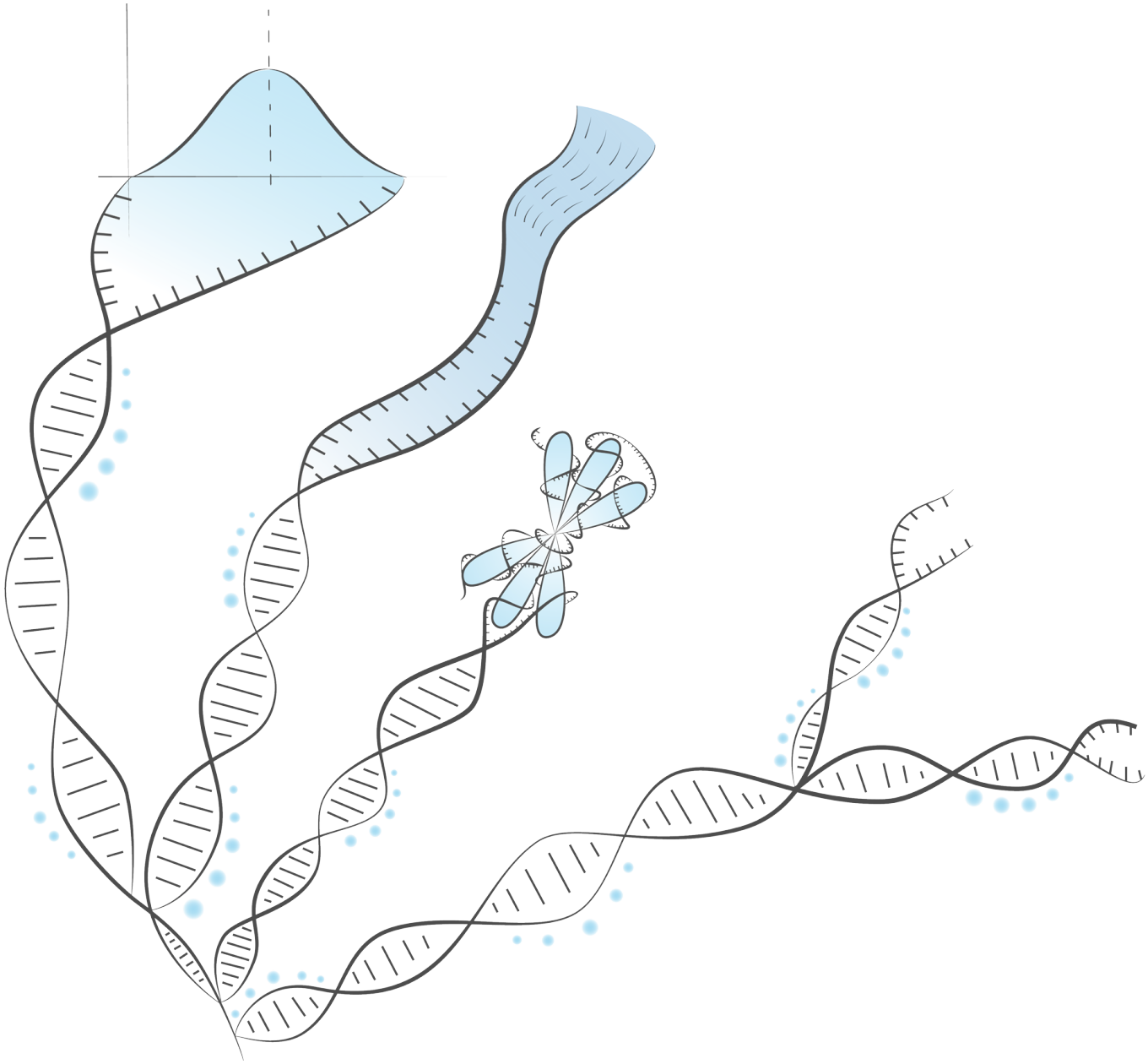
PhD thesis

Samuele Soraggi

Theory and inference on gene flow and ploidy numbers from NGS data



The PhD School of Science - The Faculty of Science - University of Copenhagen

Samuele Soraggi

Department of Mathematical Sciences
University of Copenhagen
Universitetsparken 5
DK-2100 København Ø
Denmark

samuele@math.ku.dk

This thesis has been submitted to the PhD School of The Faculty of Science, University of Copenhagen
January 31st, 2018

Academic advisors:	Carsten Wiuf Department of Mathematical Sciences University of Copenhagen, Denmark
	Anders Albrechtsen Department of Biology University of Copenhagen, Denmark
Assessment Committee:	Hans Siegismund (chair) Department of Biology University of Copenhagen, Denmark
	Jeff Wall Institute for Human Genetics University of California San Francisco, USA
	Thomas Mailund Bioinformatics Research Center University of Aarhus, Denmark

ISBN: 978-87-7078-905-9

ABSTRACT

Next-Generation Sequencing technologies have been a revolution for researchers in genetics, providing them quickly and at low cost with large amounts of DNA data from many individuals. This new flow of information has helped in revealing unanswered questions in many branches of genetics. However, NGS data suffers of intrinsical errors and quality issues due to the sequencing process, therefore SNP and genotype calling are not reliable. Such an uncertainty can bias research results, leading to the impossibility of making conclusions based on data, or even worse, leading to wrong results.

The first part of this thesis explores two different ways of handling uncertainty in NGS data by analyzing and implementing two computational tools. The first tool is illustrated in a tool called D-statistic, that is used for testing the genetic relationship amongst four populations. Here we implemented and studied an improved version of the D-statistic that does not need to call genotypes or SNP, and uses all reads from all available genomes. This results in a more powerful and reliable instrument to test genetic relationships.

The second tool integrates information about coverage and unobserved genotypes into a Hidden Markov Model to infer ploidy levels in a genome. The application on a dataset of whole genomes of the fungus Batrachochytrium dendrobatis, which is a parasitic fungus of frogs, shows inferred ploidy levels compatible with the ones that can be detected from the sequencing coverage.

In the second and last part of this thesis, a mathematical background for genetic relationships between populations is laid out. A genetic relationship between populations is typically modelled through a type of graph called admixture graph, that takes into account migrations between populations. Computational methods to test or infer a genetic relationship are now a standard in research publications, but the necessary mathematical background has not been laid out. Here we formalize a mathematical theory that connects to the current applications in population genetics, and creates a relationship between the topology of the graph and the parameters that characterize a genetic relationship between populations.

RESUME PÅ DANSK

Next-Generation Sequencing (NGS) data har været en revolution for forskere i genetik. NGS data har gjort det muligt hurtigt og billigt at generere store mængder DNA-data fra mange individer. Selvom denne nye informationsstrøm har hjulpet med at afsløre ubesvarede spørgsmål indenfor genetikken, lider NGS-data af iboende fejl og kvalitetsproblemer på grund af sekventeringsprocessen. Derfor er bestemmelsen af SNPs og genotyper ikke altid pålidelig. Sådan en usikkerhed kan gøre det vanskeligt at drage konklusioner baseret på data, eller endnu værre, føre til forkerte resultater.

Den første del af denne afhandling analyserer to forskellige måder at håndtere problemer i NGS data på, ved implementeringen af to forskellige stykker software. Det første implementering er D-statistikken. Det bruges til at teste det genetiske slægtskab imellem fire populationer. Her implementerer jeg en forbedret version af D-statistikken, der ikke bruger genotype og SNP bestemmelse, men indlæser alle sekvensdata. Denne forbedrede D-statistik er en mere robust og pålidelig måde at teste genetiske slægtskabsforhold.

