Testing for population decline using maximal linkage disequilibrium blocks

Elise Kerdoncuff[1,2,*], Amaury Lambert[2,3] and Guillaume Achaz[1,2]

April 1, 2020

1: Atelier de Bioinformatique, UMR 7205 ISYEB, Sorbonne Université, CNRS, EPHE, Muséum National d'Histoire Naturelle, Paris, France

2: SMILE (Stochastic Models for the Inference of Life Evolution), UMR 7241 CIRB, Collège de France, CNRS, INSERM, PSL Research University, Paris, France

3: Laboratoire de Probabilités, Statistique et Modélisation (LPSM), UMR 8001, CNRS, Sorbonne Université, Paris, France

* corresponding author: elise.kerdoncuff@college-de-france.fr

Keywords: coalescent theory, recombination, demography, conservation biology.

Abstract

Only 6% of known species have a conservation status. Methods that assess conservation statuses are often based on individual counts and are thus too laborious to be generalized to all species. Population genomics methods that infer past variations in population size are easy to use but limited to the relatively distant past. Here we propose a population genomics approach that tests for recent population decline and may be used to assess species conservation statuses. More specifically, we study Maximal Recombination Free (MRF) blocks, that are segments of a sequence alignment inherited from a common ancestor without recombination. MRF blocks are relatively longer in small than in large populations. We use the distribution of MRF block lengths rescaled by their mean to test for recent population decline. However, because MRF blocks are difficult to detect, we also consider Maximal Linkage Disequilibrium (MLD) blocks, which are runs of single nucleotide polymorphisms compatible with a single tree. We develop a new method capable of inferring a very recent decline (e.g. with a detection power of 50% for populations which size was halved to N, $0.05 \times N$ generations ago) from rescaled MLD block lengths. Our framework could serve as a basis for quantitative tools to assess conservation status in a wide range of species.

1 Introduction

The severe and rapid changes imposed by human activities upon living organisms are suspected to be a major factor leading to short-term mass extinctions (Barnosky et al., 2011).The most comprehensive list of endangered species is the Red List of the International Union for Conservation of Nature (IUCN) (Rodrigues et al., 2006). Criteria used in the list to assess the species conservation status are based on geographical range, population trends, threats to habitat and ecology. Despite being very robust and reliable, these criteria are hard to establish for many species.To quantify the ongoing crisis for a wider range of organisms, there is a crucial need to develop quantitative measures of extinction risk to efficiently monitor species in real time and at a global scale. Previous attempts were developed to estimate quantitatively extinction rates, including by two of the present authors, based on occurrence data (Régnier et al., 2015; Ceballos et al., 2017; Sánchez-Bayo and Wyckhuys, 2019) or genetic data (from museum specimen Díez-del Molino et al. (2018); van der Valk et al. (2019)). The genetic methods measure the genetic diversity at different time to estimate the population size at these times and conclude on a general trend. The limitation of these methods is the difficulty to obtain time series data.

A handful of genomes sampled in a population at a single time point can help infer the past demography of this population (Gutenkunst et al., 2009; Li and Durbin, 2011; Excoffier et al., 2013; Harris and Nielsen, 2013; Sheehan et al., 2013; Schiffels and Durbin, 2014; Lapierre et al., 2017; Ringbauer et al., 2017; Terhorst et al., 2017; Beichman et al., 2018). In standard population genetic inferences, the periods when variations of population size can be estimated are of the order of N_e generations back in time. N_e denotes the so-called effective population size (Wright, 1931). Recent methods such as MSMC (Schiffels and Durbin, 2014) can provide inferences on more recent past but hardly scale up to large datasets of complete genomes because of their computational load. With the development of next generation sequencing, complete genomes from multiple individuals of the same species are now released routinely (Gibbs et al., 2015; Alonso-Blanco et al., 2016). Actual methods can not be applied to test for recent decline of populations, the models and methods we present in this manuscript specifically target very recent past when considering small populations and are meant to be applied to datasets of arbitrary size.

Methods using whole genome sequences to infer demography use different measures of genomic polymorphism. One of these measures is the so-called Site Frequency Spectrum,

or SFS (Fu, 1995). The SFS, that is the genome wide distribution of the frequencies of polymorphic alleles in a sample of the population, is strongly distorted by the demographic history of the species (Adams and Hudson, 2004; Marth et al., 2004). SFS-based methods (e.g. Gutenkunst et al. (2009)) can handle arbitrarily large numbers of loci and genomes but disregard correlations between sites caused by genetic linkage. Using genetic linkage information may help overcoming the SFS-based methods limitations (e.g. difficulty to discriminate between different scenarios Lapierre et al. (2017) and to infer recent demography).

Recombination is the process by which two DNA sequences are intermixed to create a new sequence that combines segments of different ancestries. When two homologous regions of the genome are inherited from the same ancestor without having undergone recombination, they are said IBD: *Identical By Descent*. The probability distribution and the length of IBD regions passed through generations have been studied (Stam, 1980; Chapman and Thompson, 2003; Stefanov, 2000).

Recombination patterns are also characterized by Linkage Disequilibrium (LD). LD arises when individuals of a finite population share chunks of DNA inherited from a common ancestor (IBD blocks). Specifically, two variants located at two distinct sites are in linkage disequilibrium (LD) when their joint frequency differs from what is expected under independence. More specifically, LD is defined as the covariance $f_{A_1B_1} - f_{A_1}f_{B_1}$, where f_{A_1} is the frequency of allele 1 at locus A (Lewontin and ichi Kojima, 1960). When $f_{A_1} \times f_{B_1} = f_{A_1B_1}$, the two variants are said in complete linkage equilibrium. On average, LD decreases exponentially with genetic distance due to recombination. The pattern of LD is distorted by demography (Hill and Robertson, 1968) and thus can be used to infer the past demography of a population (Hollenbeck et al., 2016; Patin et al., 2014).

Importantly, despite the fact that breakpoints between IBD blocks are usually not observable when comparing two homologous regions, "long enough" IBD blocks can be retrieved by applying one of several recent methods to a pair of sequences (Purcell et al., 2007; Gusev et al., 2009; Browning and Browning, 2010). These methods are based on detecting long identical shared segments (Gusev et al., 2009) or shared regions that harbor multiple rare variants (Purcell et al., 2007; Browning and Browning, 2010). If two individuals share the same rare variant, they may also share the surrounding chromosomal region, particularly because rarer variants are more likely to be relatively recent. Most methods take sequencing errors into account, allowing IBD blocks to not be totally identical. The

accuracy of IBD block detection depends on the algorithm used (Browning and Browning, 2013).

Some demographic inference methods are based on the distribution of lengths of inferred pairwise IBD blocks in a population. Palamara et al. (2012) have calculated the distribution of expected lengths of pairwise IBD blocks for a given parameterized demographic model. Browning and Browning (2015) have calculated the expected time to the most recent common ancestor (TMRCA) of an IBD block as a function of its length. Then they use the empirical density of IBD block lengths to estimate the distribution of TMRCA and thus the variations of effective population size through time.

Other methods use the length of identical shared segments of chromosome within a diploid individual (Hayes et al., 2003). Two identical shared segments may be inherited from a common ancestor without recombination event (and then be IBD) or may not be IBD as there are invisible recombination events that may have occurred within it. The probability that the two haplotypes of an individual share identical alleles for a given number of adjacent positions can be predicted (Hayes et al., 2003; MacLeod et al., 2009). Tools have been developed to apply these methods to infer demographic inference from genomic data (MacLeod et al., 2013; Harris and Nielsen, 2013).

Yet another approach to infer demographic history from IBD blocks is to reconstruct the genome-wide distribution of the TMRCA between two haploid genomes. In PSMC, Li and Durbin (2011) devised a Hidden Markov Model that infers the TMRCA from the positions of heterozygous sites along a pair of sequences and then estimate a step-wise demographic pattern. MSMC, the extension of PSMC (Schiffels and Durbin, 2014), analyzes the heterozygosity pattern from multiple individuals and uses first coalescence events between any two haploid genomes of the sample. These methods are computationally intensive (as of today, MSMC cannot infer the demographic history of more than 8 individuals) and pool the diversity on windows of 100 bp, that are assumed to form a single locus with two states, heterozygous or homozygous.

Importantly, the previous methods infer stepwise changes of the "effective population size" ($N_e(t)$) that are estimated from the density of coalescence events. This motivated Mazet et al. (2015, 2016); Chikhi et al. (2018); Rodríguez et al. (2018) to propose to replace $N_e(t)$ by the more explicit *Inverse Instantaneous Coalescence Rate*. IICR only matches the instantaneous population size when the population is panmictic. It is nonetheless always possible to find a population model with constant size but spatial structure that corre-

sponds to any IICR of a size-changing population for the TMRCA of 2 sequences (Chikhi et al., 2018). For larger samples, the joint distribution of coalescence events $[T_2, T_3, \cdots]$ can be used, in theory, to disentangle structure from demography (Grusea et al., 2019).

Existing methods for demographic inference using recombination information often use the whole genome of few individuals (less than 10) or use a smaller part of the genome. These methods only consider the joint history of two individuals (e.g. the pairwise IBD length distribution or the time of the first coalescence event between any two haploid genomes) which algorithmic complexity increases drastically with the number of individuals (e.g. detection of pairwise IBD blocks is quadratic) and generates a computational load limiting in most cases the application of the methods to a larger number of individuals. On the other hand, with few individuals, demographic inferences are unable to detect recent changes of population size.

