article). We show a way to modify such neutrality tests to accommodate for conditional spectra.

The development of tests for a conditional spectrum was aimed for cases where a naturally outstanding mutation can be taken as the focal mutation such as a chromosomal introgression or inversion. Introgressions can occur, if sister species are not yet entirely separated and matings lead to a certain hybridization [Hedrick, 2013]. Modern humans, for instance, have experienced introgressions from Neanderthals and Denisovans, identified and scanned for selection with a specialized set of summary statistics by Racimo et al. [2017]. Additionally, there is evidence that polymorphic inversions abound in humans [Sudmant et al., 2015b]. However, the amount of reliable data for inversions lags behind other structural variants such as copy number variation, for which more effective experimental methods are available [Feuk et al., 2006] and on which already whole-genome scans for selection have been undertaken [Sudmant et al., 2015a]. In principle, inversions can be recognized by comparing two alternative assemblies of the human genome [Tuzun et al., 2005; Feuk et al., 2005; Vicente-Salvador et al., 2017]. This is laborious, though, and within the 1000 genomes project they were inferred computationally from the alignment of short and long reads to the reference sequence, yielding 768 predicted polymorphic inversions of size 250b-50kb [Sudmant et al., 2015b]. The false positive rate of this approach is high, though, and a validation by other experimental methods needed [Vicente-Salvador et al., 2017]. A few individual inversions have been characterized in detail [Antonacci et al., 2009]. A particularly well studied case is an inversion at the chromosomal position 17q21.31, which was discovered through its extended strong linkage disequilibrium, caused presumably by substantially reduced (but not entirely absent [Deng et al., 2011]) recombination between the two arrangements [Stefansson et al., 2005]. One of the orientations turned out to be rare in Africans and Asians, yet obtaining frequencies up to 33.7% in Southern Europeans [Alves et al., 2015]. This variant is clearly of medical relevance: it is positively correlated with fertility in women [Stefansson et al., 2005] while at the same time, it is prone to a further structural mutation, a so-called *microdeletion*, causing mental retardation [Rao et al., 2010; Boettger et al., 2012].

Figure 3.2 supplements the evolutionary scenarios of Figure 1.2 by an inversion. The standard frequency spectrum would be dominated by fixed differences between introgressed and non-introgressed parts and

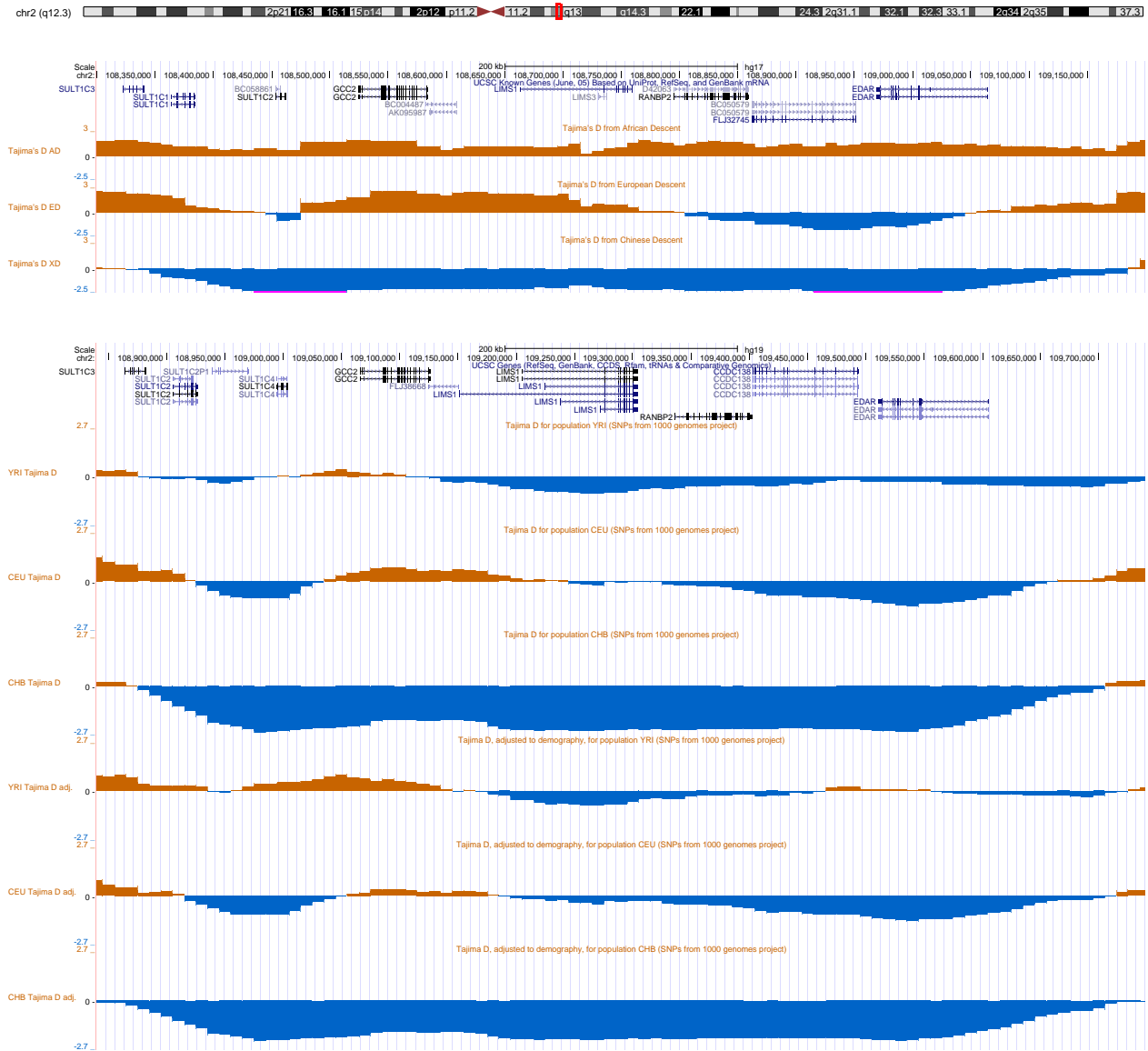


Figure 3.1: TAJIMA'S D values in a genomic region around the gene *EDAR* on human chromosome 2, rendered by the UCSC Genome Browser [Kent et al., 2002]. Top: tracks by Carlson et al. [2005] for Americans of European (ED), African (AD) and Chinese (XD) ancestry, coordinates 108.300.000-109.200.000 in hg17. Middle/bottom: tracks of original and adjusted values from Rafajlović et al. [2014] for populations YRI, CEU and CHB, coordinates 108.840.000-109.740.000 in hg19. The values were calculated in sliding windows of size 100kb with an offset of 10kb.

3 Discussion

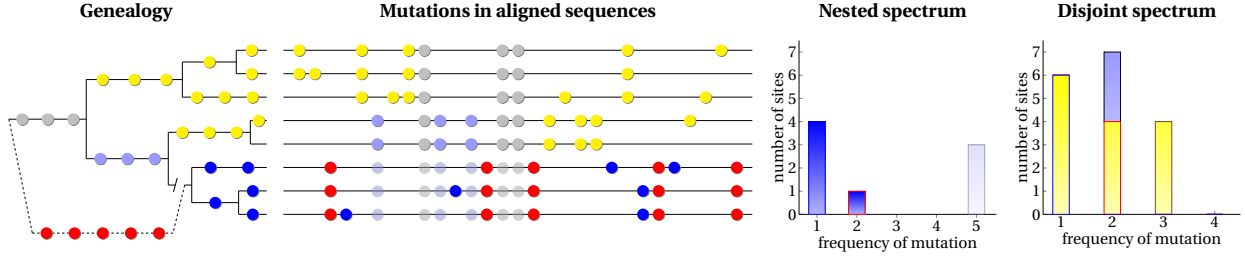


Figure 3.2: Schema of an introgression. The dashed line at the root symbolizes a split from a common ancestor to another species not entirely separated, which hybridizes at a later stage. Both species accumulate fixed mutations during this supposedly relatively long time, marked in grey and red, respectively. The introgressed region itself or fixed mutations on it can be taken as focal mutation(s). In reality fixed mutations would make up the bulk of variation seen in the sample, and in fact, enable the inference of an occurred introgression. However they do not confer information about the evolution of that region after the event. This information is contained only in the conditional spectra of nested and disjoint mutations. A further complication arises because mutations that are in reality “enclosing” cannot be recognized as such from the aligned sequences. In order to model the observable spectra correctly they have to be ascribed as depicted to disjoint mutations.

tests based on it would be likely to infer mistakenly balancing selection. Since mainly the fate of the introgression in its now host population is of interest, only mutations that still segregate within either the introgressed or the non-introgressed part of the sample are informative. These we baptized *strictly nested* and *strictly disjoint* respectively, the focal mutation consisting of the inversion itself (see Figure 1 of the article). They constitute the two frequency spectra seen on the right of Figure 3.2. Inversions, by contrast, could in principle simply be re-oriented to yield a standard full frequency spectrum. However, only 20% of the putative inversions found by Sudmant et al. [2015b] conformed to a model of clear-cut re-orientation, while the bulk was accompanied by further structural rearrangements.

In the following, an example is given on how a test based on the frequency spectrum, yet conditioned on a specific focal mutation might look like.

The expected spectra of a conditional spectrum are given in the article. For a focal mutation of size k and further mutations of size l the “strictly nested” and “strictly disjoint” mutations have expectation values of respectively

$$E \left[\xi_{l|k}^{(n)} \right] = \frac{1}{2} (\beta_n(l) - \beta_n(l+1)) k \theta \quad \text{for } l = 1, \dots, k-1 \quad (3.1)$$

$$E \left[\xi_{l|k}^{(d)} \right] = \left(\frac{1}{l} - \frac{1}{2} (\beta_n(k) - \beta_n(k+1) + \beta_n(l) - \beta_n(l+1)) k \right) \theta \quad \text{for } l = 1, \dots, n-k-1. \quad (3.2)$$

with $\beta_n(i)$ as in Eqs. (1.24), (1.25).