Den anden software integrerer information om sekventeringsdækning og uobserverede genotyper i en Hidden Markov Model for at udlede ploiditetsniveauer i et genom. En test på et genom af svampen Batrachochytrium dendrobatis, som er en parasitisk svamp på frøer, viste udledte ploiditetsniveauer, der stemmer overens med dem, der blev estimeret ud fra sekventeringsdækningen alene.

I den anden og sidste del af denne afhandling vises der en matematisk baggrund for det genetiske slægtskabsforhold mellem populationer. Et genetisk slægtskabsforhold mellem populationer modelleres typisk gennem en type graf der kaldes en admixture graph. En admixture graph modellerer også migrationer mellem populationer. Implementering af software til at teste eller aflede et genetisk slægtskabsforhold er nu standard i forskningspublikationer, men den nødvendige matematiske baggrund er ikke blevet lagt ud. Her formaliserer jeg en matematisk teori, der forbinder de nuværende applikationer i populationsgenetik, og definerer et forhold mellem grafens topologi og de parametre, der karakteriserer et genetisk slægtskabsforhold.

Contents

Introduction

The overall focus of this thesis is the theoretical study, the statistical analysis and implementation of models targeted to genetics data. I will analyze two methods implemented for Next Generation Sequencing (NGS) data, where the issues related to such data are tackled in different ways, and illustrate a background theory for graphs used to relate populations.

Overview of the Thesis

The first analyzed and implemented method is an extension of the D-statistic, and has been published on the February 2018 issue of G3: Genes, Genomes, Genetics. The D-statistic is used to define a formal test, also known as the four-populations test or the ABBA-BABA test, to verify the fulfillment of the hypothesized genetic relationship in Figure 1 between four populations H_1, H_2, H_3, H_4. Here we use multiple genomes per population to reduce the bias in the calculated value of D, and both SNP and genotype calling are avoided. Moreover the implementation illustrated in this thesis is able to correct for errors due to deamination of the genome and to accomodate the introgression caused by a population external to the hypothesized tree, in order to unbias the four-population test.

Closely related to this method is the theoretical analysis of admixture graphs and F-statistic. A manuscript on this topic ready for submission in the Bulletin of Mathematical Biology is part of this thesis. The admixture graphs are used to describe the genetic relationship between populations, where each population is represented by a node (see Figure 2). With the use of moment statistics, namely F-statistics, calculated from genetic data, it is possible to infer a graph or test its fitness to the data. In this thesis, a background theory for the admixture graphs and the F-statistics is proposed, in connection with the population genetics framework. The F-statistics are the basis of many methods based on admixture graphs, including the four-population test, where the graph of Figure 1 is described by an admixture graph.

The second method discussed and implemented is a preliminary and minor work of this Ph.d. thesis, and originates from an exchange period at Imperial College London under the supervision of Dr. Matteo Fumagalli. The method proposes a Hidden Markov Model (HMM) for detecting and inferring variations in the ploidy number (or ploidy) from NGS data, where ploidy is the number of sets of chromosomes in a cell. The implementation is able to detect the ploidy and uses genotype likelihoods as an aid to achieve the result (see Figure 3). From another point of view, the method can also be used to detect errors in mapping sequenced data if the true ploidy numbers are already known.

A page illustrating scientific contributions and future perspectives of the three aforementioned works is in the page preceding each manuscript at the end of this introduction.

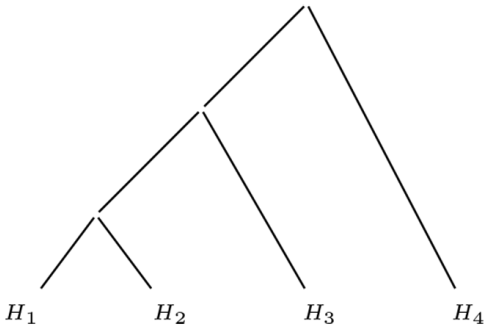


Figure 1: **Tree topology for the D-statistic**. Hypothesis of genetic relationship between four populations H_1, H_2, H_3, H_4, on which the four-population test is developed. Note that H_4 is assumed to be an outgroup.

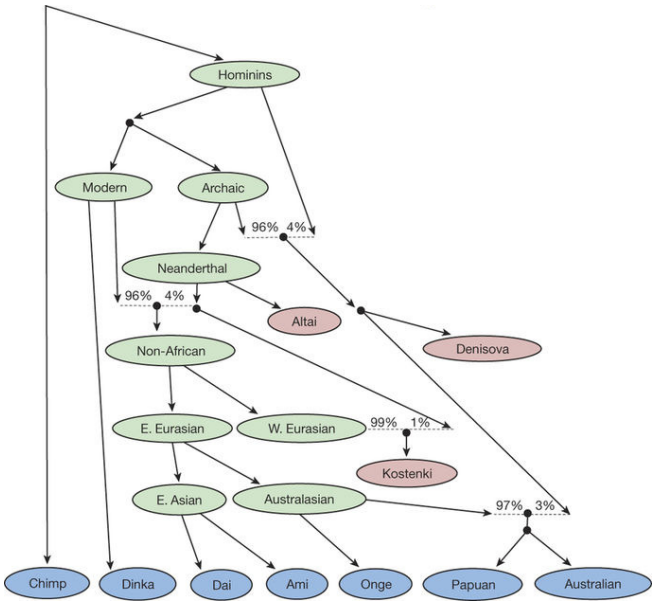


Figure 2: **Example of admixture graph.** Admixture graph (with four admixture events) representing the ancestry of some present-day populations. Source: [1].

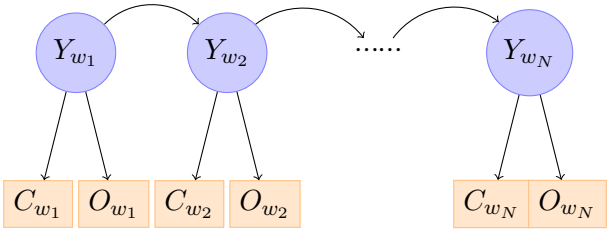


Figure 3: **Hidden Markov Model for ploidy inference**. Graphical representation of the Hidden Markov Model used to infer the ploidy numbers. The Markov chain $\{Y_{w_i}\}_{i=1}^{N}$ represents the unknown ploidy numbers on N windows of loci. The observations C_{w_i}, O_{w_i} are the average coverage and sequenced bases at window i, respectively, for $i = 1, \ldots, N$.

Background

Elements of Biology for Beginners

This section contains the key definitions related to the biology of DNA. This background section is necessary to understand the terminology of the topics of this thesis.

Cells and DNA

Cells are the basic element of living organisms. They give structure to the body, intake nutrients and convert them into energy, and carry out special tasks. Cells contain the hereditary material of an organism and can copy themselves. A cell is composed by many parts and organs, amongst which the nucleus. The nucleus is the control room of the cell, and contains the DNA (deoxyribonucleic acid), in which the hereditary information of the organism is stored. The DNA present in a cell is called the genome. The organisms whose cells have a nucleus are called eukaryotic (e.g. mammals), otherwise prokaryotic (e.g. bacteria).

The DNA consists of small molecules called nucleotides. There are four possible nucleotides decoded by four letters corresponding to four chemical bases: **A** (Adenine), **C** (Cytosine), **G** (Guamine), **T** (Thymine). We can consider the DNA molecule as a word of a certain length over the set of letters $\{A, C, G, T\}$ that is characterized (for chemical reasons) by a direction: from the $5'$ side to $3'$ side, where the numbers $5', 3'$ are due to chemical conventions.