Following the idea of Tiret and Hospital (2017), we decided to study the IBD concept extended to a multilocus segment and a larger number of individuals ($n > 2$). Some studies have been conducted on the amount of genetic material shared IBD with $n > 2$, considering closely related individuals (Donnelly, 1983; Ball and Stefanov, 2005). We extend the concept at a population level while relaxing the need for *identical* sequence (without mutation), which is why we decided to define a new term. We call 'MRF blocks' homologous segments that are entirely inherited from the same ancestor without recombination; these segments may or may not harbour different alleles, because of mutations. An MRF block is a segment of an alignment of haploid genomes that share the same coalescent tree. A recombination event along the sequence cuts the genome alignment into two MRF blocks, one on each side of the recombination point. By definition there is no recombination within MRF blocks so that all variants located within an MRF block are necessarily in complete linkage disequilibrium. The reciprocal is not true, as variants in complete LD do not necessarily belong to the same MRF block. MRF blocks carry the information of any recombination event that happened among the sampled individuals. As for IBD blocks, MRF blocks are usually not observable.

Outline We have developed a new test to detect very recent population declines of endangered species. We first consider the full length distribution of MRF blocks in a sample

of haploid genomes ($n \geq 2$). Second, as MRF blocks are not directly observable from sequence alignments, we devised a simple and efficient algorithm to chop an alignment of $n \geq 4$ haploid genomes in Maximal Linkage Disequilibrium (MLD) blocks, that are segments which variants are in complete LD. From the length distributions of MRF blocks or MLD blocks, we devised a summary statistic to test whether a population has been declining in the very recent past. Our method is not limited by the number of genomes in the sample.

2 Model and Methods

In the absence of recombination, ancestral relationships between genomes can be represented in the form of a genealogical tree. Individual haploid genomes at present time are the leaves of the tree, the MRCA of these individuals is the root. The fusion of two lineages into one (a common ancestor) is named coalescence event (Kingman, 1982), hence the name of "coalescent tree". The sum of all branch lengths that separates two genomes up to their common ancestor is the time of divergence between them, usually expressed in generations. In the Wright-Fisher model with constant population size N, branch lengths measured in number of generations scale like N. In particular, if we define T as the total length of the coalescent tree, the expectation of T is proportional to N. Large populations generate coalescent trees with deep nodes, whereas small populations have shallow coalescent trees.

In the presence of recombination, two loci of an alignment have the same coalescent tree only if no recombination event happened since their MRCA. We name MRF block, a maximal interval along the alignment of sites sharing the same coalescent tree. MRF blocks are consequently separated by recombination points, corresponding to recombination events. It is standard to assume that conditional on the total length T of the coalescent tree of a site, the length L of its MRF block is exponentially distributed with rate ρT, where ρ is the recombination rate (expressed in a arbitrary unit proportional to Morgan). Then for a fixed ρ, T and L are negatively correlated: recombination is more likely to occur in deep trees, which thus are carried by shorter blocks. As mentioned above, as T is proportional to N, MRF blocks are also shorter in larger populations. More accurately, because the law of T/N does not depend on N, neither does the law of NL (for $n = 2$ it alludes to results of Carmi et al. (2014)). In other words, if population 1 has size N_1

and population 2 has size N_2, the distribution of MRF block lengths in population 2 can be deduced from that in population 1 by a scaling factor N_1/N_2, both populations having identical demography otherwise. For example if $N_2 = 2N_1$, the MRF blocks in population 2 are twice smaller than those of population 1.

Note that for a given N the lengths $(L_1, L_2, \ldots)$ of successive adjacent blocks have the same distribution, but they are not independent, because the coalescent trees of adjacent MRF blocks are not. The dependencies between these trees is encoded in the so-called Ancestral Recombination Graph (ARG) (Griffiths and Marjoram, 1997). Because these dependencies have a complex structure (Wiuf and Hein, 1999), a popular way of approximating them is the Sequentially Markovian Coalescent (SMC) (McVean and Cardin, 2005; Marjoram and Wall, 2006). This approximation neglects coalescences between lineages with no overlapping ancestral material and assumes Markovian dependencies of coalescent trees along the sequence: the genealogy of an MRF block only depends on the genealogy of the adjacent ones.

Although genealogies of different MRF blocks are not independent, they are asymptotically independent as the distance between them increases.

Throughout this article, we use msprime to generate MRF blocks directly from the ARG (Kelleher et al., 2016) but very similar results were obtained with a local SMC implementation. We assume constant recombination and mutation rate along the genome. We simulated the alignment of $n = 10$ haploid genomes at present time.

Demographic scenario. We consider a single change of population size (Fig 1a). Here N_t represents the population size at time t, $t = 0$ is the present time and positive values represent the past. We denote by κ the ratio of the two sizes: $\kappa = N_\infty/N_0$, and by τ the time at which the population size changes in coalescent units of N_0 generations. If $\kappa = 1$, $N_\infty = N_0$: there is no change. If $\kappa = 10$, $N_\infty = 10N_0$: the population size has been divided by 10, τN_0 generations in the past.

3 MRF blocks

3.1 Distribution of block lengths

Impact of population decline on tree length. For declining populations ($\kappa > 1$), the coalescent trees have two distinct time scales: a first one for the shallow part of the tree ($t < \tau$), expressed in N_0 generations, and a second one for the deep part of the tree ($t > \tau$) that is expressed in κN_0 generations. When the declining population tree is compared to a standard coalescent tree (constant population size), it has shorter external branches if the reference time scale is expressed in κN_0 generations *or* longer internal ones if the reference time scale is expressed in N_0 generations. When it is compared to a reference tree with population size chosen so as to have the same T_{MRCA}, its external branches are too short *and* its internal branches are too long. Similarly, the distribution of the total length T of the tree is overdispersed when compared to the length of the standard coalescent tree with the same mean.

Impact of population decline on lengths of MRF blocks. For a declining population, the distribution of the length L of MRF blocks will depend not only on ρ and N_0 but also on κ and τ. As the tree relative branch lengths are distorted and the distribution of T is overdispersed, so is the distribution of L. In a declining population, the distribution of L can be seen as a mixture of the two distributions that correspond to the two population sizes, N_0 and κN_0. The strength of the decline (κ) tunes the difference between the distributions; the date of decline (τ) tunes in what proportion the two distributions are mixed. When $\tau \to 0$ (practically, $\tau < 10^{-4}$ times N_0 generations for a sample size $n \in [10, 100]$), the distribution of L is indistinguishable from that of block lengths in a population with constant size equal to κN_0. At the opposite, for $\tau \to \infty$ (practically, $\tau > 10$ times N_0 generations for a sample size $n \in [10, 100]$), the distribution of L is indistinguishable from that of block lengths in a population with constant size equal to N_0. As a result for $\tau \in [10^{-4}, 10]$, the distribution of L has an excess of MRF blocks smaller than the N_0 reference and an excess of MRF blocks longer than the N_∞ reference (Fig 1b). The small blocks correspond to the trees which total length T is mostly driven by the distant N_∞ time scale and the long ones to the trees which total length T is mostly driven by the recent N_0 time scale.

As mentioned in the previous section, in a population with constant size N, the distribution of L, briefly denoted L_N, scales like $1/N$, in the sense that the distribution of $\tilde{L} := NL_N$ does not depend on N. In particular, the distribution of $L' := L/E[L]$ does not