As mentioned above, introgressions cause the further complication that all “enclosing” mutations “survive” only on branches disjoint to the introgression. These mutations have to be added to the disjoint spectrum (see right panels of Figure 3.2) to yield a new disjoint spectrum for mutations of size $l = 1, \dots, n-k-1$:

$$E \left[\xi_{l|k}^{(d|Intro)} \right] = E \left[\xi_{l|k}^{(d)} \right] + E \left[\xi_{(l+k)|k}^{(n)} \right]. \quad (3.3)$$

We can combine now both spectra as $\xi = (\xi^{(n)}, \xi^{(d)}) = (\xi_1^{(n)}, \dots, \xi_{k-1}^{(n)}, \xi_1^{(d)}, \dots, \xi_{n-k-1}^{(d)})$ and use the frame-

work of Achaz [2008], presented in section 1.4 to construct tests using this spectrum in form of Eq. (1.12)

$$T_{\Omega} = \frac{\hat{\Theta} \cdot \Omega}{\text{Var}(\hat{\Theta} \cdot \Omega)}, \quad (3.4)$$

with estimators $\hat{\Theta} = (\frac{\xi_1}{\xi_0^0}, \dots, \frac{\xi_{n-2}}{\xi_{n-2}^0})$ relying on the combined expected frequency spectra $\xi_1^0, \dots, \xi_{n-2}^0$ of nested and disjoint mutations. For any weighting Ω the nominator can be calculated by Eq. (8) of the first reprinted article

$$\text{Var}(\hat{\Theta} \cdot \Omega) = \theta \sum_{i=1}^{n-1} \frac{\Omega_i^2}{\xi_i^0} + \theta^2 \sum_{i,j=1}^{n-1} \frac{\Omega_i}{\xi_i^0} \sigma_{ij}^0 \frac{\Omega_j}{\xi_j^0}, \quad (3.5)$$

where σ_{ij}^0 are the quadratic terms of the covariance matrix $\text{Cov}[\xi_i, \xi_j]$ for $\theta = 1$. Under neutrality, these covariances are a corollary of the third moments presented in the third article.

3.3 The third moments of the site frequency spectrum

The main result of this article is an extension of Eq. (1.22) to the third moments:

$$E[\xi_h \xi_i \xi_j] = \delta_{h=i=j} \frac{1}{i} \theta + (\delta_{h=i} \tau_{ij} + \delta_{i=j} \tau_{hj} + \delta_{j=h} \tau_{hi}) \theta^2 + \tau_{hij} \theta^3 \quad (3.6)$$

with terms τ_{ij} as defined by Eq. (1.23) and new terms τ_{hij} given by

$$\tau_{hij} = \sum_{\text{Permutations}(h,i,j)} t_{aa}(h, i, j) + t_{ab}(h, i, j) + t_{ba}(h, i, j) + t_{bb}(h, i, j). \quad (3.7)$$

The functions t_{aa} , t_{ab} , t_{ba} and t_{bb} contain less closed expressions than the corresponding functions t_a and t_b of Eqs. (1.24) and (1.25) for the second moments, but are computationally tractable for sample size n up to the order 10^3 .

A simple corollary of the third moments are the second moments of conditional spectra, i.e. the covariance of two mutations of size i and j , given a third mutation of size h , which can be subdivided into nested and disjoint parts: $E[\xi_{i|h}^{(n)} \xi_{j|h}^{(n)}]$, $E[\xi_{i|h}^{(n)} \xi_{j|h}^{(d)}]$, $E[\xi_{i|h}^{(d)} \xi_{j|h}^{(n)}]$ and $E[\xi_{i|h}^{(d)} \xi_{j|h}^{(d)}]$, given by Eq. (25) of the article.

A weighting scheme for a test on neutrality using the combined nested and disjoint conditional spectra as in Eq. (3.4) can thus be calculated by the following way: let $\xi_i^0 := E[\xi_i | \theta = 1]$ be any expected spectrum (conditional or not) under a null-hypothesis, C the matrix $c_{ij} = c_{ij}(\theta) = \text{Cov}[\xi_i, \xi_j] = \delta_{ij} \xi_i^0 \theta + \sigma_{ij}^0 \theta^2$ and ξ_i^A the spectrum of an alternative scenario. A test of the form 1.12 can be “optimized” for the detection of the alternative scenario by the weights given in Ferretti et al. [2010b, Eq. (S21)]

$$\Omega_i = \frac{\sum_j \xi_i^0 c_{ij}^{-1} \xi_j^A}{\sum_i \sum_j \xi_i^0 c_{ij}^{-1} \xi_j^A} - \frac{\sum_j \xi_i^0 c_{ij}^{-1} \xi_j^0}{\sum_i \sum_j \xi_i^0 c_{ij}^{-1} \xi_j^0}, \quad (3.8)$$

which reduces in the limit $\theta \rightarrow 0$ to

$$\Omega = \frac{1}{\sum_j \xi_j^A} \xi^A - \frac{1}{\sum_j \xi_j^0} \xi^0. \quad (3.9)$$

In the following, a small, analytically tractable example of such a test will be given. Assume a very simplified scenario of two variants under long-term balancing selection so strong that the affected region behaves like that of two separated populations without migration. Assume further that each “population”

3 Discussion

evolves neutrally, so that we can select any of the two variants as focal mutation. Then, strictly nested as well as strictly disjoint variants both display a neutral spectrum. If we finally assume that the sample frequency of the focal variant $\frac{k}{n}$ approximates its population frequency, then the conditional spectrum has expectation values

$$E[\xi_{l|k}^{(n)}] = \frac{1}{l} \frac{k}{n} \theta \quad \text{for } l = 1, \dots, k-1, \quad (3.10)$$

$$E[\xi_{l|k}^{(d)}] = \frac{1}{l} \frac{n-k}{n} \theta \quad \text{for } l = 1, \dots, n-k-1. \quad (3.11)$$

This scenario will be contrasted with one where the focal mutation rose purely by genetic drift to its current frequency. In this case, nested and disjoint spectra are given by Eqs. (3.1, 3.2).

The above test will be compared with a simpler one which goes back to an idea of Stefansson et al. [2005]. They compared the amount of variation contained within the two opposite parts of an inversion, respectively. Critical values were obtained by simulations. It is possible to integrate this idea into our general framework. Let k be the frequency of the focal mutation, $S^{(n)}$ the number of strictly nested variants and $S^{(d)}$ the number of strictly disjoint sites. Under long term balancing selection we expect to yield $\frac{S^{(n)}}{S^{(n)}+S^{(d)}} \approx \frac{k}{n}$ or $\frac{n}{k} S^{(n)} - \frac{n}{n-k} S^{(d)} \approx 0$. We can create two estimators of θ analogous to Watterson's $\hat{\theta}_S$:

$$\hat{\theta}_{S^{(n)}} = \frac{S^{(n)}}{\sum_{i=1}^{k-1} \xi_{i|k}^{0(n)}} = \frac{\sum_{i=1}^{k-1} \xi_{i|k}^{(n)}}{\sum_{i=1}^{k-1} \xi_{i|k}^{0(n)}} = \sum_{i=1}^{k-1} \omega_i^{(n)} \frac{\xi_{i|k}^{(n)}}{\xi_{i|k}^{0(n)}} = \sum_{i=1}^{k-1} \omega_i^{(n)} \hat{\theta}_{i|k}^{(n)} \quad (3.12)$$

$$\hat{\theta}_{S^{(d)}} = \frac{S^{(d)}}{\sum_{i=1}^{n-k-1} \xi_{i|k}^{0(d)}} = \frac{\sum_{i=1}^{n-k-1} \xi_{i|k}^{(d)}}{\sum_{i=1}^{n-k-1} \xi_{i|k}^{0(d)}} = \sum_{i=1}^{n-k-1} \omega_i^{(d)} \frac{\xi_{i|k}^{(d)}}{\xi_{i|k}^{0(d)}} = \sum_{i=1}^{n-k-1} \omega_i^{(d)} \hat{\theta}_{i|k}^{(d)} \quad (3.13)$$

with $\omega_i^{(n)} = \frac{\xi_{i|k}^{0(n)}}{\sum_{j=1}^{k-1} \xi_{j|k}^{0(n)}}$ for $i = 1, \dots, k-1$ and $\omega_i^{(d)} = \frac{\xi_{i|k}^{0(d)}}{\sum_{j=1}^{n-k-1} \xi_{j|k}^{0(d)}}$ for $i = 1, \dots, n-k-1$. The difference of the two estimators is described by the weighting $\Omega = (\omega_1^{(n)}, \dots, \omega_{k-1}^{(n)}, -\omega_1^{(d)}, \dots, -\omega_{n-k-1}^{(d)})$, thereby subsuming the "simple" test into the general framework of Achaz [2009].

The tests presented in introductory section 1.4 all assume neutral evolution as null hypothesis and some selection scenario as alternative. However, in contrast to selective sweeps, the time scales of neutral evolution and balancing selection largely overlap, since two variants can co-exist for a long time even without any selection. Hence any test designed to distinguish between the two scenarios is expected to have limited power. Simulations show, that it is easier to reject balancing selection than neutrality and for this reason, the power of the tests, using the former as null hypothesis, is presented in Figure 3.3. The first proposed test can be applied on the unfolded spectrum of nested and disjoint variants or on the corresponding folded spectra, where instead of the derived variant frequency the minor variant frequency is taken (formulas not shown). For the second, "simple" test that makes no difference. It can be seen that the folded conditional spectra does not give any added value to the mere number of nested and disjoint segregating sites. For values of $k < \frac{n}{2}$, the first test can exploit the extra information contained in the unfolded conditional frequency spectrum to yield a higher power.