Each nucleotide of the DNA is chemically bonded with a complementary one, specifically A with T and G with C, to form a basepair (bp). The DNA is then seen as a word written from the direction $5'$ to $3'$ complemented by a word written in the opposite direction, that is, from $3'$ to $5'$. Basepairs are found sequentially on a DNA and are tied together by two backbones of sugar and phospate. The position of a basepair on the DNA is called locus and the length of a genome is its number of basepairs (see Figure 4).

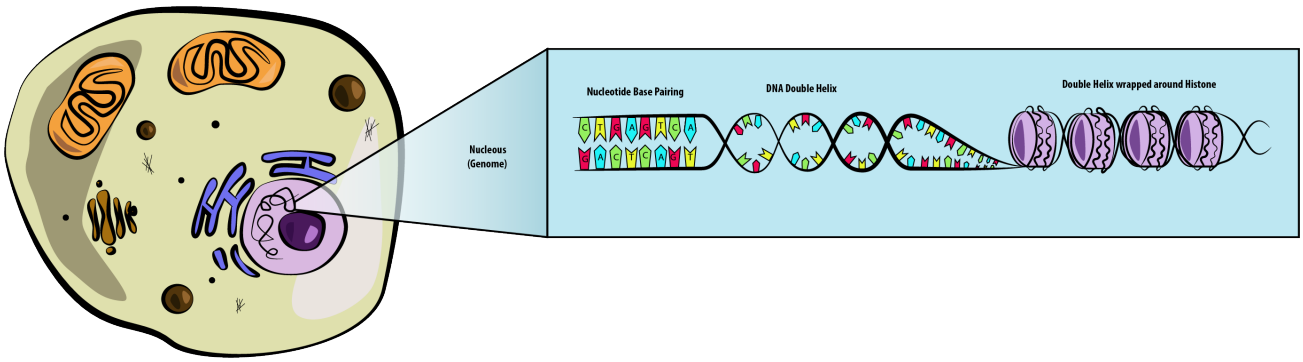


Figure 4: **Representation of a cell and the DNA contained in the nucleus.** Illustration of a cell's structure and detail of the basepairs in a section of the DNA helix, finally wrapped around a histone to form a chromosome.

Chromosomes and ploidy

The DNA is wrapped around proteins called histones to form structured threads called chromosomes (see Figure 4). Each chromosome is grouped with its homologue, e.g. in singletons, pairs, triplets, etc., and the organism is then called haploid, diploid, triploid, etc. The bases of the $5'$-to-$3'$ DNA sequences at the same locus in grouped chromosomes form the genotype.

Sexual organisms such as mammals are usually diploid, that is, they have N paired copies of chromosomes (N = 22 for humans, plus two sex chromosomes), where each chromosome of a pair comes from each mating parent. After male and female gametes (haploid sex cells) are generated through a process called cell division,

they can mate and form a new organism (zygote). Starting from the union of the two haploid gametes, the zygote will develop into a diploid organism, essentially through a process of cell replication.This process of reproduction happens essentially in all eukaryotes organisms (plants, animals, fungi, humans, etc.), with some minor differences.

The diploid state seems to be the favoured one in nature to enable sexual reproduction. However, genomes more than diploid have been observed in plants and fungi already more than one hundred years ago. Such property is called polyploidy and is considered being a very important mechanism in speciation of organisms. Haploidy, diploidy and polyploidy are prevalently observed in plants and fungi, ranging from haploid (some types of fungi) up to dodecaploid plants, while animals are in general diploid.

In some cases it can happen that some steps of cell division prior to mating happen erroneously, leading for example to a wrong number of chromosomes in a gamete and causing aneuploidy (abnormal number of chromosomes in a cell), that can cause death or developmental problems of some organism (e.g. Down syndrome in humans, where chromosome 21 is triploid). Other variations can lead to small aneuploidy portions of the genome without consequences.

Cancer cells are often characterized by aneuploidy in the host organisms. Cancer cells are essentially cells that do not respond anymore to the normal signals governing their growth and death. Normally, a cell reproduces a finite number of times, and destroy itself when its genetic material results too damaged. This does not happen in cancer cells. Here, mutations (see next section) in the parts of DNA governing those mechanism lead to abnormal behaviours: accelerated cell replication, fast generation of new mutations, altered cell duplication resulting in aneuploidy, etc.

Sources of Genetic variations in a Population

The DNA can undergo changes that are cause of genetic variation, that is, variation of genomes between members of species, or between groups of species located in different parts of the world. Genetic variations can be essential elements in the future survival of organisms over different geographical locations and environmental conditions. Through the study of genetic variations scientists aim for example at tracking history of past populations, characterizing pathologies, determining the lineage dynamic of species of organisms.

Genetic variations are first introduced through mutations. Mutations can be of different types:

- **single nucleotide variation (SNV)**: inheritable base substitution at one or more loci of the DNA,
- **insertion or deletion**: insertion or deletion of a string of DNA sequence,
- **copy number variation (CNV)**: replication of a section of DNA a certain number of times.

Once mutations are introduced, ulterior variability is introduced through recombinations. This is the exchange of information between chromosomes in the process of creating of a zygote. In such a way the correlation between different loci can be changed and eventually broken, and mutations can change their position in a genome. Loci that are physically close to each other on the DNA are more unlikely to be separated by effect of recombination. The more those loci keep being close through time, the more they are said to be genetically linked. Two loci are said to be unlinked when they are found on two different groups of chromosomes.

The rate at which an SNV happens at each nucleotide is of the order 10^{-9}/year in humans [2, 3]. A Single Nucleotide Polymorphism (SNP) is a variation at a single locus in a DNA sequence between individuals. Usually, if more than 1% of a population does not carry the same nucleotide at a specific position in the DNA sequence, then this variation can be classified as a SNP.

Mutations can be useful, e.g. when caused by the pressure for adapting to an environment. In this case they are said to be advantageous mutation. A mutation can otherwise be neutral (no effect in terms of adaptation) or deleterious (negative effect in terms of adaptation).

Mathematical Modeling of Genetic Data: The Wright-Fisher Model

This section illustrates the Wright-Fisher model for genetic data at a single locus [4, 5]. This is a basic mathematical model to explain how a population of N individuals (genes) evolves through non-overlapping generations. Here we assume a population of N haploid individuals with alleles (types) A and B. We overlook some details that in reality influence the behaviour of the system, e.g. population structure, population size distribution, selection etc. Main references for a deeper mathematical treatment of this model are [6, 7].

The Wright-Fisher model illustrates how the allele frequencies evolve in a population of finite size N. Each individual is of one type (A or B) and we ignore the effects of mutations. At each non-overlapping generation the population of N parents is sampled with replacement with probability $1/N$ to form children in the next generation.

Let Z_i be the r.v. that describes the number of offspring of individual $i \in \{1, \dots, N\}$; the multivariate random variable $(Z_1, \dots, Z_N)$ is multinomially distributed with sampling probabilities $1/N$. Therefore each Z_i is a $\mathcal{B}in(N, \frac{1}{N})$. Consider the random variables $\{C_r = \text{number of A alleles at generation } r\}_{r\in\mathbb{N}}$. Given $C_r = i$ for some $r \geq 0$ and $x \in \{1, \dots, N\}$, let $\frac{i}{N} =: x_i$ be the observed frequency of A alleles. Then

$$C_{r+1}|C_r = i \backsim \mathcal{B}in(N, x_i) \tag{1}$$

defines a time-homogeneous Markov Chain $\{C_r\}_{r\geq 0}$ with state space $\mathcal{S} = \{0, \dots, N\}$. States 0 and N are absorbing states for the chain (see Figure 5B). The change in the frequency of allele A through this random process is called drift. Figure 5A shows an example of Wright-Fisher model.