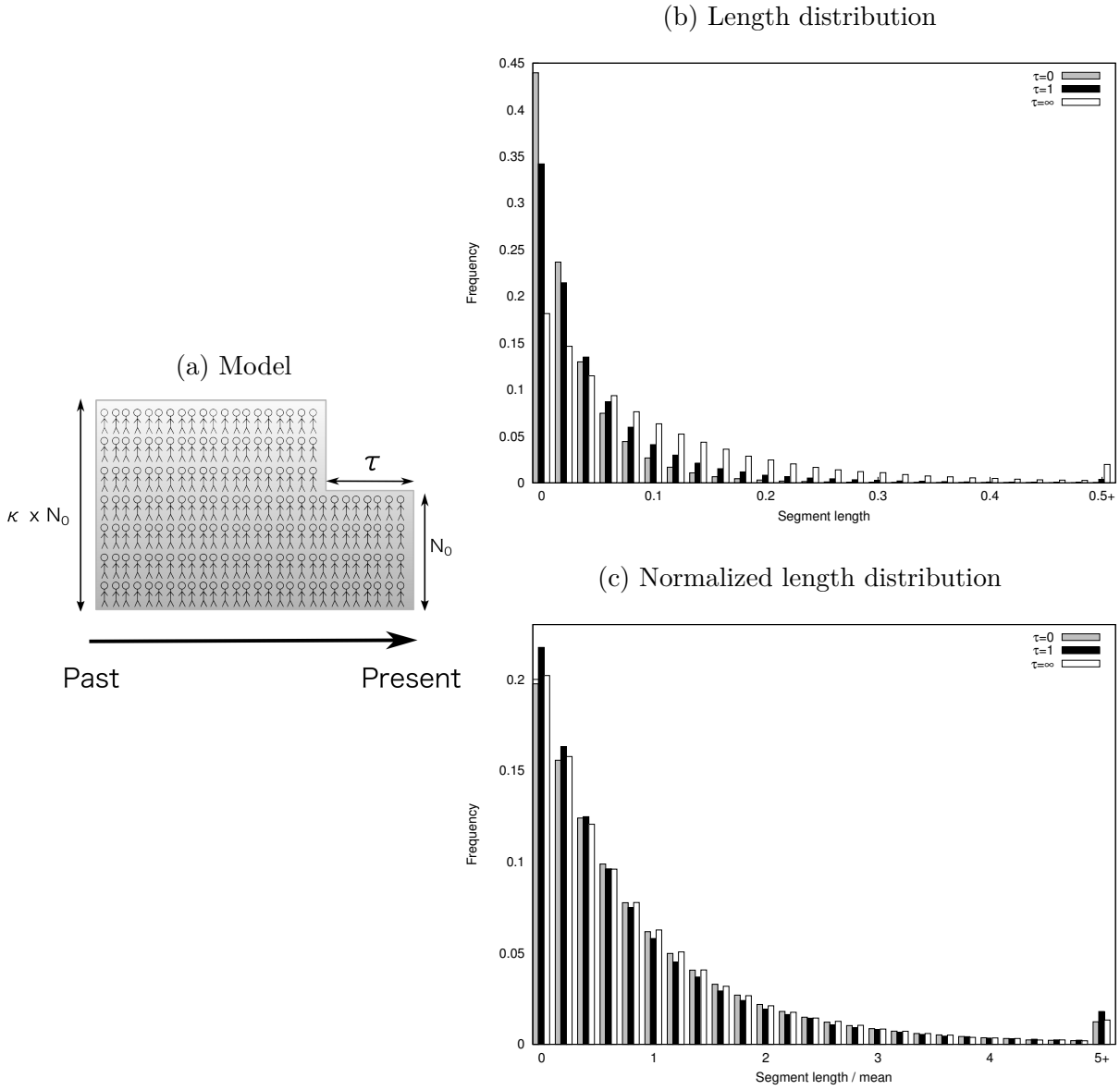


Figure 1: Impact of the demography on the distribution of MRF block lengths. (a) The demography considered here is a sudden size change tuned by 3 parameters : N_0, the actual population size, τ the date of decline (in backward time) and κ the strength of decline. Time is expressed in N_0 generations. (b) Distribution of L for $\rho = 1$, $\kappa = 3$ with $\tau = \{0, 1, \infty\}$. When $\tau = 0$ (grey) or $\tau = \infty$ (white), population size is constant. (c) Distribution of $L' = L/E[L]$ under the same values of $\rho = 1$, κ and τ. In case of a decline, the distribution is overdispersed, with an excess of both short and long normalized MRF blocks.

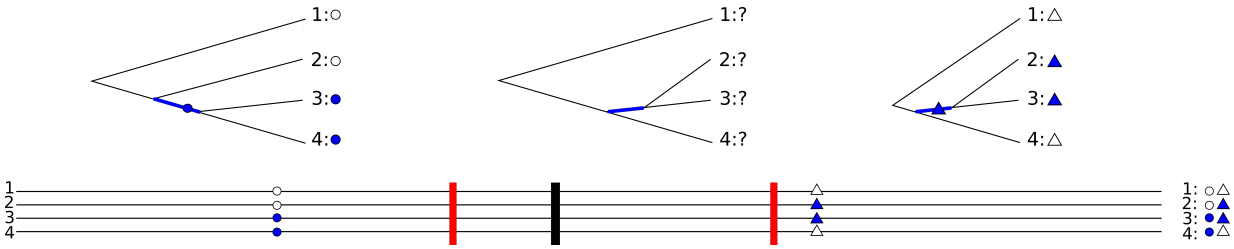


Figure 2: **Detection of recombination in three MRF blocks**. The four lines represent four haploid genomes, circles and triangles are mutation events, red lines are the true recombination events delimiting MRF blocks and above each MRF block is represented its true tree. Mutation events occur on certain lineages as represented on the trees. The first recombination event generates an incompatibility between the blue branches of the first two MRF blocks, but as no mutation occurs on the second MRF block, this recombination event cannot be detected. The second recombination event does not change the topology of the tree and thus this second event cannot be detected either. However, the first and the third MRF blocks carry mutations that are not compatible; thus a minimum of one recombination event can be inferred between the two mutations, as indicated by a vertical thick black line arbitrarily placed in the middle.

depend on N in a population with constant size and follows the law of $\tilde{L}/E[\tilde{L}]$. However, the distribution of L' is distorted when there is a size change. For a declining population, the distribution of L' is overdispersed, it has an excess of small blocks (*i.e.* less than 0.2) and an excess of long blocks (*i.e.* more than 5), as can be seen on Fig 1c.

Note that we always have $E[L'] = 1$, but here $E[L]$ has

$$\frac{1}{N_\infty} E[\tilde{L}] = E[L_{N_\infty}] < E[L] < E[L_{N_0}] = \frac{1}{N_0} E[\tilde{L}].$$

As the block distribution of L_N is a mixture of the one of L_{N_0} and L_{N_∞}, $E[L]$ is bounded by $E[L_{N_\infty}]$ and $E[L_{N_0}]$ that depend on the population size.

4 MLD blocks

4.1 Definition

All recombination events are not directly visible in a genome alignment. First, adjacent MRF blocks may have coalescent trees sharing the same topology and the same branch lengths, so that mutations occurring on either tree show exactly the same pattern on either block. Second, adjacent MRF blocks may have coalescent trees sharing the same

topology but not the same branch lengths, so that mutations occurring on either tree display the same bipartitions (compare the second and third tree in Figure 2). Third, even if two adjacent MRF blocks have trees with different toplogy, it is possible that branches distinguishing these topologies do not carry mutations (see the second block in Figure 2).

Importantly, recombination events that happen between the two oldest lineages do not impact the topology of the tree, so are never detectable because they do not impact the possible bipartitions.

A possibility used in the literature to detect breakpoints between MRF blocks is to detect the changes in the density of polymorphic sites along the sequence due to the change of coalescent tree (like in PSMC, Li and Durbin (2011)).

Here we used instead the incompatibilities between bipartitions displayed by polymorphic sites to place the minimal number of recombination events on the alignment. Two bipartitions are said incompatible when they are not compatible with a common tree.

In what follows, we will assume that the mutation rate μ is constant through time and along the genome.

4.1.1 The four-gamete test

From now on, we assume that each site can be hit at most once by a mutation, so that a polymorphic site is always bi-allelic, an assumption known as the "infinitely-many sites model". The four-gamete test (Hudson and Kaplan, 1985) serves to detect incompatibilities between bipartitions displayed by two polymorphic sites. For any two biallelic sites (A/a and B/b) there are at most four gamete haplotypes in the population (A-B A-b a-B and a-b). Under the infinitely-many sites model, the four possible haplotypes cannot be observed in a sample if the two sites share the same genealogy. Then if the four possible haplotypes are observed in the sample, a recombination event must have occurred between them – but not necessarily the other way round. This property can be used to compute a lower bound for the number of recombination events in a genome alignment (Hudson and Kaplan, 1985) or even to estimate the recombination rate (Hey and Wakeley, 1997). We used it to compute and place the minimal number of breakpoints in a genome alignment.

Two polymorphic sites are said *incompatible* if the four possible haplotypes are present in the sample. When a sequence of adjacent polymorphic sites contains no pairwise incompatibility, we speak of a sequence of compatible sites. Note that a sequence of compatible sites are in complete linkage disequilibrium. We thus define an MLD block, for *Maximal*

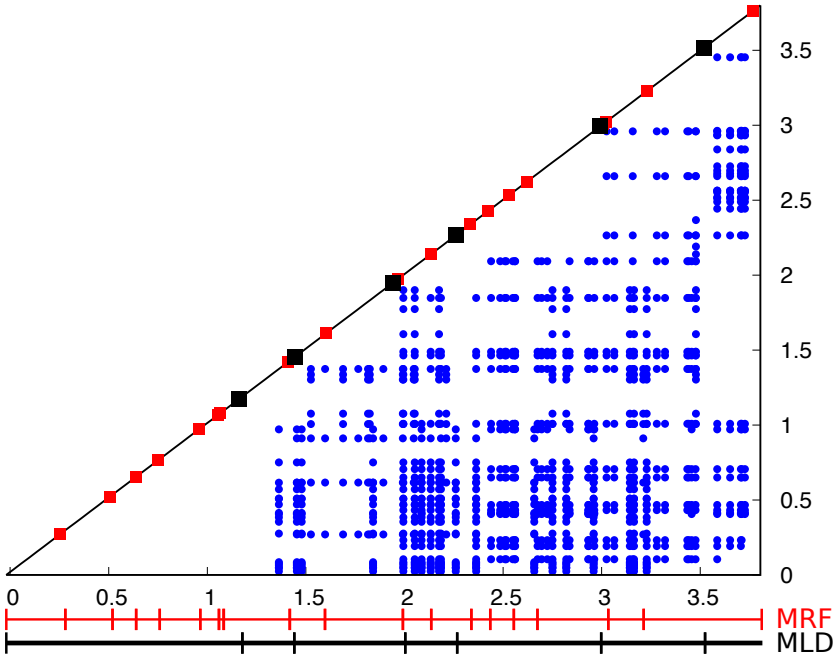


Figure 3: The incompatibility matrix and the chopping algorithm. X and Y-axis are positions on the genome alignment. Blue dots represent a pair (x,y) of incompatible sites. The red squares are the true positions of recombination events (MRF breakpoints) and the black squares are MLD breakpoints inferred by the chopping algorithm.