3.4 A reimplementaion of the R package REHH

The main difference between versions 1.xx and 2.0 of the R package REHH consists in an increase of performance. Although technical in nature, it is a prerequisite to perform analyses on data sets of currently

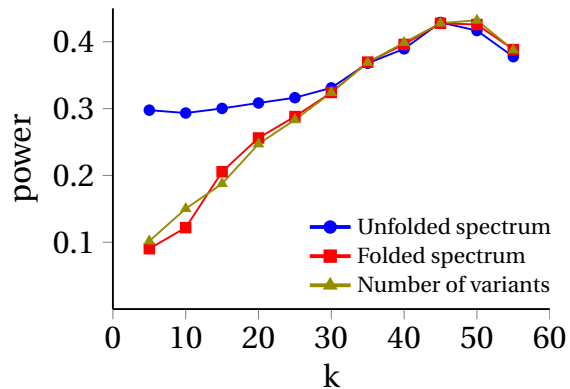


Figure 3.3: The power of the two tests defined in the text, calculated by coalescent simulations with $\theta = 100$. The null hypothesis of balancing selection is tested against the alternative of neutral evolution. The sample size is n , the frequency of the focal mutation k . The significance level chosen was $\alpha = 0.05$ on one side. Used are only strictly nested and strictly disjoint mutations. “Unfolded” means that these can be polarized, “folded” means that they cannot (first test). The second, “simple” test uses only the number of variants and ignores polarization.

available size. While the first package needed up to a month to evaluate SNP data of a single chromosome in a sample of 100 individuals, the updated package can cope with the whole genome and same sample size within hours. The jump in computing velocity owes to the implementation of an efficient algorithm as well as the introduction of multi-threading. It is unlikely that major further gains in speed can be achieved. Current lines of software development hence focus on a broadening of the range of processible data. At present, the input data, consisting usually of SNPs, needs to be polarized and phased beforehand. Although nowadays a standard procedure for most human SNPs, both cannot be easily undertaken for less well investigated species or fragmentary sequencing/genotyping. Hence the need to relax some of the assumptions underlying the implemented statistics. For instance, if sequences are not phased, the test statistics like *EHH* can still be calculated on individuals that are homozygous at the SNP under investigation [Wang et al., 2006], although such a restriction leads inevitably to an appreciable loss of statistical power.

3.5 Conclusions

Since the inception of the “neutral theory” during the 1960 years, there is an on-going dispute over the limits of its validity, namely the portion of variants evolving under the influence of selection [Kimura, 1983; Nei, 2013]. Instead of “directly addressing the problem” by application of neutrality tests [Williamson et al., 2007], the argument is now over the amount of their “false positives” [Nei et al., 2010; Barrett and Hoekstra, 2011]. The reason to doubt claims about selection based on summary statistics of genomic data owes chiefly to the scarce mutual overlap between findings of individual studies [Akey, 2009; Hermisson, 2009]. Surely, the latter can be partly explained by the somewhat arbitrary attribution of the selection scenario to extreme values of test statistics, be it with or without comparison to a simulated null model [Kelley et al., 2006]. In any case, extreme values, even if called “outliers”, are often not isolated, but part of a continuous distribution and statistical noise in the data might easily lead to a slight change in the order of values which in turn can yield very different sets of “candidate regions” as discussed in the first reprinted article. Yet behind this discussion, admittedly of little worry for the practical geneticist or medical doctor, lie different conceptions on evolution in general. I’ll try to sort them in the following way:

- α) Frequent and recurrent selective sweeps affect the whole genome [Braverman et al., 1995]. The fate of a neutral mutation is either to be driven to fixation or to be swept away, depending on whether it arises in linkage with a selected mutation or not. This concept is known under the name “genetic draft” [Gillespie, 2004] and is appropriate for very fast evolving organisms like HIV, but not for humans [Neher, 2013].
- β) Classic selective sweeps are the dominant form of evolution, but affect only a minor part of the

3 Discussion

genome, say 10% in humans [Williamson et al., 2007; Enard et al., 2014].

- γ) Classic selective sweeps are just one of several forms of evolution by selection. Others include a softening of the sweep model in the sense that selection does not continue from the emergence of a mutation until its fixation, but alternates with phases of neutral evolution. Moreover, quantitative traits may be governed by polygenic selection [Pritchard et al., 2010; Hernandez et al., 2011; Stephan, 2016].
- δ) Mutations can be advantageous only in the appropriate genomic background, which in turn evolves neutrally. In this view, it is the context that decides whether a mutation is selected for. Neutral mutations can be functionally as important as selected ones and hence the hoped-for prioritization of variants by the application of neutrality tests is flawed [Nei, 2013].

In the following, I'll comment on three specific points that are related to the above topic.

3.5.1 Whole genome scans help to assess test values at individual loci

Presumably for the relatively long time that the frequency spectrum keeps traces of past selection [Sabeti, 2006], Stoneking [2017] filed the tests based on it under species-wide detection of selection and reserved regional selection for haplotype based tests. He gave two examples of genes claimed to have been under selection in the entire human species, yet admitted that the claims had proven erroneous. Nevertheless, both cases are instructive. The first concerns a variant in the gene *PRNP*, supposedly experiencing in the past world-wide balancing selection. Since in a well-studied case in Papua New Guinea balancing selection on that gene was shown to be an indirect consequence of ritual cannibalism, this cultural practise was implied for mankind in general. It turned out, though, that simple ascertainment bias of the used world-wide SNP data set caused the gene to show an over-representation of middle frequency variants, mistaken as signal for balancing selection [Soldevila et al., 2006]. The other case regards two mutations in the gene *FOXP2*, associated with articulation and comprehension of spoken language. Allegedly, these mutations were fixed only recently in humans under the influence of strong positive selection [Enard et al., 2002]. This was spectacularly disproven by the authors themselves who found the modern variants in the sequence of Neanderthals, excluding thereby a recent origin [Krause et al., 2007]. The reason why the author's case for selection failed, seems not to have been investigated so far. However, a closer look on their data suggests that neglect of population structure is to blame (Figure 3.4).

It is highly noteworthy that both errors could have been avoided, had whole-genome data been available for comparison at the time. Although anecdotal, these two cases are well in line with the observation of Sabeti [2006] that many early studies, analyzing a single gene, had come up with similar claims, of which most could not be corroborated later by genome-wide data. Incidentally, the two genes *G6PD* and *CD40LG*, used by Sabeti et al. [2002] to introduce the *LRH* test and demonstrate its power, are not among the, admittedly restrictive, set of candidate loci found by its whole genome application by The international HapMap Consortium [2005].

3.5.2 Recent completed selective sweeps are rare in humans

Some studies explicitly claim to detect completed selective sweeps. For instance the SWEEPfinder “considers a model of a complete selective sweep in which the beneficial allele reaches a frequency of 100%” [Williamson et al., 2007]. Yet a single completed selective sweep should lead to at least dozens of fixed derived variants. Assuming that recent sweeps are predominantly caused by regional selective pressures, as is sub-understood in studies that concentrate on regional populations, these sweeps should lead to many SNPs with extreme differentiation among subpopulations. However, as tables 3.1 and 3.2 show, only a handful of SNPs do so except for comparisons between African and East Asian populations. But

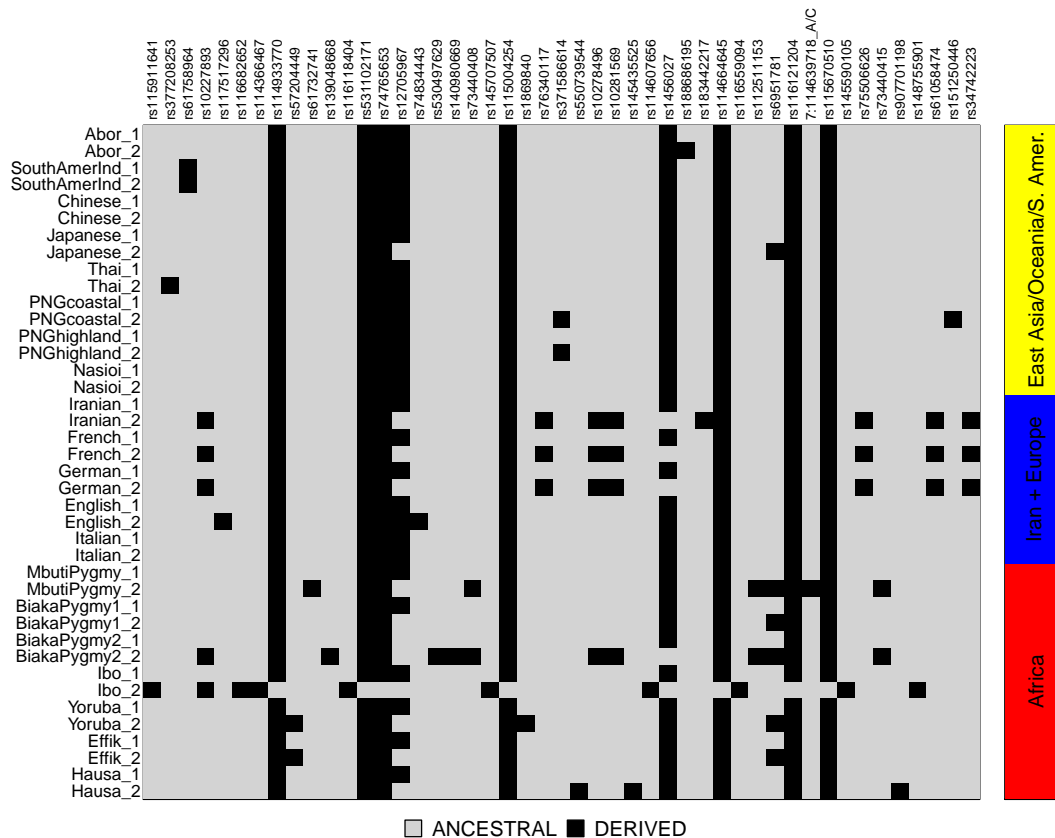


Figure 3.4: Haplotype structure within the gene *FOXP2*. Enard et al. [2002] sequenced a region of 14kb within an intron of the gene. Here, the variants found (*ibid.*, spreadsheet in supplementary material) are given in graphical form. On the y axis are the 20 sequenced individuals. On the x axis are the variant identifiers of dbSNP [Sherry, 2001]. One variant is not (yet) contained in this database and its coordinates with respect to human assembly hg38 are given instead. There are 7 derived variants of size $n-1=39$ and another high frequency derived variant of size 35. These cause the value of *FAY&WU's H* to be significant at the 5% level. As main evidence for strong selection served the value of *TAJIMA's D* which was with -2.2 significant at the 1% level. Although the possibility of confounding population structure was acknowledged, it was deemed negligible. However, a closer view on geographical regions shows that it is not: *TAJIMA's D* in the regions East Asia/Pacific/South America, Europe+Iran and Africa has a value of -1.7 , 0.6 and -1.8 respectively. The positive sign for the Europe+Iran subsample is inconsistent with species-wide selection.