Let X_r denote the frequency of the A allele at generation r. The expected frequency of the A alleles at generation $r + 1$, conditionally on the count at generation r, remains the same as in generation r:

$$\mathbb{E}[X_{r+1}|X_r] = X_r. \tag{2}$$

It follows that the expected frequency at each generation is the same as the one at generation $r = 0$. Let h_r be the heterozygosity at the r-th generation, that is, the probability that two random individuals from the population at generation r have different alleles. The heterozygosity at the r-th generation is $h_r = \lambda^r h_0$, implying that $h_r \longrightarrow_r 0$ (see Figure 5B). Therefore the genetic drift reduces a population's diversity and increases the divergence between different populations.

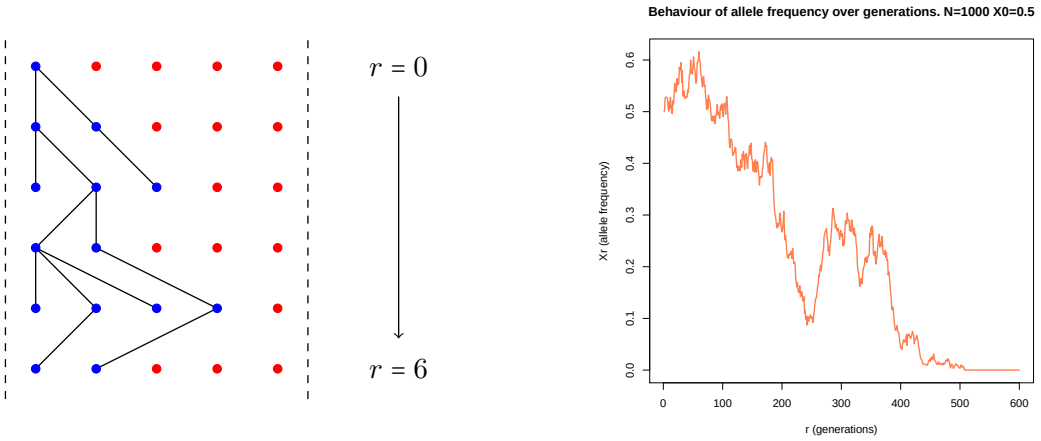


Figure 5: **Wright-Fisher model and allele frequency.** (A) A possible Wright-Fisher model with $N = 6$ individuals and following the sampling in (1). Note that it is possible that some individuals never get sampled. (B) Behavior of the allele frequency over generations for $N = 1000$ individuals and a proportion of $1/2$ for the two types at generation $r = 0$. Note that one of the two types is not present anymore in the population after around 500 generations, in accordance with the fact that the heterozygosity tends to zero as $r \to +\infty$.

Ancestral Process and Coalescent Times

One can assume a different perspective when studying the Wright-Fisher model, by using a bottom-to-top point of view (thus backward in time) in the genealogy of the individuals. In this way, we can try to answer to different questions, e.g. Where did an individual come from? What are the ancestral relationships?

Denote by g_{kj} the probability of having j different ancestors for k individuals. This is given by

$$g_{kj} = \frac{N(N-1)(N-2)\cdots(N-(j-1))}{N^k} S_k^{(j)}, \tag{3}$$

where $S_k^{(j)}$ is the Stirling number of the first kind, that is, the way of assigning k children to the j fathers.

Let t be the time variable. Denote by $\{A_n^N(t)\}_{t\geq 0}$ the process representing the number of ancestors of n individuals at time t, given a population size N. The ancestral process is given by

$$\mathbb{P}(A_n^N(t+1) = j | A_n^N(t) = k) = g_{kj},$$

with border condition $A_n^N(0) = N$. The approximation $g_{kk} + g_{k,k-1} \approx 1$ holds for (3). It is therefore highly probable to remain in the same state of the ancestral process (i.e. to have the same number of parents) or to jump to the next state (i.e. to have one parent less) as in Figure 6.

Consider a large population size N, ideally $N \to \infty$. Rescale the time in unit of N generation by $r = \lfloor N \cdot t \rfloor$. Let T_k be, for $k \geq 2$, the time while a sample of size k has exactly k ancestors. In other words, the time until which k individuals coalesce. It follows that for $N \to +\infty$ the distribution of T_k is exponential of parameter $\binom{k}{2}$. Note that $E(T_2) = 1$ and $\sum_{k=2}^{N} T_k = 2$, hence almost half of the time spent in coalescing N individuals is necessary to coalesce the two main ancestral branches (see Figure 6).

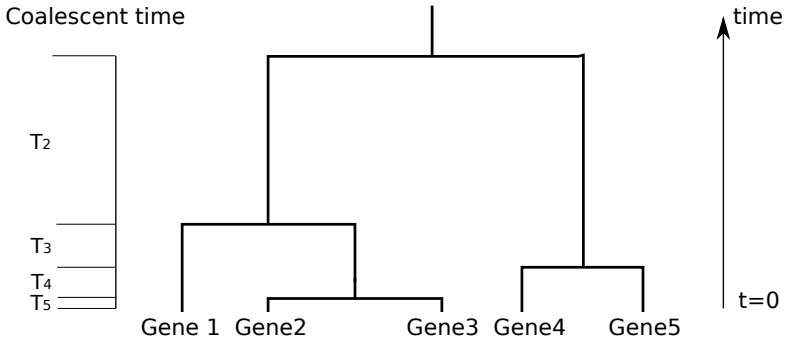


Figure 6: **Example of coalescent times.** Example of coalescent process of five genes present at time $t = 0$. Much time is spent in T_2, and exponentially distributed coalescent times decrease when the number of genes grow.

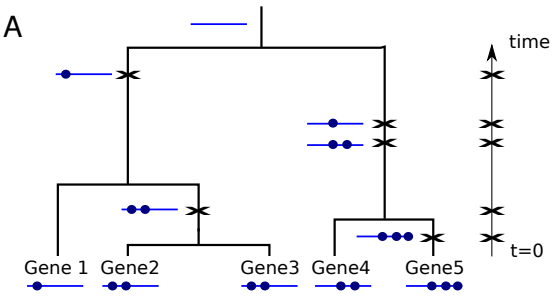
Wright-Fisher Infinitely-many Sites Model

The Wright-Fisher Infinitely-many Sites Model [8, 9] maintains the binomial sampling nature of the Wright-Fisher model. Here each gene of the Wright-Fisher model is considered as a sequence with infinite number of sites, where each allele is drawn from the set $\{0, 1\}$. The sampling of allele 1 is a mutation, and happens with probability u, called mutation rate. Whenever a mutation is verified at a locus, a new type is created and randomly sampled from an uniform variable in the interval $[0, 1]$. In this way it is possible to keep track of the mutations that happen along the lineages (see Figure 7).

The infinitely-many sites model is considered to be a reliable mathematical explanation of genetic data. Speaking in terms of biology, it is possible to observe that in a DNA sequence there are very few loci where variations happen, and those correspond often to one or two alleles, suggesting that at most one mutation can happen at a locus.