Linkage Disequilibrium block, as any maximal sequence of compatible sites.

We now explain how to extend this notion originally designed for haploid genomes (or phased diploid genomes) to an unphased diploid genome, that is, a diploid sequence lacking the linkage information. For an unphased diploid genome, the two original haplotypes can be determined if the diploid genome is homozygous at at least one the two sites:

- When the genome is homozygous at both loci (A/A-B/B), both haplotypes must be A-B.
- When the genome is homozygous at one locus and heterozygous at the second one (A/A-B/b), the haplotypes must be A-B and A-b.

The four-gamete test can then be extended to a sample of unphased diploid genomes by saying that two sites are incompatible in this sample if they are incompatible in the subsample of haplotypes that have been inferred thanks to the previous remark. When the haplotype is ambiguous, the sites are considered compatible and do not bring more information about a recombination event.

4.1.2 The chopping algorithm

We used the four-gamete test to detect incompatibilities in the genome alignment and to chop it into MLD blocks (Fig 3). To avoid computing the full matrix of pairwise incompatibilities between all polymorphic sites of the genome, we only compute the incompatibilities for sequences of P adjacent polymorphic sites (by default $P = 150$). Each pair of incompatible sites (i, j) defines an interval that contains at least one MLD breakpoint. To place the MLD breakpoint, we seek the shortest interval that is sufficient to explain the incompatibilities.

Algorithm: We retrieve all intervals and sort them in increasing order of site positions along the genome (first by i the first site position and when equal, by j the second site position). As we scan two times the list of intervals, the algorithm complexity is linear with the number of polymorphic sites:

1. **Discarding and shortening.** For this step, we scan the list in reverse order, from the last (N) to the first interval. (The algorithm can be done in the forward order, the distribution will be slightly different but it will not affect the study.) Each interval containing another entire interval is discarded: for two intervals (i_N, j_N) and (i_{N-1}, j_{N-1}), if $i_N \leq i_{N-1} \leq j_{N-1} \leq j_N$, then (i_N, j_N) is discarded. When two intervals overlap, they are replaced by their intersection (the two original ones are discarded): for the two intervals (i_N, j_N) and (i_{N-1}, j_{N-1}), if $i_{N-1} \leq i_N \leq j_{N-1} \leq j_N$, both are replaced by a new interval (i_N, j_{N-1}), that is then compared to (i_{N-2}, j_{N-2})...

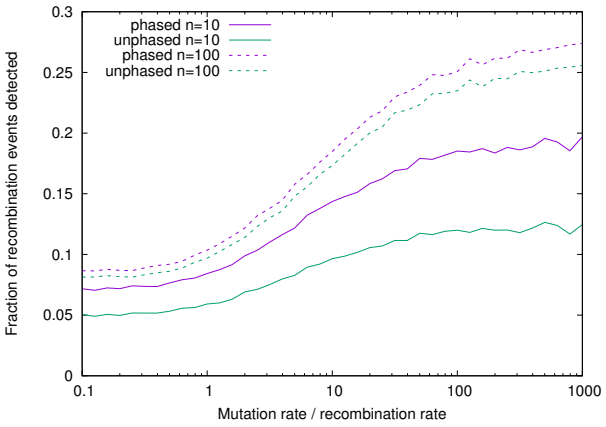
2. **Positioning.** From the final list of disjoint intervals, we place an MLD breakpoint at the middle of each interval.

MLD breakpoints partition the genome alignment into MLD blocks.

4.2 Length distribution

The distribution of the length L_c of a typical MLD block does not only depend on the distribution of L (MRF block length) but also on the fraction p of recombination events that are detected. This fraction increases with the ratio μ/ρ, as illustrated in Figure 4a. When many mutations occur in two different MRF blocks ($\mu \gg \rho$), the probability that they occur on incompatible branches of their respective coalescent trees increases and so

(a) Factors impacting the power of detection



(b) MLD block lengths

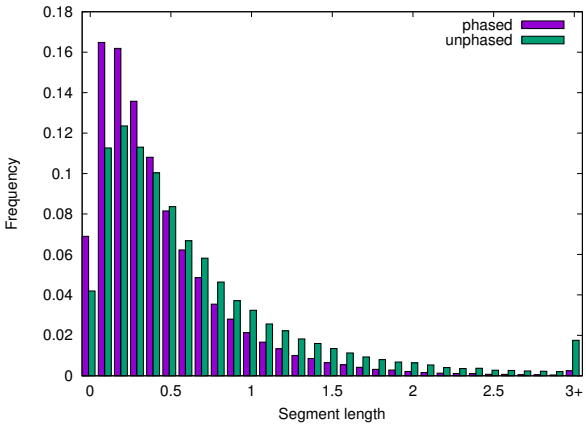


Figure 4: Detection of recombination events and its impact on MLD block length (L_c) distribution under a constant population. (a) Fraction of recombination events that are detected as a function of μ/ρ for different sample sizes ($n = 100$, dashed lines and $n = 10$ plain lines) and for phased (purple) or unphased (green) diploid genomes. (b) Distribution of MLD block lengths for phased (purple) and unphased (green) diploid genomes in a population of constant size ($\mu = 10$, $\rho = 1$, $n = 10$).

does the detection efficiency, up to a point of saturation due to cases when these MRF blocks share the same tree topology. The number of sampled individuals also impacts the efficiency of detection (Fig 4a): the larger the sample size, the higher the probability to observe incompatible mutations. The four-gamete test for unphased diploid genomes has obviously less power to detect recombination than for phased genomes (Fig 4a).

The lower the power to detect recombination, the longer the MLD blocks. In particular, phased genomes have smaller MLD blocks than unphased ones (Fig 4b). Furthermore increasing the sample size results in more detectable recombination points and thus smaller MLD blocks. In Figure 4b, the average block length, in our arbitrary unit for $n = 10$ phased haploid genomes is $\bar{L_c} = 0.497$ ($\mu = 10, \rho = 1$). Considering smaller sample size will result in larger MLD blocks (e.g. $\bar{L_c} = 1.32$ for $n = 5$). This implies that the total number of blocks can be limiting for small sample size, and that these long blocks will be harder to detect in scaffolds of partial genomes. In (very) large samples, MLD blocks are shorter: $\bar{L_c} = 0.132$ for $n = 600$ (ten times smaller than for $n = 5$) and $\bar{L_c} = 0.103$ for $n = 6,000$. On a side note, the theoretical pitfall of having too small "undetectable" blocks can always

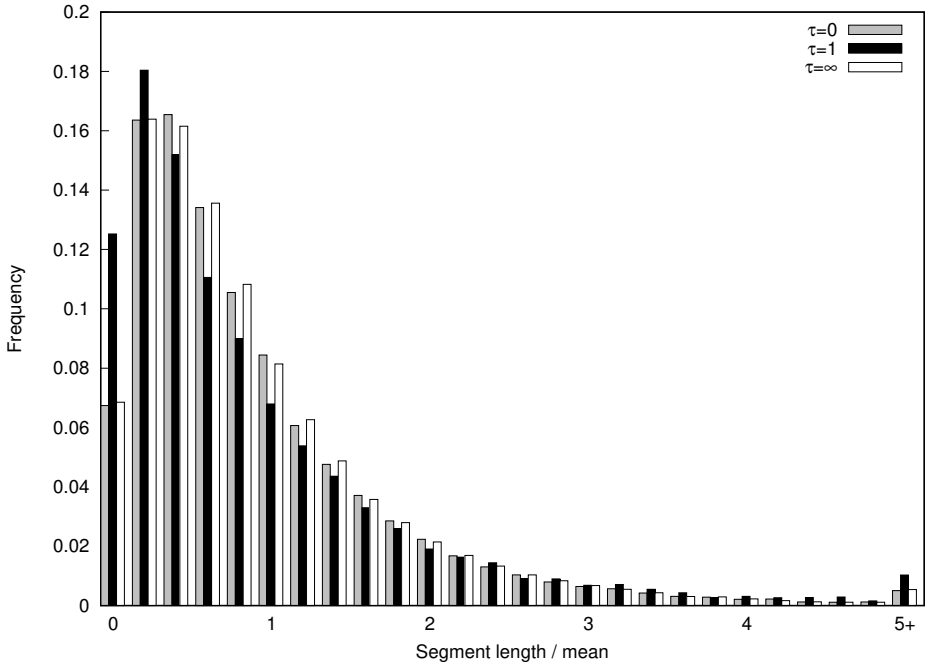


Figure 5: Distribution of L'_c for a population of constant size (white, $N = N_0 = 1$ and grey, $N = \kappa N_0$=3) and for a declining population (black for $\tau = 1$) with $\rho = 1$, $\mu = 10$ and $n = 10$.

be overcome by subsampling.

Here, we consider the block lengths normalized by the average length $L'_c = L_c/\overline{L_c}$. Similarly to the MRF blocks, the distribution of L'_c does not depend on the value of N but does depend on the demographic scenario (Fig 5). However, it still depends on our ability to detect recombination and so on the ratio μ/ρ and n the number of sampled individuals. To compare distributions, it is then important that they have the same ratio μ/ρ and the same n.