3 Discussion

	Africa					Europe					South Asia					East Asia					
	ESN	GWD	LWK	MSL	YRI	CEU	FIN	GBR	IBS	TSI	BEB	GIH	ITU	PJL	STU	CDX	CHB	CHS	JPT	KHV	
Africa	ESN	-	0	0	1	6	1	7	0	6	1	6	1	1	1	10	21	11	24	10	
	GWD	0	-	0	0	6	3	5	3	4	4	5	3	2	2	1	7	3	8	8	
	LWK	0	0	-	0	2	2	1	1	2	4	2	2	2	2	1	2	2	2	1	
	MSL	0	0	0	-	4	2	4	2	5	3	5	1	1	1	29	7	4	8	6	
	YRI	1	3	0	2	-	0	0	0	0	3	0	2	2	3	55	13	9	12	1	
Europe	CEU	6	6	2	4	0	-	0	0	0	8	1	6	7	9	2	2	1	1	2	
	FIN	1	3	2	2	0	0	-	0	0	0	0	0	0	0	1	1	1	1	1	
	GBR	7	5	1	4	0	0	0	-	1	0	6	0	5	6	2	1	1	1	1	
	IBS	0	3	1	2	0	0	0	1	-	0	1	0	0	0	1	1	1	0	1	
	TSI	6	4	2	5	0	0	0	0	-	5	0	6	5	6	1	1	1	1	1	
South Asia	BEB	1	4	4	3	3	8	0	6	1	5	-	6	0	0	0	7	0	10	8	
	GIH	6	5	2	5	0	1	0	0	0	0	6	-	7	5	0	0	0	0	0	
	ITU	1	3	2	1	2	6	0	5	0	6	0	7	-	0	0	7	0	10	7	
	PJL	1	2	2	1	2	7	0	6	0	5	0	5	0	-	0	6	0	9	7	
	STU	1	2	2	1	3	9	0	7	1	6	0	6	0	0	-	6	0	10	7	
East Asia	CDX	10	1	1	29	55	2	1	2	1	1	0	0	0	1	0	-	0	1	0	0
	CHB	21	7	2	7	13	2	1	1	1	1	7	0	7	6	6	0	-	0	0	0
	CHS	11	3	2	4	9	1	1	1	0	1	0	0	0	0	0	1	0	-	0	0
	JPT	24	8	2	8	12	1	1	1	0	1	10	0	10	9	10	0	0	0	-	0
	KHV	10	8	1	6	1	2	1	1	1	1	8	0	7	7	7	0	0	0	0	-

Table 3.1: Differentiation between non-admixed populations of the 1000 genomes project. Listed are the number of SNPs which have a F_{ST} value of 1 (autosomes only). The amount for the population pair CDX and YRI sticks out. However, 46 of these SNPs lie within the gene *EXOC6B* and 8 within the gene *DOK5*. Similarly, 24 of the SNPs from the comparison MSL with CDX lie in the gene *EXO6B*. In most other comparisons, by contrast, the SNPs appear not to be clustered.

	Africa					Europe					South Asia					East Asia					
	ESN	GWD	LWK	MSL	YRI	CEU	FIN	GBR	IBS	TSI	BEB	GIH	ITU	PJL	STU	CDX	CHB	CHS	JPT	KHV	
Africa	ESN	-	0	14	0	17	36	41	42	11	38	23	37	21	8	17	403	369	256	361	279
	GWD	0	-	12	0	17	30	22	33	4	28	8	31	8	5	5	265	241	150	284	190
	LWK	14	12	-	12	0	4	7	4	2	5	20	4	18	17	24	174	160	123	161	140
	MSL	0	0	12	-	15	30	23	27	4	31	11	35	11	5	15	297	324	288	359	244
	YRI	17	17	0	15	-	17	32	14	11	11	38	15	35	22	38	409	365	257	333	318
Europe	CEU	36	30	4	30	17	-	0	0	0	0	22	1	22	23	23	12	6	4	4	5
	FIN	41	22	7	23	32	0	-	0	0	0	13	0	13	14	13	13	3	4	4	3
	GBR	42	33	4	27	14	0	0	-	1	0	22	0	24	23	25	13	5	4	7	4
	IBS	11	4	2	4	11	0	0	1	-	0	1	0	0	0	1	7	4	2	5	3
	TSI	38	28	5	31	11	0	0	0	0	-	22	0	23	24	23	10	5	4	7	4
South Asia	BEB	23	8	20	11	38	22	13	22	1	22	-	23	0	0	0	21	28	0	30	31
	GIH	37	31	4	35	15	1	0	0	0	0	23	-	23	21	24	0	0	0	0	0
	ITU	21	8	18	11	35	22	13	24	0	23	0	23	-	0	0	23	30	0	29	30
	PJL	8	5	17	5	22	23	14	23	0	24	0	21	0	-	0	25	28	0	28	29
	STU	17	5	24	15	38	23	13	25	1	23	0	24	0	0	-	23	30	0	32	32
East Asia	CDX	403	265	174	297	409	12	13	13	7	10	21	0	23	25	23	-	0	1	0	0
	CHB	369	241	160	324	365	6	3	5	4	5	28	0	30	28	30	0	-	0	0	0
	CHS	256	150	123	288	257	4	4	4	2	4	0	0	0	0	0	1	0	-	0	0
	JPT	361	284	161	359	333	4	4	7	5	7	30	0	29	28	32	0	0	0	-	0
	KHV	279	190	140	244	318	5	3	4	3	4	31	0	30	29	32	0	0	0	0	-

Table 3.2: Differentiation between non-admixed populations of the 1000 genomes project. Listed are the number of SNPs which have a F_{ST} value of at least 0.95 (autosomes only). Almost half of the SNPs differentiating CDX and YRI stem from the genes *EXOC6B* and *DOK5*. The remaining are spread over about 20 different regions.

even there, the number of affected genomic regions is at most about 25. It may be possible that migration largely prevents the fixation of variants within a regional population. Then, still, a model that assumes fixation, is not appropriate.

However, that does not exclude the existence of on-going selective sweeps. On the contrary, the paradigmatic *LCT* locus conforms well to the sweep model. Although the variant causing lactose tolerance in Europeans, *rs4988235* [Enattah et al., 2002], is present in two African chromosomes (population GWD of the 1000 genomes project), it is likely that it arose independently in North Europeans, where its frequency is highest, and may well have been under positive selection since its emergence [Bersaglieri and Sabeti, 2004]. Among the 26 populations of the 1000 genomes project it has reached the highest frequency in CEU (74%), GBR (72%) and FIN (59%) and is completely absent in East Asians. The region repeatedly appears conspicuous in scans using a test designed to detect partial or on-going selective sweeps [Sabeti et al., 2007]. However, having a dominant effect, and assuming it will remain under selection, the variant will need another 50.000 years until it reaches the population frequency of 100% [Vitti et al., 2013]. The SWEEPfinder in fact, does detect that region, but only because “it has some power to detect recent adaptive events that deviate from the assumptions of the complete sweep model” [Williamson et al., 2007].

3.5.3 How neutral is the human genome?

Figure 3.5 shows the observed scaled SNP frequency spectrum $i\xi_i$, also referred to as variant density [The 1000 Genomes Project Consortium, 2012], of four populations from The 1000 Genomes Project Consortium [2015]. The SNPs were polarized with help of the reconstructed ancestor sequence of humans and chimpanzees, downloaded from the Ensembl webpage [Zerbinio et al., 2018], release 91. This sequence, in turn, was established by a multiple sequence alignment of the human reference and eleven other primate species. Depending on the agreement between species, the confidence of the reconstructed ancestral variant is marked as “high” or “low” (README file conveyed with the ancestral sequence). About 88% of the 1000 genome SNPs can be polarized with high confidence and another 7% with low confidence. The frequency spectra depicted include both.

Three features of the observed spectra are of note. First, all populations show an excess of low frequency variants, in particular “singletons”. This is usually ascribed to world-wide growing populations [Hernandez et al., 2007]. Although low frequency variants have the highest likelihood of being false positives [Achaz, 2008], their number is thought to be even considerably under-estimated due to the relatively low coverage used [The 1000 Genomes Project Consortium, 2015]. Second, non-African populations are known to have less variation than African populations [Stajich and Hahn, 2005; The 1000 Genomes Project Consortium, 2015], which is attributed to demography, in particular a bottleneck in non-African population during their migration out of Africa, as already mentioned in the introduction. However, this difference is limited to variants of frequency less than 0.4, while the amount of higher frequency variants is remarkably similar in all populations. Third, and most conspicuously, all frequency spectra have a peak at highest frequency variants which deserves a more-in-depth inspection in the following.

The phenomenon is known to be partly due to misidentification of ancestral variants and this part can be “corrected” [Hernandez et al., 2007]. Let d designate the average divergence from the ancestor of all modern humans to the common ancestor of humans and chimpanzees which is about half the divergence between humans and chimps, the famous 1.23% [The Chimpanzee Sequencing and Analysis Consortium, 2005]. The observed spectrum is then modelled as a function of the true, unobserved spectrum and the probability p of misidentified ancestral variants [Hernandez et al., 2007, Eq. (4)]:

$$\xi_i^{obs} = (1 - p)\xi_i^{true} + p\xi_{n-i}^{true} . \quad (3.14)$$

Since i can be replaced by $n - i$, we have in fact two equations which can be solved for the true spectrum

3 Discussion

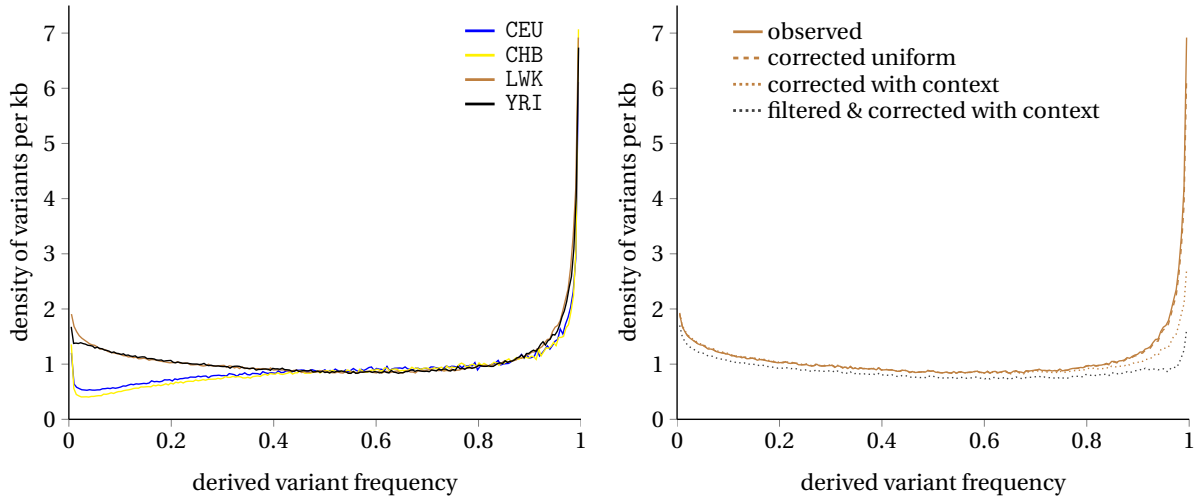


Figure 3.5: The scaled frequency spectrum or variant density $i\xi_i$. The left panel shows the observed scaled frequency spectrum of four population samples from the 1000 genomes project. The low frequency variants clearly distinguish samples of African ancestry (LWK and YRI) from samples of European (CEU) or East Asian (CHB) ancestry (after figure 2c of The 1000 Genomes Project Consortium [2012]). The right panel repeats the observed spectrum for the LWK sample together with different corrections for misidentification of ancestral variants and after filtering out regions with an exceptionally high amount of high frequency variants.