Consider the rescaling $\theta = \lim_{N\to\infty} 2Nu$, meaning that the mutation rate is of order reciprocal to the population size, and $r = Nt$. The number of mutations at time t (backward) on a lineage has $Poisson(\frac{\theta}{2}t)$ distribution for $N \to \infty$. This means that mutations are very unlikely to happen on a range of generations relatively short when the mutation rate is very small, such as when we consider recent splits between human populations. The Poisson nature of the number of mutations on a lineage implies that mutations are uniformly distributed along the lineage length.

A



B

Position:	0.13	0.28	0.62	0.7	0.91
Gene 1	1	0	0	0	0
Gene 2	1	1	0	0	0
Gene 3	1	1	0	0	0
Gene 4	0	0	1	1	0
Gene 5	0	0	1	1	1

Figure 7: **Coalescence and representation of sequences.** (A) Coalescence of five sequences subject to mutations according to the infinitely-many sites model. The sequences are represented by blue lines, and a dot is added randomly on a line to mark the locus at which a mutation happen. Each mutation on a lineage is denoted with a cross and the mutated gene. Each mutation is reported on the time line. (B) Table representing the sequences only at the random positions where mutations happen. Each mutation is characterized by the allele 1.

The Importance of Allele Frequencies in Modeling Genetic Relationships

Mutation rates alone are not responsible of many changes in allele frequencies, if not over very long time frames. Due to the Poisson nature of mutation events as a function of both time and mutation rate [7], mutations happen very rarely when mutation rates are low, and only few of them aren't lost in the population of interest. Therefore mutations are generally not considered when modeling allele frequencies within different lineages of the same species [10]. Note that this causes the expected allele frequencies to fulfill (2) under the Wright-Fisher model. Moreover the conditional variance $Var(X_{r+1}|X_r = x_i)$ is given by $x_i(1 - x_i)/N$, for $i = 1, \ldots, N$, as a consequence of the binomial sampling in (1). Hence one expects small variations in allele frequencies in a Wright-Fisher model with absence of mutations and with large population size.

However, changes in the genetic drift (the sampling process of types) are introduced when taking into account other factors [10, 11] that are function of the population size, such as:

- bottlenecks: the population size is greatly reduced. In this case allele frequencies increase their variance and the drift varies more.
- founder effect: similarly to bottlenecks, it happens when a fraction of a population becomes isolated from the rest of the individuals.

Moreover, allele frequencies can be altered as a result of the following processes [10]:

- selection: some types gain reproductive advantage, and therefore alter the random process of sampling a new generation.
- population structure: individuals in a population do not mate randomly because of geographical factors, leading them into mating preferences.
- gene flow: gene flow occur when a population of individuals intakes other individuals from a different population. Given allele frequency x and x' of the two populations, respectively, the allele frequency of the newly generated population is modeled as $\alpha x + (1 - \alpha)x'$. In other words, the gene flow is considered

as an instantaneous pulse such that two portions of genetic material are inherited from the previously separated populations [12, 13].

Allele frequencies have become of increasing interest in studying relationship between populations. The use of frequencies has given rise to different methods to estimate which relationships are wrongly hypothesized or best fitted by available genetic data [13–18]. In the first manuscript of this thesis [19], observed allele frequencies are used to characterize a computational method to detect the presence of admixture. In the second manuscript, a stochastic model for graphs relating populations is developed. Here we will also focus on the analysis of the moment statistics called F-statistics, calculated from allele frequencies [12, 13]. The F-statistics are a fundamental building block of computational applications [12, 13, 15, 20, 21] for understanding the past history of different populations. Many of the assumptions in the analysis of Manuscript 1 [19] and 2 follow the properties of allele frequencies treated in this section.

Next Generation Sequencing Data

The recent technological developments in the field of DNA sequencing data has provided scientists with a large amounts of genetic data produced faster and cheaper than in the past. In this thesis the focus is on Next Generation Sequencing (NGS) data [22–24]. More properly, this is the second generation of NGS data and includes various protocols, amongst which Illumina [25], the one primarily used for sequencing the data applied in this thesis.

Sequencing Pipeline 101

The generation of sequenced data follows an NGS protocol characterized by some essential steps [22] (see Figure 8A). Firstly, DNA is extracted and DNA fragments are prepared from it. Then fragments are subject to enrichment (essentially some of them are selected), and PCR amplification, after which a larger library of fragments is available. Thereafter, the sequencing process generates short reads.

Raw data

The output data of an NGS system is given by reads whose length is of the order of hundreds of pairs. Each base is an i.i.d. sample from the true genotype at its locus. In term of file format, the sequenced reads are collected in a `.fastq` or `.fq` file [26], in which each read takes four lines as follows:

1. identifier of the sequence and eventual description,
2. sequence of bases in the read,
3. a + symbol and eventual other identifiers and comments,
4. encoded representation of quality values, one character for each base.

Quality values are taken from the ASCII alphabet of characters that corresponds to numeric values ranging from 33 to 126. From those values it is possible to calculate the quality scores. There are two standards for the quality scores (Phred and Solexa); we work with the Phred system since it is the one our data is based on. The quality score of a base, that is, the probability that the nucleotide's base of a read is wrongly sequenced, is calculated as

$$\epsilon = 10^{\frac{-(Q-33)}{10}},$$

where Q is the ASCII integer value at that base. The obvious use of this coding system for `.fastq` files is the compression of floating numbers into single characters, even though the probabilities become discretized.

Processed data

The output of the sequencing process is given by the reads and their quality scores. This data has to undergo a preprocessing step to accomodate for various artifacts and filter reads and bases. Thereafter reads are assembled, that is, aligned to a reference sequence or de novo (see Figure 8B). At each locus the aligned bases are samples with replacement from the true genotype. The resulting file (in `.bam` format [27]) contains the reads with the coordinates of their alignment. Depending on the alignment technique that is used, every locus has assigned a mapping quality score $mapQ$, that can be converted into probabilities with the Phred score formula. This is the probability that the estimated alignment of reads is wrong at a specific locus.

The depth (or coverage) at a locus of the sequenced genome is the number of times a base is read at that locus from aligned reads. The whole genome coverage is the ratio between the total number of sequenced bases and the length of the sequenced genome. The depth can be modeled using a Poisson distribution, but often the allele counts are overdispersed, therefore the negative binomial distribution is preferred [28].