Similar to what we have observed for MRF blocks, a declining population exhibits both an excess of small blocks ($L'_c < 0.2$) and large blocks ($L'_c > 5$) (Fig 5). The shape of the distribution of L'_c (Fig 5) differs from the one for L' (Fig 1c): MLD blocks are longer than MRF blocks. Indeed, they contain a variable number of MRF blocks and below a certain size, MRF blocks are not detectable as recombination points at the edges of an MRF block can be detected only when mutations have occurred inside the block. MLD blocks are always longer than MRF blocks.

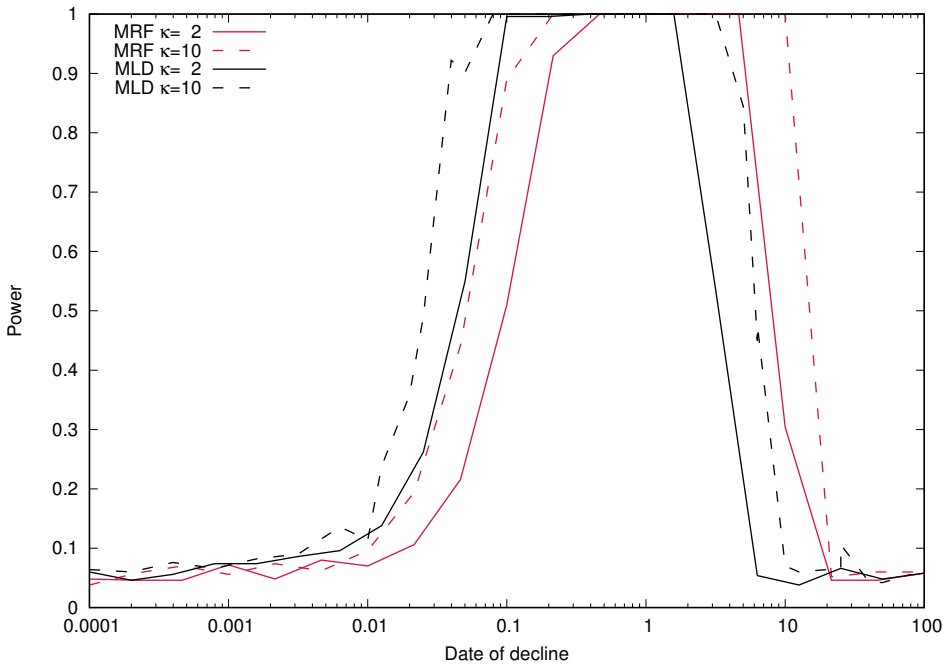


Figure 6: Power to detect population decline. The test based on MRF blocks (f) is pictured in red, whereas the one based on MLD blocks (f_c) is represented in black. We assess the power of the two tests for $\kappa = 2$ (plain line) and $\kappa = 10$ (dashed line) with $\tau \in [0.0001, 100]$.

5 Statistical tests for population decline

5.1 Test

To test for population decline, we use the excess of small and large blocks that we observe when comparing samples from a declining *vs* a constant population size. More specifically, we compute the fraction of blocks which normalized length is either smaller than 0.2 or larger than 5, both in the case of MRF blocks ($f = f_{L'<0.2} + f_{L'>5}$) and of MLD blocks ($f_c = f_{L'_c<0.2} + f_{L'_c>5}$) . To set an empirical threshold value under H_0, we simulate 10,000 genomes of 10^5 MRF blocks under a constant population size for 10 haploid genomes and compute both $f^{5\%}$ and $f_c^{5\%}$ as upper limits for one-tailed tests: $f^{5\%} = 0.214236$ and $f_c^{5\%} = 0.075824$. As the threshold is empirical, simulations need to be redone for a change in sampled size or in null model. Time needed for simulations depends on the algorithm/software used and the specific features of the model.

5.2 Power

To assess the power of this test, we simulated 1,000 replicates under population decline (H_1 with various τ and κ) and report the fraction of runs where $f^{H_1} > f^{5\%}$ for MRF blocks or $f_c^{H_1} > f_c^{5\%}$ for MLD blocks. When the power is 1, the decline was significant in all runs. When the power is 5%, the decline is not detectable, the test can not differentiate H_0 and H_1.

Without surprise, results show that the power of the test to detect population decline depends on both the decline strength (κ) and the date of decline (τ) (Fig 6). For both tests based either on MRF or on MLD blocks, the power outreaches the 5% risk only for a range of τ. The type I error of the test is 5% as expected. For both tests, the range of detection is wider when the decline is stronger (compare dashed to solid lines in Fig 6). The surprise is that the test based on MLD blocks (f_c) detects more recent declines than the test based on MRF blocks (f). Therefore, we recommend using the f_c test when searching for very recent decline even if MRF blocks are known (which is generally not the case).

6 Application to data: the case of the western lowland Gorillas

6.1 Handling the low quality of real genomes

Genomic data sets often include sequencing errors and regions that are not genotyped. Consequently, the f_c test cannot be run as is on these data sets. We present some modifications to our test to handle the poor quality of data. We show in this section that adjustments can be made to get the information from the L'_c distribution.

Simulations with lower quality

Difficulties in applying the f_c test to real data sets can stem from the low quality of DNA sequences. We replicated in the simulated genomes the two main issues, namely the interruptions of DNA tracts and the absence of genotyping for some SNPs in some individuals.

DNA tract interruptions truncate MLD blocks and make their detection difficult. The

number, size and location of these interruptions will have an effect on the detection of MLD blocks and thus will alter the L'_c distribution. To handle the effect of interruptions, we placed the interruptions at the same positions in our simulated chromosome as in the real chromosome.

As for the partial genotyping issue, we artificially lowered the genotyping quality in the simulated chromosomes. We used the empirical distribution of missing individuals (e.g. chr1 of *Gorilla gorilla*, Fig 7) to pick random positions in the simulated chromosome and erase the genotypes of some individuals.

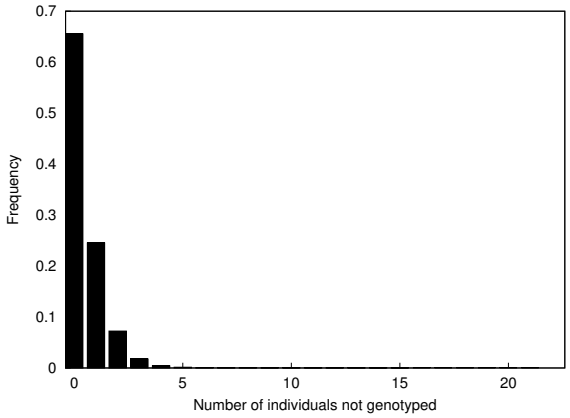


Figure 7: Distribution of the number of individuals not genotyped per SNP on chromosome 1 of the Gorillas dataset (Prado-Martinez et al., 2013).

Mutation rate and recombination rate

To cope properly with the issue of genotyping, we simulated chromosomes with the same number of mutations and the same MLD length mean as in the studied data set. We use the Watterson estimator (Watterson, 1975) for the mutation rate and fixed the recombination rate so that simulated and real chromosomes had the same average length of MLD block.

6.2 Application to Chr1 of *Gorilla gorilla gorilla*

We applied this methodology on chromosome 1 of twenty-three unrelated western lowland Gorillas (*Gorilla gorilla gorilla*) from the Great Ape Genome Project (Prado-Martinez et al., 2013). The chromosomes have 247,249,719 base pairs. The 23.1% of sites that are considered "low coverage"(Prado-Martinez et al., 2013) divide the chromosome alignment

into 6,277,293 uninterrupted stretches. The 5,388,083 interruptions due to a single site were not considered as *interruptions*. To speed up simulations, we considered stretches longer than 499 sites, as smaller stretches often carry no entire MLD block. We chopped chromosome 1 using a window of 150 polymorphic sites, into 7,082 MLD blocks with an average length of 307.897 bp.

Distribution of L'_c

The distribution of L'_c for our sample of gorilla sequences has an excess of small and long MLD blocks compared to the L'_c distribution of a constant population with the same characteristics (same number of mutations and same average length of MLD blocks) (Fig 8). The excess of small blocks is even larger than what we see in simulated declines (see above). The truncation of long MLD blocks due to the inclusion of low quality of genotyping can potentially inflate this excess.

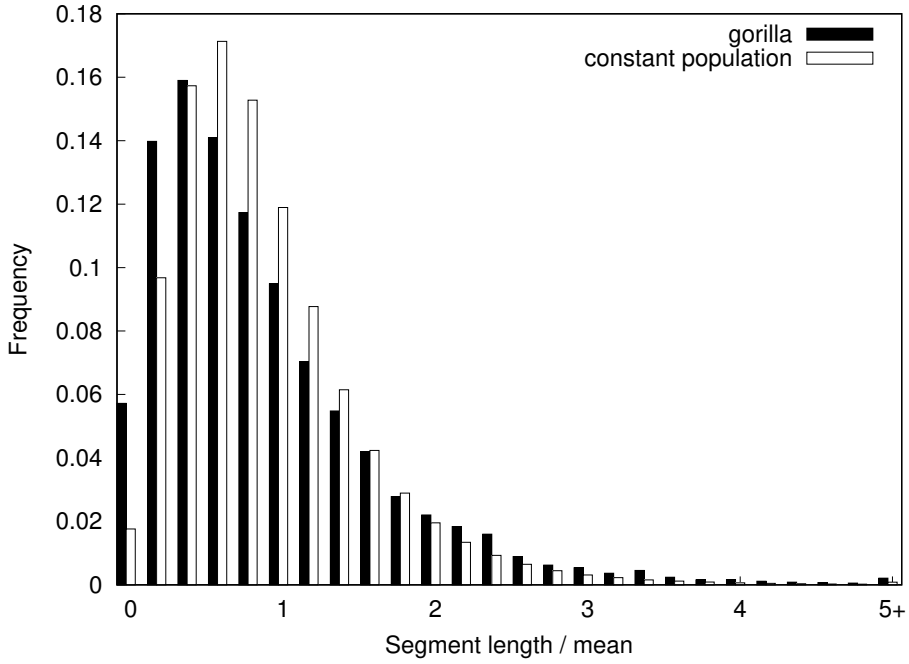


Figure 8: Distribution of L'_c for a population of constant size (white, mutation rate= 0.000375 , recombination rate= 0.012) and for the chromosome 1 of the gorillas (black)

With the low quality of genotyping and the chosen mutation and recombination rates, the threshold value $f_c^{5\%}$ is 0.041627. As we measure $f_c^{gor} = 0.0631178$ for the gorillas,

the test significantly rejects H_0. However, it is possible that other designs of similar tests (tweaking the lower or upper bounds) may be more relevant to analyze demography from low quality chromosome alignments.