to yield [Hernandez et al., 2007, Eq. (5)]

$$\xi_i^{true} = \frac{(1-p)\xi_i^{obs} - p\xi_{n-i}^{obs}}{1-2p}. \quad (3.15)$$

As explained in the box of section 1.4, considering a uniform mutation rate, p can be approximated by $\frac{d}{3}$, however, unequal mutation rates yield a higher proportion of false polarity. Hernandez et al. [2007] offered an R-program, containing estimated mutation probabilities, to allow for all 12 possible mutations, including context dependence on the previous and next site. Both the uniform context-free and the detailed context-dependent corrections are depicted in the right panel of Figure 3.5. However, about twice as much divergence would be necessary to fully “correct” the extra-amount of derived variants with very high frequency (not shown). Hence, an explanation of the remaining “peak” is still missing. In principle, false positive singletons could contribute to it, if they happen to turn an already fixed derived variant back to an ancestral version. The probability for this to occur, assuming a sequencing error rate of $e = 0.01$, yields $\frac{1}{3}de$, which is too small to contribute appreciably to the “peak”. Another possibility is that we see here a signal of incomplete selective sweeps.

Sweeps should cause a clustering of high frequency derived variants around the selected variant. For a visual inspection, the number of variants with highest sample frequency in non-overlapping intervals, ξ_{n-1} , is depicted in Figure 3.6 (no correction is applied there). The most conspicuous values are in two consecutive windows of population LWK on chromosome 17q21.31. It turns out that these belong to the minor orientation of the inversion described above: in fact, one individual, NA19042, is heterozygous for 18 of the 21 marker SNPs proposed by Donnelly et al. [2010] to distinguish both inversion arrangements. If we see here the action of a selected sweep is unclear, though, since the inversion is estimated to have happened 2.3 million years ago and the amount of diversity seems to be not vastly different between the two variants within Africans [Steinberg et al., 2012]. Exclusion of this region from the calculation of the frequency spectrum has only a very minor effect.

Do the other “spikes” seen in Figure 3.5 represent signals of nearly completed selective sweeps? If a simple partitioning of the genome in a major part, evolving neutrally, and a minor part, governed by sweeps, is appropriate, the elimination of relatively few regions with exceptionally high numbers of derived variants should yield a neutral spectrum. However, the distribution of the number of ξ_{n-1} in non-overlapping windows (not shown), has a long “right tail”, but no outliers, and argues against such a simple dichotomy. Figure 3.5, right panel, shows a frequency spectrum where 10% of 100kb windows with high amounts of high frequency variants have been filtered out. That it is possible to influence the spectrum in this way suggests that high frequency variants are indeed concentrated at least to a certain degree in the genome. In any case, the above-mentioned suggestion that 10% of the genome is influenced by selective sweeps cannot be ruled out in this way.

3 Discussion



Figure 3.6: The number of highest derived frequency variants ξ_{n-1} in non-overlapping windows of 500kb for the 1000 genomes populations CHB, CEU, LWK and YRI. The range of the y axis is [0,200] for each track. Conspicuous is the peak in population LWK in chromosome 17, where two neighbouring windows contain each twice as much variants (beyond the limits of the y axis) as the maximum number of all remaining windows of all four populations. The extended linkage disequilibrium implied arises from an inversion with a complicated evolutionary history [Steinberg et al., 2012; Boettger et al., 2012].

4 References

- Achaz, G. Testing for neutrality in samples with sequencing errors. *Genetics*, 179(3):1409–1424, 2008.
- Achaz, G. Frequency spectrum neutrality tests: one for all and all for one. *Genetics*, 183(1):249–258, 2009.
- Adams, A. M. and Hudson, R. R. Maximum-likelihood estimation of demographic parameters using the frequency spectrum of unlinked single-nucleotide polymorphisms. *Genetics*, 168(3):1699–712, 2004.
- Akashi, H. Inferring the fitness effects of DNA mutations from polymorphism and divergence data: statistical power to detect directional selection under stationarity and free recombination. *Genetics*, 151(1):221–38, 1999.
- Akey, J. M. Interrogating a high-density SNP map for signatures of natural selection. *Genome Research*, 12:1805–1814, 2002.
- Akey, J. M. Constructing genomic maps of positive selection in humans: where do we go from here? *Genome Research*, 19:711–22, 2009.
- Alves, J. M., Lima, A. C., Pais, I. A., Amir, N., Celestino, R., Piras, G., Monne, M., Comas, D., Heutink, P., Chikhi, L., Amorim, A., and Lopes, A. M. Reassessing the evolutionary history of the 17q21 inversion polymorphism. *Genome Biology and Evolution*, 7(12):3239–3248, 2015.
- Andrés, A. M., Hubisz, M. J., Indap, A., Torgerson, D. G., Degenhardt, J. D., Boyko, A. R., Gutenkunst, R. N., White, T. J., Green, E. D., Bustamante, C. D., Clark, A. G., and Nielsen, R. Targets of balancing selection in the human genome. *Molecular Biology and Evolution*, 26(12):2755–64, 2009.
- Antonacci, F., Kidd, J. M., Marques-Bonet, T., Ventura, M., Siswara, P., Jiang, Z., and Eichler, E. E. Characterization of six human disease-associated inversion polymorphisms. *Human Molecular Genetics*, 18(14):2555–2566, 2009.
- Bamshad, M. and Wooding, S. P. Signatures of natural selection in the human genome. *Nature Reviews Genetics*, 4(2):99–111, 2003.
- Barrett, R. D. H. and Hoekstra, H. E. Molecular spandrels: tests of adaptation at the genetic level. *Nature Reviews Genetics*, 12(11):767–780, 2011.
- Baudry, E. and Depaulis, F. Effect of misoriented sites on neutrality tests with outgroup. *Genetics*, 165(3):1619–1622, 2003.
- Bersaglieri, T. and Sabeti, P. C. Genetic signatures of strong recent positive selection at the lactase gene. *The American Journal of Human Genetics*, 74:1111–1120, 2004.
- Boettger, L. M., Handsaker, R. E., Zody, M. C., and McCarroll, S. A. Structural haplotypes and recent evolution of the human 17q21.31 region. *Nature Genetics*, 44(8):881–885, 2012.
- Booker, T. R., Jackson, B. C., and Keightley, P. D. Detecting positive selection in the genome. *BMC Biology*, 15(1): 1–10, 2017.
- Bozic, I., Antal, T., Ohtsuki, H., Carter, H., Kim, D., Chen, S., Karchin, R., Kinzler, K. W., Vogelstein, B., and Nowak, M. A. Accumulation of driver and passenger mutations during tumor progression. *Proceedings of the National Academy of Sciences*, 107(43):18545–50, 2010.
- Braverman, J. M., Hudson, R. R., Kaplan, N. L., Langley, C. H., and Stephan, W. The hitchhiking effect on the site frequency spectrum of DNA polymorphisms. *Genetics*, 140(2):783–796, 1995.

4 References

- Browning, S. R. and Browning, B. L. Haplotype phasing: existing methods and new developments. *Nature Reviews Genetics*, 12(10):703–714, 2011.
- Bubb, K. L., Bovee, D., Buckley, D., Haugen, E., Kibukawa, M., Paddock, M., Palmieri, A., Subramanian, S., Zhou, Y., Kaul, R., Green, P., and Olson, M. V. Scan of human genome reveals no new loci under ancient balancing selection. *Genetics*, 173(4):2165–77, 2006.
- Bustamante, C. D., Fledel-Alon, A., Williamson, S., Nielsen, R., Hubisz, M. T., Glanowski, S., Tanenbaum, D. M., White, T. J., Sninsky, J. J., Hernandez, R. D., Civello, D., Adams, M. D., Cargill, M., and Clark, A. G. Natural selection on protein-coding genes in the human genome. *Nature*, 437(7062):1153–7, 2005.
- Cadzow, M., Boocock, J., Nguyen, H. T., Wilcox, P., Merriman, T. R., and Black, M. A. A bioinformatics workflow for detecting signatures of selection in genomic data. *Frontiers in Genetics*, 5(293):1–8, 2014.
- Carlson, C. S., Thomas, D. J., Eberle, M. A., Swanson, J. E., Livingston, R. J., Rieder, M. J., and Nickerson, D. A. Genomic regions exhibiting positive selection identified from dense genotype data. *Genome Research*, 15:1553–65, 2005.
- Cavalli-Sforza, L. L. The Human Genome Diversity Project: past, present and future. *Nature Reviews Genetics*, 6(4):333–340, 2005.
- Charlesworth, B., Morgan, M. T., and Charlesworth, D. The effect of deleterious mutations on neutral molecular variation. *Genetics*, 134(4):1289–303, 1993.
- Charlesworth, D. Balancing selection and its effects on sequences in nearby genome regions. *PLoS Genetics*, 2(4):379–84, 2006.
- Cohen, J. Relative differences: the myth of 1%. *Science*, 316(5833):689–689, 2007.
- DeGiorgio, M., Lohmueller, K. E., and Nielsen, R. A model-based approach for identifying signatures of ancient balancing selection in genetic data. *PLoS Genetics*, 10(8):1–20, 2014.
- DeGiorgio, M., Huber, C. D., Hubisz, M. J., Hellmann, I., and Nielsen, R. SweepFinder2: increased sensitivity, robustness and flexibility. *Bioinformatics*, 32(12):1895–1897, 2016.
- Deng, L., Tang, X., Hao, X., Chen, W., Lin, J., Yu, Y., Zhang, D., and Zeng, C. Genetic flux between H1 and H2 haplotypes of the 17q21.31 inversion in European population. *Genomics, Proteomics and Bioinformatics*, 9(3):113–118, 2011.
- Donnelly, M. P., Paschou, P., Grigorenko, E., Gurwitz, D., Mehdi, S. Q., Kajuna, S. L. B., Barta, C., Kungulilo, S., Karoma, N. J., Lu, R. B., Zhukova, O. V., Kim, J. J., Comas, D., Siniscalco, M., New, M., Li, P., Li, H., Manolopoulos, V. G., Speed, W. C., Rajeevan, H., Pakstis, A. J., Kidd, J. R., and Kidd, K. K. The distribution and most recent common ancestor of the 17q21 inversion in humans. *American Journal of Human Genetics*, 86(2):161–171, 2010.
- Drmanac, R., Sparks, A., and Callow, M. Human genome sequencing using unchained base reads on self-assembling DNA nanoarrays. *Science*, 327(5961):78–81, 2010.
- Dumont, B. L. and Payseur, B. A. Evolution of the genomic rate of recombination in mammals. *Evolution*, 62(2):276–294, 2008.
- Eldon, B., Birkner, M., Blath, J., and Freund, F. Can the site-frequency spectrum distinguish exponential population growth from multiple-merger coalescents? *Genetics*, 199(3):841–856, 2015.
- Enard, D., Messer, P. W., and Petrov, D. A. Genome-wide signals of positive selection in human evolution. *Genome Research*, 24:885–895, 2014.
- Enard, W., Przeworski, M., Fisher, S., Lai, C. S. L., Wiebe, V., Kitano, T., Monaco, A. P., and Pääbo, S. Molecular evolution of FOXP2, a gene involved in speech and language. *Nature*, 418(6900):869–72, 2002.
- Enattah, N. S., Sahi, T., Savilahti, E., Terwilliger, J. D., Peltonen, L., and Järvelä, I. Identification of a variant associated with adult-type hypolactasia. *Nature Genetics*, 30(2):233–237, 2002.