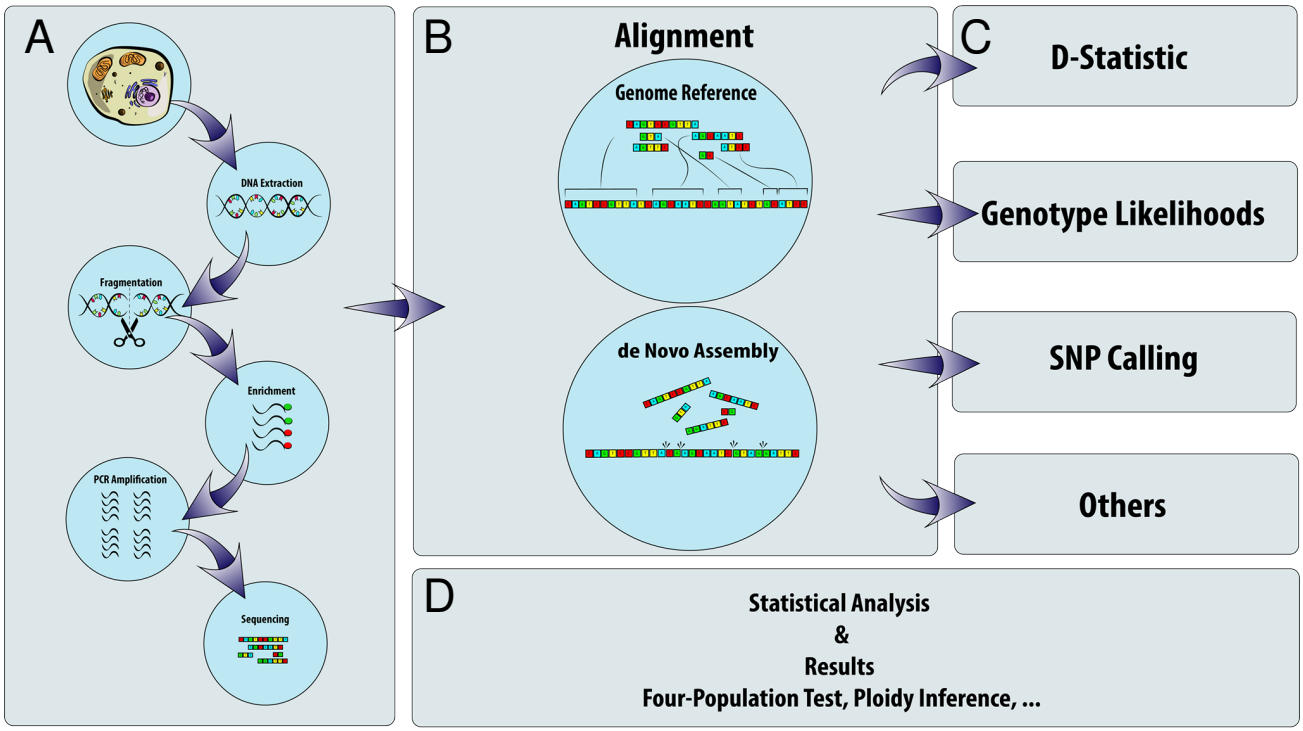


Figure 8: **NGS Data pipeline.** (A) Sequencing pipeline from a high-throughput sequencing machine. The DNA is first extracted from the cell's nucleus and reduced into small fragments. Thereafter some of the fragments can be enriched. In other words they are selected according to some characteristic and preserved for the next step. Through Polymerase Chain Reaction (PCR) the fragments are repeatedly duplicated to form a larger library. The last step is sequencing: here single base pairs on the fragments are identified and output in a digital memory as reads, that is, strings containing basepairs and other information such as quality scores (identification error for each basepair). (B) Once reads are output from the sequencing machine, they can be aligned against a reference genome or denovo. In the first case, each read is mapped to a reference sequence, and reads might be completely or partially stacked when matching the same portion of sequence. Denovo assembly builds a genome without mapping it to a reference. (C) A variety of operations can be carried out using aligned data, such as calculating the D-statistic, the genotype likelihoods, call SNPs and genotypes, etc. (D) After the necessary calculations are performed, statistical analysis (e.g. ABBA-BABA test, ploidy inference, etc.) can be carried out to obtain the necessary results (admixture detection, poliploidy, etc.).

Given the overview and background for the thesis, the following section introduces a broad overview of NGS data and some of the challenges that scientists face in its analysis.

NGS Data and its Challenges

The advances in sequencing technologies in the last decade have provided scientists with high-throughput data, for example through NGS techniques [23, 29], that have allowed increased speed and reduced cost of the sequencing process. There are many different protocols available on the market to produce NGS data (e.g. 454, SOLiD, GeneReader, Ion Torrent, Illumina) [23, 24, 29], where all of them essentially provide at the end of the sequencing scheme an output that consists of reads of a certain length (in the order of 100 bases for Illumina technology), that are either aligned to an available reference genome, or organized in scaffolds (denovo assembly) when a reference is not available.

However, the use of NGS data encounters multiple challenges that need to be addressed. In situations where large genomes are sequenced, and as long as there is no further reduction in sequencing costs, the reasonable trade-off "costs vs. size" of the samples leads to the use of low-depth data, that is, with a depth lower than $5X$. In other words each base is sequenced on average less than five times. Low sequencing depth can be even more problematic in ancient genomes, where the depth can be greatly lower than $1X$, and alleles are characterized by high error rates due to deamination of the sample (a chemical process) before sequencing [30, 31].

Why is low depth a problem in the analysis of NGS data? NGS data is affected by relatively high sequencing error. For example, current Illumina sequencing shows higher sequencing error when compared to the Sanger sequencing [32, 33], a method established in the late 1970s, and the most used one for many decades. Sequencing errors cause the wrong sequencing of an allele, and together with low depth can make it problematic to perform SNP calling (the inference of polymorphic sites), because a polymorphism might just be a sequencing error, and there are not enough bases to determine if the locus is in effect polymorphic. Similarly it becomes difficult to perform reliably genotype calling (the inference of genotypes). In fact the genotype could be inferred just looking at the proportion of alleles at a site if the depth would be high enough. For example, a diploid individual with genotype CT at a certain locus, is expected to have a 50% proportion of aligned C alleles at that locus (see Figure 9A). With low depth data this proportion is easily altered due to the lack of observed reads, sequencing and alignment errors (see Figure 9B) and immediate inference of the reference genotype is not allowed.

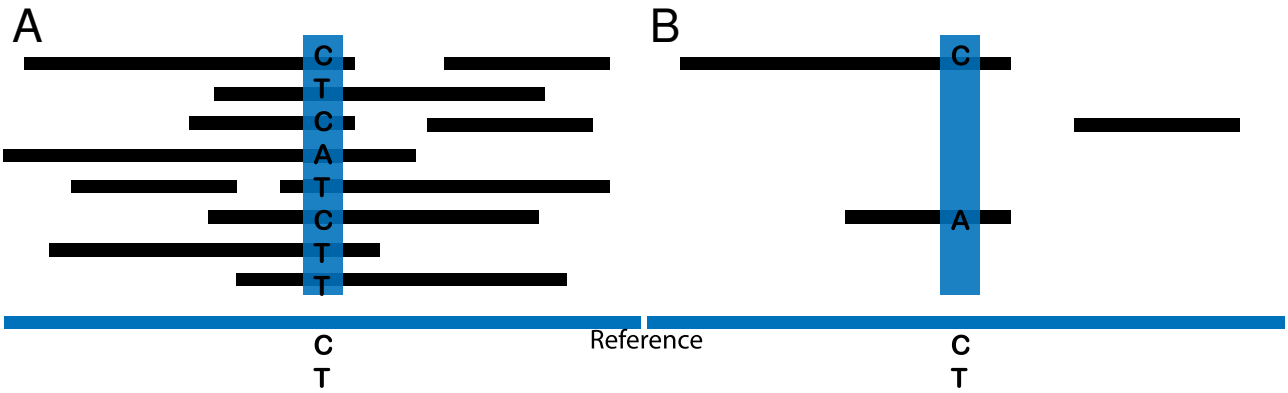


Figure 9: **High- and low-depth reads (black) aligned to a reference genome (blue).** Sequenced reads aligned to a reference genome, and details of the bases at a locus where the true genotype is CT. In figure (A) the depth is high and there are enough reads to estimate the genotype, even though there is an allele that is wrongly sequenced or misaligned. In figure (B) the estimation of the genotype is not possible due to lack of data and an erroneous base.

The two computational methods developed in this thesis will tackle the problems of NGS data in two different ways by

- using all the aligned reads from multiple genomes to determine an unbiased estimator of the allele frequency [19, 34],
- characterizing genotypes probabilistically through the so-called genotype likelihoods [35] to help determining ploidy numbers.