However, and this may be even more important, misspecifications of the model can also make the test significant. Among all, we have chosen to explore the impact of recovery after the decline and of spatial structure.

7 Misspecification of H_1

To appreciate how the f_c test, that was specifically designed to detect population decline, is sensitive to other violations of H_0, we explored their sensitivity to a scenario of bottleneck (decline followed by recovery) and to a scenario with structure but no demography.

7.1 Bottleneck

In the bottleneck scenario, we model a population that experienced a sudden strong decline ($\kappa = 10$) at time $\tau = 1$ in the past and recovered to its original size after a duration of $x \in [0, 1]$. If $x = 0$, there is no population decline. If $x = 1$, the population has not recovered and the bottleneck scenario is identical to our original H_1. When the bottleneck lasts long enough ($x > 0.02$), it is detected by the f_c test (Fig 9b). On the contrary, when the bottleneck is too short ($x < 0.02$), the distribution of L'_c is similar to the one under H_0 (Fig 9a). This shows that even if the population has recovered, the signal of decline will be observable in the excess of short and long MLD blocks.

7.2 Island-mainland structure

Structured populations generate signals of population size change, even when the population is stationary (Mazet et al., 2015). For example if the size of the sample is $n = 2$, for any population model with spatial structure, there exists a model without structure but specifically designed variations of population size which has the same distributions of coalescence times (Mazet et al., 2016). We consider here a larger sample size ($n = 10$, as in the other scenarios). We assume that genomes are sampled from an island with population size N and the island receives migrants with individual rate m from the mainland, which has population size $10N$. The shape of the distribution of L'_c is impacted by the migration rate

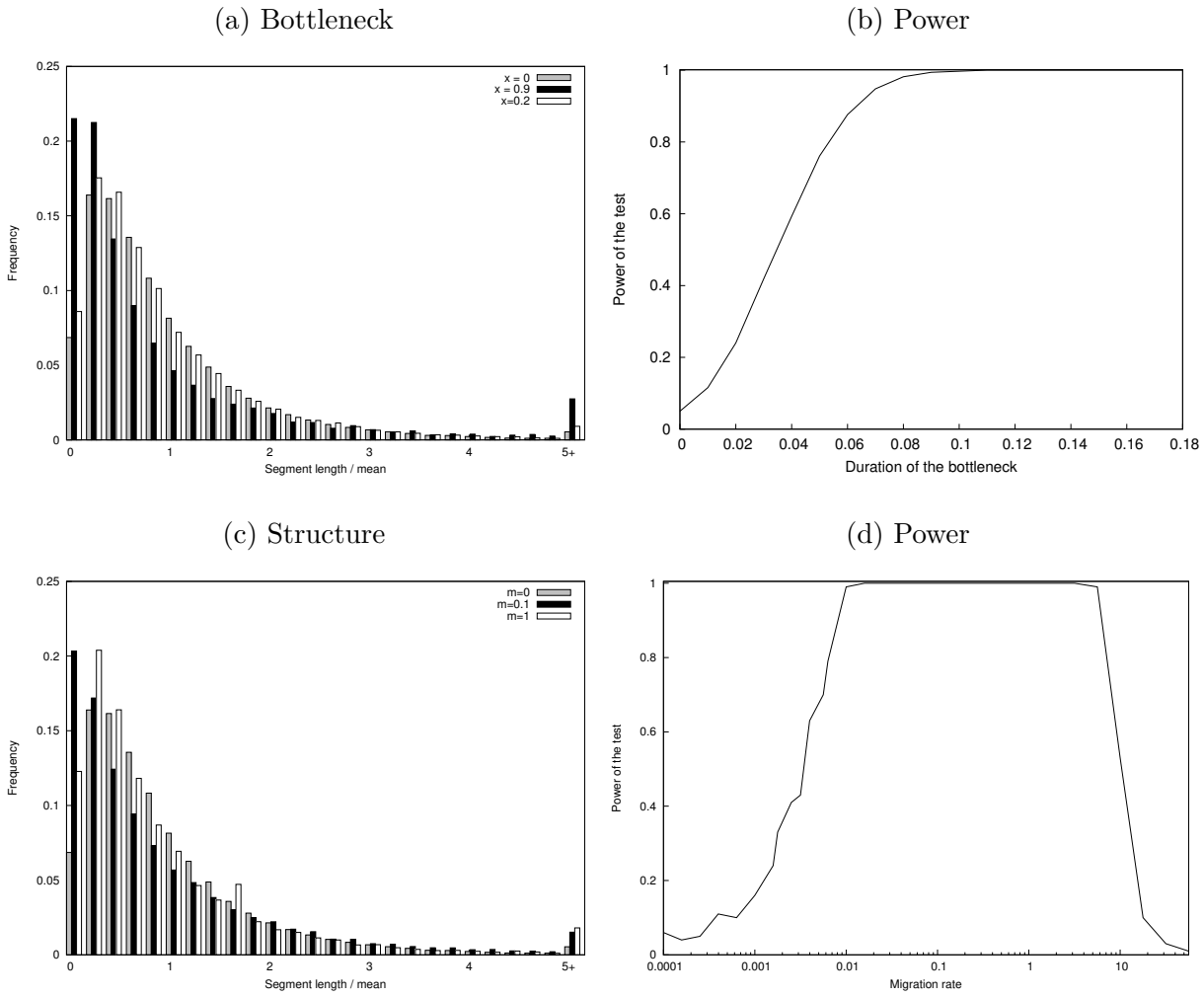


Figure 9: Distribution of L'_c and power of the f_c test under two alternative scenarios. (a) In the bottleneck scenario, the population size is constant equal to N except during the period $[1-x, 1]$ (measured in units of N generations backwards from the present) during which it equals $N/10$, with $x = \{0, 0.2, 0.9\}$ ($x = 1$ corresponds to decline without recovery). (b) Power of the f_c test on the bottleneck population as a function of the duration of the bottleneck, $x \in [0, 0.18]$. (c) In the island-mainland scenario, the population size is constant equal to N (island) and receives migrants at individual rate $m = \{0, 0.1, 1\}$ from a population of size $10N$ (mainland). (d) Power of f_c test as a function of m, the migration rate from the mainland to the island.

and the ratio of population sizes between the island and the mainland (data not shown). When the migration rate gets too small ($m < 0.001$) or too large ($m > 10$), the distribution of L'_c is the same as under H_0 (Fig 9c). For intermediate values (*i.e.* $0.001 < m < 10$), an a excess of short and long blocks will be observed (Fig 9d). However, the shape of the distribution of L'_c is visually different from the one of a declining population, that is, the excess of small blocks is higher than under a declining scenario. For a value of $m = 0.1$, the proportion of blocks between 0 and 0.1 times the mean is significantly higher than the proportion of blocks between 0.1 and 0.2 times the mean, which is not observed for the distribution of block lengths under decline. This suggests that the distribution could be used to differentiate between the effects of demography and structure.

8 Discussion

We have explored the impact of demography, more specifically recent population decline, on the pattern of recombination in a sample of n genomes, where $n \gg 2$. We have shown that the distribution of the distances between recombination breakpoints (MRF block lengths) is strongly affected by the demography. More specifically, a decline will result in an overdispersion of the distribution, that is, a relative excess of short and long blocks. As most recombination breakpoints are difficult, and sometimes impossible, to detect in a sequence alignment, we have proposed to restrict ourselves to the ones that can be detected using the four-gamete test. These detectable breakpoints delineate blocks in full linkage disequilibrium that we named MLD blocks.

Although different from the distribution of MRF blocks, the distribution of MLD block lengths is also overdispersed when the population has been declining recently. Using simple tests based on an excess of small and long blocks (f and f_c), one can detect declines for a wide range of different dates and strengths.

Surprisingly, the f_c test based on MLD blocks has more power for very recent declines ($\tau \approx 0.01$) than the f test based on MRF blocks. The past demography of the population impacts the distribution of the length L of MRF blocks but also the fraction p of MRF breakpoints that also correspond to MLD breakpoints. When recombination occurs at distant times when only $k \ll n$ ancestor lineages are present (*i.e.* the most ancient times of the tree), it rarely produces incompatibilities detectable with the four-gamete test (never when $k = 2$). For a declining population, these ancient lineages have longer branches

than the ones of a constant population scenario, so that recombination events occur more frequently in these lineages. This results in a smaller p for declining populations and thus in more numerous (ancient, small) MRF blocks per MLD block. The relative abundance of long recent MLD blocks becomes thus more important in the distribution. This effect fades away for distant declines. In summary, the effect of recent declines on the L_c distribution is the result of both a change in the L distribution and a change in the fraction p of breakpoints detected, which can explain the difference in power between the f and the f_c tests.