- Ewens, W. J. *Mathematical Population Genetics, 2nd edition*. Springer Verlag, 2004.
- Fay, J. C. and Wu, C.-I. Hitchhiking under positive Darwinian selection. *Genetics*, 155(3):1405–13, 2000.
- Ferrer-Admetlla, A., Liang, M., Korneliussen, T., and Nielsen, R. On detecting incomplete soft or hard selective sweeps using haplotype structure. *Molecular Biology and Evolution*, 31(5):1275–1291, 2014.
- Ferretti, L., Marmorini, G., and Ramos-Onsins, S. E. Properties of neutrality tests based on allele frequency spectrum. *arXiv*, 1011.1470:1–42, 2010a.
- Ferretti, L., Perez-Enciso, M., and Ramos-Onsins, S. E. Optimal neutrality tests based on the frequency spectrum. *Genetics*, 186(1):353–65, 2010b.
- Feuk, L., MacDonald, J. R., Tang, T., Carson, A. R., Li, M., Rao, G., Khaja, R., and Scherer, S. W. Discovery of human inversion polymorphisms by comparative analysis of human and chimpanzee DNA sequence assemblies. *PLoS Genetics*, 1(4):0489–0498, 2005.
- Feuk, L., Carson, A. R., and Scherer, S. W. Structural variation in the human genome. *Nature Reviews Genetics*, 7(2): 85–97, 2006.
- Fu, Y.-X. Statistical properties of segregating sites. *Theoretical Population Biology*, 48:172–197, 1995.
- Fu, Y.-X. New statistical tests of neutrality for DNA samples from a population. *Genetics*, 143(1):557–70, 1996.
- Fu, Y.-X. Statistical tests of neutrality of mutations against population growth, hitchhiking and background selection. *Genetics*, 147(2):915–925, 1997.
- Fu, Y.-X. and Li, W.-H. Statistical tests of neutrality of mutations. *Genetics*, 133(3):693–709, 1993.
- Garrigan, D., Kingan, S. B., Geneva, A. J., Andolfatto, P., Clark, A. G., Thornton, K. R., and Presgraves, D. C. Genome sequencing reveals complex speciation in the *Drosophila simulans* clade. *Genome Research*, 22:1499–1511, 2012.
- Gillespie, J. H. *Population Genetics. A concise Guide, 2nd edition*. John Hopkins University Press, 2004.
- Griffiths, R. C. and Tavaré, S. Sampling theory for neutral alleles in a varying environment. *Philosophical Transactions of the Royal Society B: Biological Sciences*, 344(1310):403–410, 1994.
- Grossman, S. R., Shlyakhter, I., Karlsson, E. K., Byrne, E. H., Morales, S., Frieden, G., Hostetter, E., Angelino, E., Garber, M., Zuk, O., Lander, E. S., Schaffner, S. F., and Sabeti, P. C. A composite of multiple signals distinguishes causal variants in regions of positive selection. *Science*, 327(5967):883–6, 2010.
- Grossman, S. R., Andersen, K. G., Shlyakhter, I. a., Tabrizi, S., Winnicki, S., Yen, A., Park, D. J., Griesemer, D., Karlsson, E. K., Wong, S. H., Cabili, M., Adegbola, R. A., Bamezai, R. N. K., Hill, A. V. S., Vannberg, F. O., Rinn, J. L., Lander, E. S., Schaffner, S. F., and Sabeti, P. C. Identifying recent adaptations in large-scale genomic data. *Cell*, 152(4): 703–13, 2013.
- Gutenkunst, R. N., Hernandez, R. D., Williamson, S. H., and Bustamante, C. D. Inferring the joint demographic history of multiple populations from multidimensional SNP frequency data. *PLoS Genetics*, 5(10):1–11, 2009.
- Haasl, R. J. and Payseur, B. A. Fifteen years of genomewide scans for selection: Trends, lessons and unaddressed genetic sources of complication. *Molecular Ecology*, 25(1):5–23, 2016.
- Hedrick, P. W. Adaptive introgression in animals: examples and comparison to new mutation and standing variation as sources of adaptive variation. *Molecular Ecology*, 22(18):4606–18, 2013.
- Hermisson, J. Who believes in whole-genome scans for selection? *Heredity*, 103(4):283–284, 2009.
- Hernandez, R. D., Williamson, S. H., and Bustamante, C. D. Context dependence, ancestral misidentification, and spurious signatures of natural selection. *Molecular Biology and Evolution*, 24(8):1792–800, 2007.
- Hernandez, R. D., Kelley, J. L., Elyashiv, E., Melton, S. C., Auton, A., McVean, G. A. T., Sella, G., and Przeworski, M. Classic selective sweeps were rare in recent human evolution. *Science*, 331(6019):920–924, 2011.

4 References

- Hinds, D. A., Stuve, L. L., Nilsen, G. B., Halperin, E., Eskin, E., Ballinger, D. G., Frazer, K. A., and Cox, D. R. Whole-genome patterns of common DNA variation in three human populations. *Science*, 307(5712):1072–1079, 2005.
- Holsinger, K. E. and Weir, B. S. Genetics in geographically structured populations: defining, estimating and interpreting Fst. *Nature Reviews Genetics*, 10(9):639–50, 2009.
- Hu, T. T., Pattyn, P., Bakker, E. G., Cao, J., Cheng, J. E., Clark, R. M., Fahlgren, N., Fawcett, J. A., Grimwood, J., Gundlach, H., Haberer, G., Hollister, J. D., Ossowski, S., Ottillar, R. P., Salamov, A. A., Schneeberger, K., Spannagl, M., Wang, X., Yang, L., Nasrallah, M. E., Bergelson, J., Carrington, J. C., Gaut, B. S., Schmutz, J., Mayer, K. F., Van De Peer, Y., Grigoriev, I. V., Nordborg, M., Weigel, D., and Guo, Y. L. The Arabidopsis lyrata genome sequence and the basis of rapid genome size change. *Nature Genetics*, 43(5):476–483, 2011.
- Huang, M., Tu, J., and Lu, Z. Recent advances in experimental whole genome haplotyping methods. *International Journal of Molecular Sciences*, 18(9):1–15, 2017.
- Hudson, R. R. and Kaplan, N. L. The coalescent process in models with selection and recombination. *Genetics*, 120(3):831–40, 1988.
- Jensen, J. D., Kim, Y., DuMont, V. B., Aquadro, C. F., and Bustamante, C. D. Distinguishing between selective sweeps and demography using DNA polymorphism data. *Genetics*, 170(3):1401–1410, 2005.
- Kelley, J. L., Madeoy, J., Calhoun, J. C., Swanson, W., and Akey, J. M. Genomic signatures of positive selection in humans and the limits of outlier approaches. *Genome Research*, 16:980–989, 2006.
- Kent, W., Sugnet, C., Furey, T., Roskin, K., Pringle, T., Zahler, A., and Haussler, D. The human genome browser at UCSC. *Genome Research*, 6:996–1006, 2002.
- Kim, Y. and Stephan, W. Detecting a local signature of genetic hitchhiking along a recombining chromosome. *Genetics*, 160(2):765–777, 2002.
- Kimura, M. Diffusion Models in Population Genetics. *Journal of Applied Probability*, 1(2):177–232, 1964.
- Kimura, M. *The Neutral Theory of Molecular Evolution*. Cambridge University Press, 1983.
- Kimura, M. and Ohta, T. The average number of generations until fixation of a mutant gene in a finite population. *Genetics*, 61(692):763–771, 1969.
- Kingman, J. On the genealogy of large populations. *Journal of Applied Probability*, 19:27–43, 1982.
- Klein, J., Sato, A., Nagl, S., and O’Higin, C. Molecular trans-species polymorphism. *Annual Review of Ecology, Evolution and Systematics*, 29:1–21, 1998.
- Krause, J., Lalueza-Fox, C., Orlando, L., Enard, W., Green, R. E., Burbano, H. a., Hublin, J.-J., Hänni, C., Fortea, J., de la Rasilla, M., Bertranpetit, J., Rosas, A., and Pääbo, S. The derived FOXP2 variant of modern humans was shared with Neandertals. *Current Biology*, 17(21):1908–12, 2007.
- Lachance, J. and Tishkoff, S. A. Biased gene conversion skews allele frequencies in human populations, increasing the disease burden of recessive alleles. *American Journal of Human Genetics*, 95(4):408–420, 2014.
- Lapierre, M., Lambert, A., and Achaz, G. Accuracy of demographic inferences from the site frequency spectrum: The case of the Yoruba population. *Genetics*, 206(1):139–449, 2017.
- Li, J. Z., Absher, D. M., Tang, H., Southwick, A. M., Casto, A. M., Ramachandran, S., Cann, H. M., Barsh, G. S., Feldman, M. W., Cavalli-Sforza, L. L., and Myers, R. M. Worldwide human relationships inferred from genome-wide patterns of variation. *Science*, 319(5866):1100–4, 2008.
- Liu, H., Prugnolle, F., and Manica, A. A geographically explicit genetic model of worldwide human-settlement history. *The American Journal of Human Genetics*, 79(2):230–237, 2006.
- Liu, X. and Fu, Y.-X. Exploring population size changes using SNP frequency spectra. *Nature Genetics*, 47(5):555–559, 2015.