The genotype likelihood is the probability of observing a particular genotype given the sequenced data. In the simplest form, this can be calculated by taking into account individual base qualities as probabilities of observing an incorrect nucleotide, and assuming the bases to be independent over reads.

Let R be the number of sequenced reads at a locus, O the observed data, o_r and q_r the observed nucleotide and the Phred base quality for read r, $r = 1, \dots, R$, respectively. The i-th base of genotype g is denoted by g_i, $i \in 1, \dots, y$. The genotype likelihood of g for ploidy number y is expressed as

$$\ln p(O|g, y) = \sum_{r=1}^{R} \ln \Big(\sum_{i=1}^{y} \frac{1}{y} p(o_r|g_i, q_r, y) \Big)$$

where

$$p(o_r|g_i, q_r, y) = \begin{cases} 1 - \epsilon_r, & \textit{if } o_r = g_i \\ \frac{\epsilon_r}{3} & \textit{otherwise} \end{cases}$$

and ϵ_r is the phred probability for the base at read r. The probability ϵ of observing an incorrect nucleotide is considered homogeneous through the possible nucleotides.

Population Genetics: NGS Data and Methods

Many different fields of genetics have found beneficial the vast amount of information provided by NGS technologies. With the new information it has been possible to untangle long asked issues. Amongst those fields, there is population genetics. Population genetics is the analysis of the genetic variations amongst populations and the evolutionary processes that influence them. A crucial role in learning and understanding the genetic variation of populations and their history is played by the detection of contacts between them in past times.

Such contacts can result in gene flow and admixture between populations and therefore might leave traces of a population's history in the DNA of individuals. With the term gene flow we denote the migration of individuals from one population of a species to another, with transfer of genetic material to the receiving population through interbreeding of individuals. The consequence of gene flow is admixture, that is, the generation of a new lineage in the population receiving the gene flow. A gene flow is often denoted archaic (or ancient) when it involves both modern and ancient populations.

There has been a growing research focus in both validating and inferring scenarios of gene flows and admixtures between both moderns and ancient populations, where the term populations includes - and is not only restricted to - human and ancient human populations [12, 15, 16, 30, 36–43].

In fact not only NGS technologies have made it possible to obtain a large amount of sequenced DNA data from modern individuals, but this has happened even in the case of ancient DNA, of which there are many examples amongst humans. The genome of a more than $10,000$ years old Anzick-Clovis (from North America) was sequenced with an average depth of $14.4X$ [40]. A draft sequence of the Neandertal genome was built using samples from three Neandertal individuals [15]. It was also possible to sequence the genome of a more than 8000 years old individual found in Kennewick (Washington) [44].

However, it has to be remembered that NGS data suffers of the drawbacks induced by sequencing errors and low depth. Those biases have shown for example to affect many summary statistics that are of common use in population genetics [45], including the D-statistic [46]. When possible, part of this bias is avoided by setting a reasonable lower boundary to the depth and to the quality of the bases in each locus. For example the authors of [39] set a lower bound of $10X$ for the depth when they call genotypes to apply the D-statistic.

A large number of statistical methods have been developed to study the relationship between populations through genetic data. A wide class of methods embraces the model-free methods. In this case the genetic

relationship between populations is based on probabilistic assumptions not related to any sort network or graph structure. Some of those methods are based on probabilistically assigning an individual to a certain number of admixing populations (clusters) without specifying a model for the populations' history. This results in an assignment of admixture proportions to the clusters. Even though it is not possible to infer which structure relates the different populations, those methods provide an interpretation of the clusters as originating populations, or simply the proof of admixture between populations.

Some methods that fall in this category are `STRUCTURE` [47], `ADMIXTURE` [48], `BAPS` [49], `iADMIX` [50] and `fastNGSADMIX` [51, 52]. Such tools estimate in which proportion the genome of an individual results from the admixture of K ancestral populations. All those methods are based on allele frequencies,while `iADMIX` and `fastNGSADMIX` use also genotype likelihoods from NGS data. The background model of all these methods is the admixture pulse. Here each j-th ancestral populations admixes with proportion α_j into the individual of interest, providing the admixed individual with a proportion of alleles that corresponds to a fraction α_j of the allele frequency. Thus for a locus $i = 1, \ldots, M$, where M is the number of available loci, the allele frequency x_i of the individual of interest is the linear combination of the allele frequencies x_{ij}, $j = 1, \ldots, K$ of the K admixing populations:

$$x_i = \sum_{j=1}^{K} \alpha_j x_{ij}.$$

Moreover it is assumed that there is ideally no time span between the time of admixture and the time at which data was acquired. The software `STRUCTURE` first assigns each individual randomly to one of the K predefined populations. Thereafter the variant allele frequencies are estimated for each individual, and admixture proportions are re-estimated for the K admixing populations. The process is repeated until a convergence criteria is met.

The implementation of `BAPS` is very similar but assumes K as an unknown variable, that is then estimated to avoid an excess or lack of fragmentation in the admixing groups. The software `ADMIXTURE` tries to achieve analogous results by maximization of the likelihood of the assumed model. The implementation of `iADMIX` and `fastNGSADMIX` start by calculating the genotype likelihoods. Afterwards they perform the EM algorithm [53, 54] on the likelihood of the admixture proportions.

Despite the popularity of the clustering techniques in population genetics, they do not work well with a limited number of individuals per population and they are not appropriate to detect ancient gene flow. In fact population frequencies are not well estimated when few individuals are available, and the assumption on the time of admixture is influenced by genetic drift. For example, the application of `fastNGSADMIX` on the configuration in Figure 10B (taken from Figure 2A in Manuscript 1) shows no sign of admixture in Figure 10A, because the admixture happened 8000 generations in the past.

Another category of methods is used to reveal patterns of population structure, based on a suitable measure of dissimilarity. A widely applied technique is the PCA [55], that reveals those patterns through the eigenvectors of a matrix measuring the pairwise genetic dissimilarity between individuals. In such a way different groups that are genetically similar can be seen as being close in the patterns. PCA is performed by tools such as `EIGENSOFT` [56] and `TASER-PC` [57]. The former is able not only to perform the PCA on provided samples, but also implements a formal test to detect the presence of underlying structure between the individuals. The latter has been developed for NGS data and makes use of genotype likelihoods.

A Model-based Method: The D-statistic

The model-free methods mentioned above are not well suited in applications involving ancient admixtures. Other methods, called model-based methods, are often used to describe ancient gene flow. In such tools, the probabilistic formulation of the relationship between populations is based on a representation through a type of graph or network, whose nodes represent in which way populations are supposed to be genetically related.