We also show that using the f_c statistic, the decline can still be detected even if the population has recently recovered its original size (bottleneck scenario). Finally, we showed that local sampling of a small deme with constant size also leads to rejection of H_0 for f_c but that the distribution of block lengths seems distorted in a way that can help distinguish the two scenarios. We leave this for future work.

One interesting advantage of using the f and f_c tests are their efficiency in computing time, such that they can scale up to a very large sample of long genomes. For example, the chopping of the entire human chromosome 1 (1,636,975 SNPs) for a sample of 10 unphased genomes takes 16 seconds on a laptop (with an Intel Core i7 processor running macOS High Sierra). This very short computing time is an interesting asset of this test compared to other methods based on variations (e.g. in SNP density) induced by recombination events (Li and Durbin, 2011; Palamara et al., 2012; MacLeod et al., 2013; Harris and Nielsen, 2013; Browning and Browning, 2015). The main choice that influences the computation time of the chopping algorithm is the number of sites considered for the chopping window. An increase in the chopping window size will increase the number of sites to test for incompatibility.

In the theoretical assessment of the f_c test, we have made the assumption that the recombination rate is constant along the genome and that entire genomes are aligned. Let us discuss the limits of these assumptions.

First, the recombination rate is known to vary along the genome, especially in regions of high recombination known as recombination hot-spots. It could be possible to integrate these variations via the knowledge of the recombination map. Indeed, if the recombination rate is twice higher in a given region of the genome, MRF blocks will be twice smaller, so

we can correct this distortion by multiplying all MRF block lengths by 2.

Another issue of the test based on MLD block lengths is the need of whole genome data. For normalisation of the MLD block distribution, the average length of a block is needed. If the whole distribution of MLD block is not available, it can compromise the estimate of the average length, and so can compromise the test based on the normalised distribution. The test requires genome data with good SNP quality for all the individuals.

The f_c test is a genome-wide approach that can detect population decline that started even very recently, down to orders of $\tau = 0.01N_0$, where N_0 is the current (effective) population size. This corresponds to very recent times, in particular when considering endangered populations. For example, there are approximately 600 mature mountain Gorilla individuals alive (IUCN Red List of 31 July 2018). Assuming that the current effective population size is a third of the mature individuals, $N_0 \approx 200$, the f_c test will detect decline as recent as $0.01 * 200 = 2$ generations ago. Great apes populations (Bonobos, Chimpanzees, Orangutans) have been sequenced (Prado-Martinez et al., 2013) and are actively re-sequenced (Gordon et al., 2016). The coverage used to sequence the data currently available is not high enough to apply our test. To infer MLD blocks, the sequenced DNA tracts need to be uninterrupted. Using these whole-genome data in higher quality, we will be able to confirm their decline thanks to the f_c test.

Giant pandas have a 'vulnerable' conservation status in the Red List. Recently they have seen their population increased (around 500 mature individuals, from IUCN website 2019). Applying the f_c test on some genomes of theirs (Zhao et al., 2013) sequenced in higher quality, will give some precise information on their demography. As the test is influenced by duration and strength of a bottleneck, the strength and the date of the increase in the population size impact the result of the test. Applying the f_c test to mammals with approximately known demography will be interesting to verify the method. However, the real asset of this test is its possible application to a much wider range of organisms. Whole-genome data start to become more and more common for non-model, non-vertebrate organisms like honeybee (Wallberg et al., 2014), as well as organisms with no conservation status such as mimicry butterflies (Zhang et al., 2017).

The chopping algorithm detects incompatibilities among trees along the genome. All the sites in a MLD block are compatible with one topology. We developed the algorithm

to detect a recent change in population size. However, its use is not limited to population demography. Conflicting genealogies are also present in phylogenetic inference (Maddison, 1997). This algorithm could be used to partition the genome according to compatible trees before estimating the trees.

Recombination and mutation events leave a joint imprint on genomes which depends notably on the demography of the population. Their frequency and locations carry information about the past history of this population (decline, bottleneck, structure...). Using MLD breakpoints to chop genomes gives insights into this history and may be used to gain further information on other aspects impacting the frequency of recombination events through time and along the genome (e.g. hitch-hiking due to selection).

Acknowledgements

E.K. is funded by the PhD program 'Interfaces pour le Vivant' of Sorbonne Université. G.A., A.L. and E.K. thank the *Center for Interdisciplinary Research in Biology* and the *Fondation François Sommer* for funding.

References

Adams, A. M. and Hudson, R. R. (2004). Maximum-likelihood estimation of demographic parameters using the frequency spectrum of unlinked single-nucleotide polymorphisms. *Genetics*, 168(3):1699–712.

Alonso-Blanco, C., Andrade, J., Becker, C., Bemm, F., Bergelson, J., Borgwardt, K. M., Cao, J., Chae, E., Dezwaan, T. M., Ding, W., Ecker, J. R., Exposito-Alonso, M., Farlow, A., Fitz, J., Gan, X., Grimm, D. G., Hancock, A. M., Henz, S. R., Holm, S., Horton, M., Jarsulic, M., Kerstetter, R. A., Korte, A., Korte, P., Lanz, C., Lee, C.-R., Meng, D., Michael, T. P., Mott, R., Muliyati, N. W., Nägele, T., Nagler, M., Nizhynska, V., Nordborg, M., Novikova, P. Y., Picó, F. X., Platzer, A., Rabanal, F. A., Rodriguez, A., Rowan, B. A., Salomé, P. A., Schmid, K. J., Schmitz, R. J., Seren, Ü., Sperone, F. G., Sudkamp, M., Svardal, H., Tanzer, M. M., Todd, D., Volchenboum, S. L., Wang, C., Wang, G., Wang, X., Weckwerth, W., Weigel, D., and Zhou, X. (2016). 1,135 genomes

reveal the global pattern of polymorphism in arabidopsis thaliana. *Cell*, 166(2):481 – 491.

Ball, F. and Stefanov, V. T. (2005). Evaluation of identity-by-descent probabilities for half-sibs on continuous genome. *Mathematical Biosciences*, 196(2):215 – 225.

Barnosky, A. D., Matzke, N., Tomiya, S., Wogan, G. O. U., Swartz, B., Quental, T. B., Marshall, C., McGuire, J. L., Lindsey, E. L., Maguire, K. C., Mersey, B., and Ferrer, E. A. (2011). Has the earth's sixth mass extinction already arrived? *Nature*, 471(7336):51–7.

Beichman, A. C., Huerta-Sanchez, E., and Lohmueller, K. E. (2018). Using genomic data to infer historic population dynamics of nonmodel organisms. *Annual Review of Ecology, Evolution, and Systematics*, 49(1):433–456.

Browning, B. L. and Browning, S. R. (2013). Improving the accuracy and efficiency of identity-by-descent detection in population data. *Genetics*, 194(2):459–71.

Browning, S. R. and Browning, B. L. (2010). High-resolution detection of identity by descent in unrelated individuals. *Am J Hum Genet*, 86(4):526 – 539.

Browning, S. R. and Browning, B. L. (2015). Accurate non-parametric estimation of recent effective population size from segments of identity by descent. *Am J Hum Genet*, 97(3):404–18.

Carmi, S., Wilton, P. R., Wakeley, J., and Pe'er, I. (2014). A renewal theory approach to ibd sharing. *Theoretical Population Biology*, 97:35–48.

Ceballos, G., Ehrlich, P. R., and Dirzo, R. (2017). Biological annihilation via the ongoing sixth mass extinction signaled by vertebrate population losses and declines. *Proceedings of the National Academy of Sciences*, 114(30):E6089–E6096.

Chapman, N. and Thompson, E. (2003). A model for the length of tracts of identity by descent in finite random mating populations. *Theor Popul Biol*, 64(2):141 – 150.

Chikhi, L., Rodríguez, W., Grusea, S., Santos, P., Boitard, S., and Mazet, O. (2018). The iicr (inverse instantaneous coalescence rate) as a summary of genomic diversity: insights into demographic inference and model choice. *Heredity*, 120(1):13–24.

Díez-del Molino, D., Sánchez-Barreiro, F., Barnes, I., Gilbert, M. T. P., and Dalén, L. (2018). Quantifying temporal genomic erosion in endangered species. *Trends in Ecology and Evolution*, 33(3):176–185.

Donnelly, K. P. (1983). The probability that related individuals share some section of genome identical by descent. *Theoretical Population Biology*, 23(1):34 – 63.

Excoffier, L., Dupanloup, I., Huerta-Sánchez, E., Sousa, V. C., and Foll, M. (2013). Robust demographic inference from genomic and snp data. *PLoS Genet*, 9(10):e1003905.

Fu, Y. X. (1995). Statistical properties of segregating sites. *Theor Popul Biol*, 48(2):172–97.

Gibbs, R. A., Boerwinkle, E., Doddapaneni, H., Han, Y., Korchina, V., Kovar, C., Lee, S., Muzny, D., Reid, J. G., and et al. (2015). A global reference for human genetic variation. *Nature*, 526(7571):68–74.