- Liu, X., Maxwell, T. J., Boerwinkle, E., and Fu, Y.-X. Inferring population mutation rate and sequencing error rate using the SNP frequency spectrum in a sample of DNA sequences. *Molecular Biology and Evolution*, 26(7):1479–1490, 2009.
- Mahmoud, H. M. *Pólya Urn Models*. Chapman & Hall/CRC, Boca Raton, Florida, USA, 2008.
- Marth, G. T., Czabarka, E., Murvai, J., and Sherry, S. T. The allele frequency spectrum in genome-wide human variation data reveals signals of differential demographic history in three large world populations. *Genetics*, 166(1):351–72, 2004.
- Mathieson, I., Lazaridis, I., Rohland, N., Mallick, S., Patterson, N., Roodenberg, S. A., Harney, E., Stewardson, K., Fernandes, D., Novak, M., Sirak, K., Gamba, C., Jones, E. R., Llamas, B., Dryomov, S., Pickrell, J., Arsuaga, J. L., De Castro, J. M. B., Carbonell, E., Gerritsen, F., Khokhlov, A., Kuznetsov, P., Lozano, M., Meller, H., Mochalov, O., Moiseyev, V., Guerra, M. A., Roodenberg, J., Vergès, J. M., Krause, J., Cooper, A., Alt, K. W., Brown, D., Anthony, D., Lalueza-Fox, C., Haak, W., Pinhasi, R., and Reich, D. Genome-wide patterns of selection in 230 ancient Eurasians. *Nature*, 528(7583):499–503, 2015.
- Maynard Smith, J. and Haigh, J. The hitch-hiking effect of a favourable gene. *Genetical Research*, 23:23–35, 1974.
- McVean, G. A. T., Myers, S. R., Hunt, S., Deloukas, P., Bentley, D. R., and Donnelly, P. The fine-scale structure of recombination rate variation in the human genome. *Science*, 304(5670):581–584, 2004.
- Neher, R. A. Genetic draft, selective interference, and population genetics of rapid adaptation. *Annual Review of Ecology, Evolution and Systematics*, 44:195–215, 2013.
- Nei, M. *Mutation-driven evolution*. Oxford University Press, 2013.
- Nei, M., Suzuki, Y., and Nozawa, M. The neutral theory of molecular evolution in the genomic era. *Annual Review of Genomics and Human Genetics*, 11:265–289, 2010.
- Nielsen, R. Statistical tests of selective neutrality in the age of genomics. *Heredity*, 86:641–7, 2001.
- Nielsen, R. Molecular signatures of natural selection. *Annual Review of Genetics*, 39:197–218, 2005.
- Oleksyk, T. K., Smith, M. W., and O'Brien, S. J. Genome-wide scans for footprints of natural selection. *Philosophical Transactions of the Royal Society B: Biological Sciences*, 365(1537):185–205, 2010.
- Pavlidis, P. and Alachiotis, N. A survey of methods and tools to detect recent and strong positive selection. *Journal of Biological Research-Thessaloniki*, 24(1):7, 2017.
- Pavlidis, P., Živković, D., Stamatakis, A., and Alachiotis, N. SweeD: likelihood-based detection of selective sweeps in thousands of genomes. *Molecular Biology and Evolution*, 30(9):2224–34, 2013.
- Pickrell, J. K., Coop, G., Novembre, J., Kudaravalli, S., Li, J. Z., Absher, D. M., Srinivasan, B. S., Barsh, G. S., Myers, R. M., Feldman, M. W., and Pritchard, J. K. Signals of recent positive selection in a worldwide sample of human populations. *Genome Research*, 19:826–37, 2009.
- Pritchard, J. K., Pickrell, J. K., and Coop, G. The genetics of human adaptation: hard sweeps, soft sweeps, and polygenic adaptation. *Current biology*, 20(4):R208–15, 2010.
- Przeworski, M. The signature of positive selection at randomly chosen loci. *Genetics*, 160(3):1179–1189, 2002.
- Pybus, M., Dall'Olio, G. M., Luisi, P., Uzkudun, M., Carreño-Torres, A., Pavlidis, P., Laayouni, H., Bertranpetit, J., and Engelken, J. 1000 Genomes Selection Browser 1.0: a genome browser dedicated to signatures of natural selection in modern humans. *Nucleic acids research*, 42(Database issue):D903–9, 2014.
- Racimo, F., Marnetto, D., and Huerta-Sánchez, E. Signatures of archaic adaptive introgression in present-day human populations. *Molecular Biology and Evolution*, 34(2):296–317, 2017.
- Rafajlović, M., Klassmann, A., Eriksson, A., Wiehe, T., and Mehlig, B. Demography-adjusted tests of neutrality based on genome-wide SNP data. *Theoretical Population Biology*, 95:1–12, 2014.

4 References

- Rao, P. N., Li, W., Vissers, L. E., Veltman, J. A., and Ophoff, R. A. Recurrent inversion events at 17q21.31 microdeletion locus are linked to the MAPT H2 haplotype. *Cytogenetic and Genome Research*, 129(4):275–279, 2010.
- Rasmussen, M. D., Hubisz, M. J., Gronau, I., and Siepel, A. Genome-wide inference of ancestral recombination graphs. *PLoS Genetics*, 10(5), 2014.
- Ronen, R., Udpa, N., Halperin, E., and Bafna, V. Learning natural selection from the site frequency spectrum. *Genetics*, 195(1):181–193, 2013.
- Sabeti, P. C. Positive natural selection in the human lineage. *Science*, 312(5780):1614–1620, 2006.
- Sabeti, P. C., Reich, D. E., Higgins, J. M., Levine, H. Z. P., Richter, D. J., Schaffner, S. F., Gabriel, S. B., Platko, J. V., Patterson, N. J., McDonald, G., Ackerman, H. C., Campbell, S. J., Altshuler, D., Cooper, R., Kwiatkowski, D., Ward, R., and Lander, E. S. Detecting recent positive selection in the human genomes from haplotype structure. *Nature*, 419(6909):832–7, 2002.
- Sabeti, P. C., Varilly, P., Fry, B., Lohmueller, J., Hostetter, E. B., Cotsapas, C., Xie, X., Byrne, E. H., McCarroll, S. A., Gaudet, R., Schaffner, S. F., Lander, E. S., and The international HapMap Consortium. Genome-wide detection and characterization of positive selection in human populations. *Nature*, 449(7164):913–8, 2007.
- Sherry, S. T. dbSNP: the NCBI database of genetic variation. *Nucleic Acids Research*, 29(1):308–311, 2001.
- Simonsen, K. L., Churchill, G. A., and Aquadro, C. F. Properties of statistical tests of neutrality for DNA polymorphism data. *Genetics*, 141(1):413–29, 1995.
- Soldevila, M., Andrés, A. M., Ramírez-Soriano, A., Marquès-Bonet, T., Calafell, F., Navarro, A., and Bertranpetit, J. The prion protein gene in humans revisited: Lessons from a worldwide resequencing study. *Genome Research*, 16:231–239, 2006.
- Stajich, J. E. and Hahn, M. W. Disentangling the effects of demography and selection in human history. *Molecular Biology and Evolution*, 22(1):63–73, 2005.
- Stefansson, H., Helgason, A., Thorleifsson, G., Steinthorsdottir, V., Masson, G., Barnard, J., Baker, A., Jonasdottir, A., Ingason, A., Gudnadottir, V. G., Desnica, N., Hicks, A., Gylfason, A., Gudbjartsson, D. F., Jonsdottir, G. M., Sainz, J., Agnarsson, K., Birgisdottir, B., Ghosh, S., Olafsdottir, A., Cazier, J.-B., Kristjansson, K., Frigge, M. L., Thorgeirsson, T. E., Gulcher, J. R., Kong, A., and Stefansson, K. A common inversion under selection in Europeans. *Nature genetics*, 37(2):129–37, 2005.
- Steinberg, K. M., Antonacci, F., Sudmant, P. H., Kidd, J. M., Campbell, C. D., Vives, L., Malig, M., Scheinfeldt, L., Beggs, W., Ibrahim, M., Lema, G., Nyambo, T. B., Omar, S. A., Bodo, J.-M., Froment, A., Donnelly, M. P., Kidd, K. K., Tishkoff, S. A., and Eichler, E. E. Structural diversity and African origin of the 17q21.31 inversion polymorphism. *Nature Genetics*, 44(8):872–80, 2012.
- Stephan, W. Genetic hitchhiking versus background selection: the controversy and its implications. *Philosophical Transactions of the Royal Society B: Biological Sciences*, 365(1544):1245–53, 2010.
- Stephan, W. Signatures of positive selection: From selective sweeps at individual loci to subtle allele frequency changes in polygenic adaptation. *Molecular Ecology*, 25(1):79–88, 2016.
- Stoneking, M. *An introduction to molecular anthropology*. John Wiley & Sons, 2017.
- Sudmant, P. H., Mallick, S., Nelson, B. J., Hormozdiari, F., Krumm, N., Huddleston, J., Coe, B. P., Baker, C., Nordensfeldt, 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., 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., Vilems, R., Starikovskaya, E. B., Ayodo, G., Beall, C. M., Di Rienzo, A., 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. Global diversity, population stratification, and selection of human copy-number variation. *Science*, 349(6253), 2015a.