Gordon, D., Huddleston, J., Chaisson, M. J. P., Hill, C. M., Kronenberg, Z. N., Munson, K. M., Malig, M., Raja, A., Fiddes, I., Hillier, L. W., Dunn, C., Baker, C., Armstrong, J., Diekhans, M., Paten, B., Shendure, J., Wilson, R. K., Haussler, D., Chin, C.-S., and Eichler, E. E. (2016). Long-read sequence assembly of the gorilla genome. *Science*, 352(6281).

Griffiths, R. and Marjoram, P. (1997). An ancestral recombination graph. In *Progress in population genetics and human evolution*, pages 257 – 270. Springer.

Grusea, S., Rodríguez, W., Pinchon, D., Chikhi, L., Boitard, S., and Mazet, O. (2019). Coalescence times for three genes provide sufficient information to distinguish population structure from population size changes. *J Math Biol*, 78(1-2):189–224.

Gusev, A., Lowe, J. K., Stoffel, M., Daly, M. J., Altshuler, D., Breslow, J. L., Friedman, J. M., and Pe'er, I. (2009). Whole population, genome-wide mapping of hidden relatedness. *Genome research*, 19(2):318–26.

Gutenkunst, R. N., Hernandez, R. D., Williamson, S. H., and Bustamante, C. D. (2009). Inferring the joint demographic history of multiple populations from multidimensional snp frequency data. *PLoS Genet*, 5(10):e1000695.

Harris, K. and Nielsen, R. (2013). Inferring demographic history from a spectrum of shared haplotype lengths. *PLoS Genet*, 9(6):e1003521.

Hayes, B. J., Visscher, P. M., McPartlan, H. C., and Goddard, M. E. (2003). Novel multilocus measure of linkage disequilibrium to estimate past effective population size. *Genome research*, 13(4):635–43.

Hey, J. and Wakeley, J. (1997). A coalescent estimator of the population recombination rate. *Genetics*, 145(3):833–46.

Hill, W. G. and Robertson, A. (1968). Linkage disequilibrium in finite populations. *Theor Appl Genet*, 38(6):226–31.

Hollenbeck, C. M., Portnoy, D. S., and Gold, J. R. (2016). A method for detecting recent changes in contemporary effective population size from linkage disequilibrium at linked and unlinked loci. *Heredity*, 117(4):207–16.

Hudson, R. R. and Kaplan, N. L. (1985). Statistical properties of the number of recombination events in the history of a sample of dna sequences. *Genetics*, 111(1):147–64.

Kelleher, J., Etheridge, A. M., and McVean, G. (2016). Efficient coalescent simulation and genealogical analysis for large sample sizes. *PLoS Comput Biol*, 12(5):e1004842.

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

Lapierre, M., Lambert, A., and Achaz, G. (2017). Accuracy of demographic inferences from the site frequency spectrum: The case of the yoruba population. *Genetics*, 206(1):439–449.

Lewontin, R. C. and ichi Kojima, K. (1960). The evolutionary dynamics of complex polymorphisms. *Evolution*, 14(4):458–472.

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

MacLeod, I. M., Larkin, D. M., Lewin, H. A., Hayes, B. J., and Goddard, M. E. (2013). Inferring demography from runs of homozygosity in whole-genome sequence, with correction for sequence errors. *Mol Biol Evol*, 30(9):2209–23.

MacLeod, I. M., Meuwissen, T. H. E., Hayes, B. J., and Goddard, M. E. (2009). A novel predictor of multilocus haplotype homozygosity: comparison with existing predictors. *Genetics research*, 91(6):413–26.

Maddison, W. P. (1997). Gene trees in species trees. *Systematic Biology*, 46(3):523–536.

Marjoram, P. and Wall, J. D. (2006). Fast "coalescent" simulation. *BMC Genet*, 7:16.

Marth, G. T., Czabarka, E., Murvai, J., and 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):351–72.

Mazet, O., Rodríguez, W., and Chikhi, L. (2015). Demographic inference using genetic data from a single individual: Separating population size variation from population structure. *Theor Popul Biol*, 104:46–58.

Mazet, O., Rodríguez, W., Grusea, S., Boitard, S., and Chikhi, L. (2016). On the importance of being structured: instantaneous coalescence rates and human evolution–lessons for ancestral population size inference? *Heredity*, 116(4):362–71.

McVean, G. A. T. and Cardin, N. J. (2005). Approximating the coalescent with recombination. *Philos Trans R Soc Lond B Biol Sci*, 360(1459):1387–93.

Palamara, P. F., Lencz, T., Darvasi, A., and Pe'er, I. (2012). Length distributions of identity by descent reveal fine-scale demographic history. *Am J Hum Genet*, 91(5):809–22.

Patin, E., Siddle, K. J., Laval, G., Quach, H., Harmant, C., Becker, N., Froment, A., Régnault, B., Lemée, L., Gravel, S., and et al. (2014). The impact of agricultural emergence on the genetic history of african rainforest hunter-gatherers and agriculturalists. *Nature Communications*, 5(1).

Prado-Martinez, J., Sudmant, P. H., Kidd, J. M., Li, H., Kelley, J. L., Lorente-Galdos, B., Veeramah, K. R., Woerner, A. E., O'Connor, T. D., Santpere, G., and et al. (2013). Great ape genetic diversity and population history. *Nature*, 499(7459):471–475.

Purcell, S., Neale, B., Todd-Brown, K., Thomas, L., Ferreira, M. A., Bender, D., Maller, J., Sklar, P., de Bakker, P. I., Daly, M. J., and Sham, P. C. (2007). Plink: A tool set

for whole-genome association and population-based linkage analyses. *Am J Hum Genet*, 81(3):559 – 575.

Régnier, C., Achaz, G., Lambert, A., Cowie, R. H., Bouchet, P., and Fontaine, B. (2015). Mass extinction in poorly known taxa. *Proceedings of the National Academy of Sciences*, 112(25):7761–6.

Ringbauer, H., Coop, G., and Barton, N. H. (2017). Inferring recent demography from isolation by distance of long shared sequence blocks. *Genetics*, 205(3):1335–1351.

Rodrigues, A. S. L., Pilgrim, J. D., Lamoreux, J. F., Hoffmann, M., and Brooks, T. M. (2006). The value of the iucn red list for conservation. *Trends Ecol Evol*, 21(2):71–6.

Rodríguez, W., Mazet, O., Grusea, S., Arredondo, A., Corujo, J. M., Boitard, S., and Chikhi, L. (2018). The iicr and the non-stationary structured coalescent: towards demographic inference with arbitrary changes in population structure. *Heredity*, 121(6):663–678.

Sánchez-Bayo, F. and Wyckhuys, K. A. (2019). Worldwide decline of the entomofauna: A review of its drivers. *Biological Conservation*, 232:8–27.

Schiffels, S. and Durbin, R. (2014). Inferring human population size and separation history from multiple genome sequences. *Nat Genet*, 46(8):919–25.

Sheehan, S., Harris, K., and Song, Y. S. (2013). Estimating variable effective population sizes from multiple genomes: A sequentially markov conditional sampling distribution approach. *Genetics*, 194(3):647–662.

Stam, P. (1980). The distribution of the fraction of the genome identical by descent in finite random mating populations. *Genetical Research*, 35(2):131–155.

Stefanov, V. T. (2000). Distribution of genome shared identical by descent by two individuals in grandparent-type relationship. *Genetics*, 156(3):1403–10.

Terhorst, J., Kamm, J. A., and Song, Y. S. (2017). Robust and scalable inference of population history from hundreds of unphased whole genomes. *Nat Genet*, 49(2):303–309.

Tiret, M. and Hospital, F. (2017). Blocks of chromosomes identical by descent in a population: Models and predictions. *PLoS one*, 12(11):e0187416.

van der Valk, T., Díez-del Molino, D., Marques-Bonet, T., Guschanski, K., and Dalén, L. (2019). Historical genomes reveal the genomic consequences of recent population decline in eastern gorillas. *Current Biology*, 29(1):165–170.e6.

Wallberg, A., Han, F., Wellhagen, G., Dahle, B., Kawata, M., Haddad, N., Simões, Z. L. P., Allsopp, M. H., Kandemir, I., De la Rúa, P., Pirk, C. W., and Webster, M. T. (2014). A worldwide survey of genome sequence variation provides insight into the evolutionary history of the honeybee apis mellifera. *Nat Genet*, 46(10):1081–8.

Watterson, G. A. (1975). On the number of segregating sites in genetical models without recombination. *Theor Popul Biol*, 7(2):256–76.

Wiuf, C. and Hein, J. (1999). The ancestry of a sample of sequences subject to recombination. *Genetics*, 151(3):1217–28.

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

Zhang, W., Westerman, E., Nitzany, E., Palmer, S., and Kronforst, M. R. (2017). Tracing the origin and evolution of supergene mimicry in butterflies. *Nature Communications*, 8(1).

Zhao, S., Zheng, P., Dong, S., Zhan, X., Wu, Q., Guo, X., Hu, Y., He, W., Zhang, S., Fan, W., Zhu, L., Li, D., Zhang, X., Chen, Q., Zhang, H., Zhang, Z., Jin, X., Zhang, J., Yang, H., Wang, J., Wang, J., and Wei, F. (2013). Whole-genome sequencing of giant pandas provides insights into demographic history and local adaptation. *Nat Genet*, 45(1):67–71.