- Sudmant, P. H., Rausch, T., Gardner, E. J., Handsaker, R. E., Abyzov, A., Huddleston, J., Zhang, Y., Ye, K., Jun, G., Hsi-Yang Fritz, M., Konkil, M. K., Malhotra, A., Stütz, A. M., Shi, X., Paolo Casale, F., 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., Jasmine Mu, X., 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. T., 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, E., 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 Korb, J. O. An integrated map of structural variation in 2,504 human genomes. *Nature*, 526(7571):75–81, 2015b.
- Tajima, F. Evolutionary relationship of DNA sequences in finite populations. *Genetics*, 105(2):437–60, 1983.
- Tajima, F. The effect of change in population size on DNA polymorphism. *Genetics*, 123(3):597–601, 1989.
- Tajima, F. Infinite-allele model and infinite-site model in population genetics. *Journal of Genetics*, 75(1):27–31, 1996.
- Tang, K., Thornton, K. R., and Stoneking, M. A new approach for using genome scans to detect recent positive selection in the human genome. *PLoS Biology*, 5(7):1587–1602, 2007.
- Tenesa, A., Navarro, P., Hayes, B. J., Duffy, D. L., Clarke, G. M., Goddard, M. E., and Visscher, P. M. Recent human effective population size estimated from linkage disequilibrium. *Genome Research*, 17:520–526, 2007.
- Těšický, M. and Vinkler, M. Trans-species polymorphism in immune genes: general pattern or MHC-restricted phenomenon? *Journal of Immunology Research*, 2015:1–10, 2015.
- The 1000 Genomes Project Consortium. A map of human genome variation from population-scale sequencing. *Nature*, 467(7319):1061–1073, 2010.
- The 1000 Genomes Project Consortium. An integrated map of genetic variation from 1,092 human genomes. *Nature*, 491(7422):56–65, 2012.
- The 1000 Genomes Project Consortium. A global reference for human genetic variation. *Nature*, 526(7571):68–74, 2015.
- The Chimpanzee Sequencing and Analysis Consortium. Initial sequence of the chimpanzee genome and comparison with the human genome. *Nature*, 437(7055):69–87, 2005.
- The international HapMap Consortium. A haplotype map of the human genome. *Nature*, 437(7063):1299–1320, 2005.
- The international HapMap Consortium. A second generation human haplotype map of over 3.1 million SNPs. *Nature*, 449(7164):851–862, 2007.
- The international HapMap Consortium. Integrating common and rare genetic variation in diverse human populations. *Nature*, 467(7311):52–8, 2010.
- Thompson, E. E., Lovstad, J., Venkat, A., Susan, W., Moyse, J., Ross, S., Gamble, K., Sella, G., Ober, C., and Przeworski, M. Correction for Segurel et al., The ABO blood group is a trans-species polymorphism in primates. *Proceedings of the National Academy of Sciences*, 110(16):6607–6607, 2013.
- Thorisson, G. A. and Stein, L. D. The SNP consortium website: Past, present and future. *Nucleic Acids Research*, 31(1):124–127, 2003.
- Thornton, K. R., Jensen, J. D., Becquet, C., and Andolfatto, P. Progress and prospects in mapping recent selection in the genome. *Heredity*, 98(6):340–8, 2007.
- 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 Eichler, E. E. Fine-scale structural variation of the human genome. *Nature Genetics*, 37(7):727–32, 2005.

4 References

- Vicente-Salvador, D., Puig, M., Gayà-Vidal, M., Pacheco, S., Giner-Delgado, C., Noguera, I., Izquierdo, D., Martínez-Fundichely, A., Ruiz-Herrera, A., Estivill, X., Aguado, C., Lucas-Lledó, J. I., and Cáceres, M. Detailed analysis of inversions predicted between two human genomes: Errors, real polymorphisms, and their origin and population distribution. *Human Molecular Genetics*, 26(3):567–581, 2017.
- Vitti, J. J., Grossman, S. R., and Sabeti, P. C. Detecting natural selection in genomic data. *Annual Review of Genetics*, 47:97–120, 2013.
- Voight, B. F., Kudaravalli, S., Wen, X., and Pritchard, J. K. A map of recent positive selection in the human genome. *PLoS Biology*, 2006.
- Wakeley, J. *Coalescent theory: an introduction*. W. H. Freeman, 2008.
- Wall, J. D. Recombination and the power of statistical tests of neutrality. *Genetical Research*, 74:65–79, 1999.
- Wang, E. T., Kodama, G., Baldi, P., and Moyzis, R. K. Global landscape of recent inferred Darwinian selection for *Homo sapiens*. *Proceedings of the National Academy of Sciences*, 103(1):135–40, 2006.
- Watterson, G. A. On the number of segregating sites in genetical models without recombination. *Theoretical Population Biology*, 7:256–276, 1975.
- Williamson, S. H., Hubisz, M. J., Clark, A. G., Payseur, B. A., Bustamante, C. D., and Nielsen, R. Localizing recent adaptive evolution in the human genome. *PLoS Genetics*, 2007.
- Wiuf, C., Zhao, K., Innan, H., and Nordborg, M. The probability and chromosomal extent of trans-specific polymorphism. *Genetics*, 168(4):2363–72, 2004.
- Yi, X., Liang, Y., Huerta-Sánchez, E., Jin, X., Cuo, Z. X. P., Pool, J. E., Xu, X., Jiang, H., Vinckenbosch, N., Korneliussen, T. S., Zheng, H., Liu, T., He, W., Li, K., Luo, R., Nie, X., Wu, H., Zhao, M., Cao, H., Zou, J., Shan, Y., Li, S., Yang, Q., Asan, Ni, P., Tian, G., Xu, J., Liu, X., Jiang, T., Wu, R., Zhou, G., Tang, M., Qin, J., Wang, T., Feng, S., Li, G., Huasang, Luosang, J., Wang, W., Chen, F., Wang, Y., Zheng, X., Li, Z., Bianba, Z., Yang, G., Wang, X., Tang, S., Gao, G., Chen, Y., Luo, Z., Gusang, L., Cao, Z., Zhang, Q., Ouyang, W., Ren, X., Liang, H., Huang, Y., Li, J., Bolund, L., Kristiansen, K., Li, Y., Zhang, Y., Zhang, X., Li, R., Yang, H., Nielsen, R., and Wang, J. Sequencing of 50 human exomes reveals adaptation to high altitude. *Science*, 329(5987):75–78, 2010.
- Zeng, K., Fu, Y.-X., Shi, S., and Wu, C.-I. Statistical tests for detecting positive selection by utilizing high-frequency variants. *Genetics*, 174(3):1431–9, 2006.
- Zerbino, D. R., Achuthan, P., Akanni, W., Amode, M. R., Barrell, D., Bhai, J., Billis, K., Cummins, C., Gall, A., Girón, C. G., Gil, L., Gordon, L., Haggerty, L., Haskell, E., Hourlier, T., Izuogu, O. G., Janacek, S. H., Juettemann, T., To, J. K., Laird, M. R., Lavidas, I., Liu, Z., Loveland, J. E., Maurel, T., McLaren, W., Moore, B., Mudge, J., Murphy, D. N., Newman, V., Nuhn, M., Ogeh, D., Ong, C. K., Parker, A., Patricio, M., Riat, H. S., Schuilenburg, H., Sheppard, D., Sparrow, H., Taylor, K., Thormann, A., Vullo, A., Walts, B., Zadissa, A., Frankish, A., Hunt, S. E., Kostadima, M., Langridge, N., Martin, F. J., Muffato, M., Perry, E., Ruffier, M., Staines, D. M., Trevanion, S. J., Aken, B. L., Cunningham, F., Yates, A., and Flicek, P. Ensembl 2018. *Nucleic Acids Research*, 46(D1):D754–D761, 2018.
- Živković, D. and Wiehe, T. Second-order moments of segregating sites under variable population size. *Genetics*, 180(1):341–57, 2008.

5 Eigene Beteiligung an den Publikationen

Erklärung zur eigenen Beteiligung an den Publikationen bzw. Manuskripten:

1. Frau M. Rafajlović hat die zweiten Momente des Frequenz-Spektrums hergeleitet (Appendix B). Die Schätzung der Demographien (Abschnitte 2.2, 2.3, 3.1, 3.3) wurde von uns gemeinsam durchgeführt. Die Anpassung der Tests (Abschnitt 2.1 und Appendix A) sowie die genomischen Scans (Abschnitte 2.4, 3.4 und 3.5) sind von mir vorgenommen worden.
2. Die Formeln zum stetigen Spektrum („population SFS“) sind von Herrn L. Ferretti hergeleitet worden, die Formeln zum diskreten Spektrum („sample SFS“) von mir.
3. Die Formeln der Abschnitte 2.5 und 2.8 sowie ein Teil des Abschnitts 2.2 sind der Beitrag von Herrn L. Ferretti. Einleitung und Diskussion wurden gemeinsam verfasst. Alle übrigen Abschnitte und Berechnungen stammen von mir.
4. Der Algorithmus zur Indizierung der Haplotypen ist von mir ausgedacht und implementiert worden und ich habe eine erste Restrukturierung des Codes aus der Vorgänger-Version vorgenommen. Der Artikel wurde etwa zu gleichen Teilen von den drei Autoren erstellt.

Die Initiative und Konzeption der Artikel 1 und 3 ging von mir aus.

Zu Artikel 1 gehört ein Programm `ntx`. Der Quelltext in der Programmiersprache C++ ist frei verfügbar unter <http://sourceforge.net>. Die Funktionen für die Neutralitäts-Tests stammen ursprünglich von Guillaume Achaz (Paris), wurden von mir aber umfassend restrukturiert und erweitert.

Zu den Artikeln 2 und 3 gehört ein Bündel von Programmen, ebenfalls in C++ geschrieben, das unter dem Namen `coat1i` bei Sourceforge herunterladbar ist. Ich bin alleiniger Autor dieses Paketes.

6 Erklärung

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

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

L. Ferretti, A. Klassmann, A. M. Ferrer, E. Raineri, S. E. Ramos-Onsins, T. Wiehe and G. Achaz. The neutral frequency spectrum of linked sites. *Theoretical Population Biology*, 2018 (in press).

A. Klassmann and L. Ferretti. The third moments of the site frequency spectrum. *Theoretical Population Biology*, **120**:16-28, 2018.

M. Gautier, A. Klassmann and R. Vitalis. REHH 2.0: a reimplementation of the R package REHH to detect positive selection from haplotype structure. *Molecular Ecology Resources*, **17**:78-90, 2017.