A model-based method of recent development is the D-statistic. The D-statistic is used to define a formal test to verify the fulfillment of the hypothesized genetic relationship in Figure 1. Here, H_1, H_2, H_3, H_4 are four

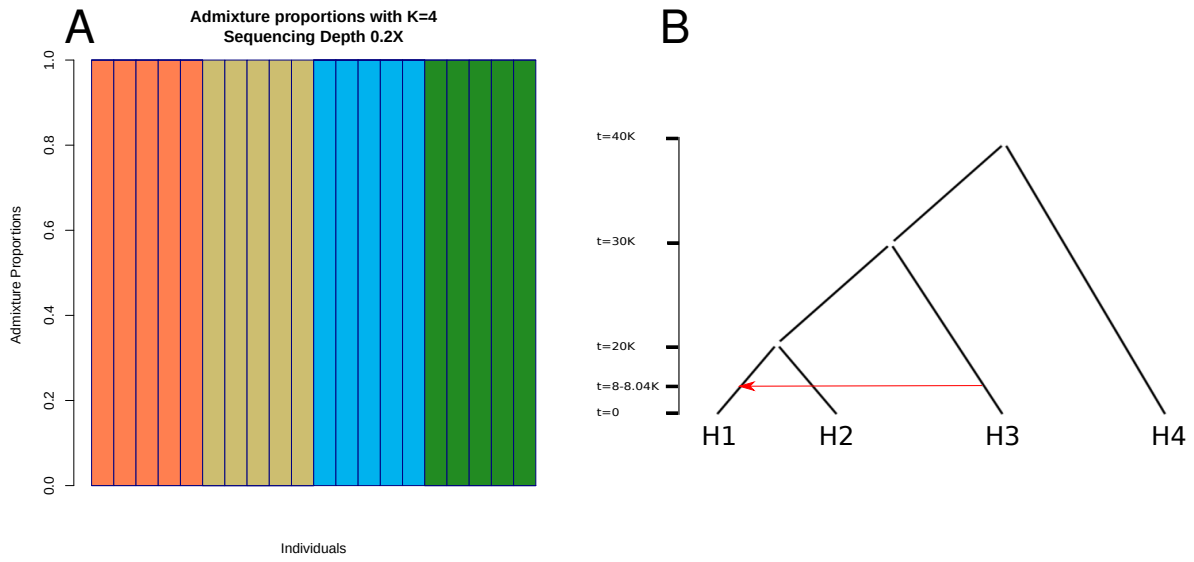


Figure 10: **Failed detection of ancient admixture from a clustering method.** (A) Application of the clustering method `fastNGSADMIX` on the tree $(((H_1, H_2)H_3)H_4)$ in the configuration in Figure (B), Taken from Figure 2A of [19] and simulated with the same parameters. Here there is migration from an external population with rate $m = 0.2\%$. The method fails because it assumes recent - ideally at zero time in the past - gene flow, and is influenced by the genetic drift from the time of admixture until present.

populations represented by the leaves of the tree, where H_4 is an outgroup population. The first application of the D-statistic can be found in [12]. Here, using a slightly different quantity called F_4-statistic, the authors are able to verify in which proportion the Indian populations of the Cline group are affected by external gene flow. In [15] the D-statistic is used to discover and quantify the genetic affinity between three non-african individuals and a Neandertal. It has been furthermore deduced that humans in Eastern Asia are subject to a higher proportion of gene flow from the Neanderthals, if compared to non-African populations located farther west [38]. The application of the four-population test to many different configurations of the hypothesis tree lead to the possibility that certain Native American populations were the result of admixture, e.g. it has been detected that Australasian populations admixed into New World Populations [16, 39].

In order to avoid SNP and genotype calling, often problematic when working with NGS data, the D-statistic relies on sampling a base at each locus according to the relative frequency of each allele in the reads [15]. This technique is the one available in widely used computational tools. For example the sampling approach is implemented for NGS data in the `doAbbababa` program of `ANGSD` [35]. An implementation for di-allelic genotype data of multiple individuals can be found in the routine `qpDstat` of `ADMIXTOOLS` [36], while the `fourpop` program of `TreeMix` [20] supports also microsatellite data.

Since the scenarios in which the D-statistic is applied often involve ancient admixtures, the DNA used in the analysis might be affected by deamination. Deamination is a process through which the DNA of an organism degrades post-mortem and results in low sequencing depth, low quality scores of the sequenced data and high frequency of base transitions. Therefore the available methods for the D-statistic cannot be always relied upon in applications involving ancient data, due to the uncertainty in calling procedures and the bias in relative frequencies of the alleles at each locus.

This part of the thesis focuses on addressing the problematics that are encountered when applying the D-statistic to NGS data. In the method we propose - called extended D-statistic - and implemented in the program `doAbbababa2` of `ANGSD` for low-depth NGS data, we calculate the D-statistic using all reads of the genomes. Differently from what happens in the sampling approach, the use of multiple individuals for each population

is allowed, and furthermore there is no requirement on the sequencing depth of the different individuals. The extended D-statistic is approximated by a standard normal distribution and our improvements do not alter the unbiasedness of the estimated frequencies used to calculate D.

In order to address the issue of type-specific errors, we correct for type-specific error rates in the data, so that the reads used to calculate the D-statistic will not bias the result. Moreover, we show how to remove the effect of known introgression from an external population into either H_1, H_2 or H_3, and indirectly estimate the admixture rate using the D-statistic. In the results section it is shown through simulated and real data that this approach amplifies the test's sensitivity in detecting the presence of gene flow, hence it makes the method more reliable compared to the sampling approach or the methods based on calling procedures.

Standard D-statistic

The D-statistic is applied to formally test if the relationship between four populations H_1, H_2, H_3, H_4, represented in Figure 1 is fulfilled by the data. Population H_4 is assumed to be an outgroup and the correctness of the tree is stated as the absence of gene flow between either H_3 and H_2 or H_3 and H_1. In what follows, a statistical test based on the allele frequencies and the null hypothesis $\mathcal{H}_0$ that the four-population tree is correct, is developed.

Here, the j-th population consists of N_j individuals sequenced without error, with n_j^i observed bases at locus i from aligned reads. In order to keep the notation uncluttered, the treatment of the D-statistic is limited to a di-allelic model with alleles A and B, where B is the non-outgroup allele, but the extension to four alleles is straightforward. Only the M loci with at least a sequenced base from aligned reads in each population are considered, where M is allowed to grow to infinity. Each j-th population has frequency of the A allele x_j^i at locus i, with $j = 1, 2, 3, 4$, and $i = 1, \ldots, M$.

The idea behind the D-statistic is to characterize the differences within the pairs of populations (H_1, H_2) and (H_3, H_4) represented in the tree of Figure 1. Given a random allele drawn independently from each population, consider two specific combinations of alleles: ABBA and BABA. In the former pattern, populations H_1, H_4 share allele A and the non-outgroup allele is shared by H_2, H_3. In the latter combination allele A is shared by H_1, H_3, while H_2, H_4 share allele B.

In the model of four-population tree considered in Figure 1 we assume that each branch undergoes independently a genetic drift. Therefore the ABBA and BABA patterns, conditionally to the populations' frequencies, will be verified in rare occasions.

Consider the probabilities of ABBA and BABA patterns at locus i. Those can be calculated as the following expectations:

$$\mathbb{P}(ABBA_i) = \mathbb{E}[x_1^i x_4^i (1 - x_2^i)(1 - x_3^i) + (1 - x_1^i)(1 - x_4^i) x_2^i x_3^i] \tag{4}$$

$$\mathbb{P}(BABA_i) = \mathbb{E}[(1 - x_1^i) x_2^i (1 - x_3^i) x_4^i + x_1^i (1 - x_2^i) x_3^i (1 - x_4^i)]. \tag{5}$$

The formal statement of the null hypothesis passes through the two equations above. In fact the objective is to test if the A allele is shared equally between the pairs H_1, H_3 and H_1, H_2. The idea is to study when the difference between (4) and (5) is equal to zero, leading to the null hypothesis:

$$\mathcal{H}_0 : \mathbb{E}[(x_1^i - x_2^i)(x_3^i - x_4^i)] = 0 \ \text{ for } \ i = 1, \ldots, M.$$

Let $\hat{x}_j^i$ be the empirical allele frequency for population j at locus i. The D-statistic is defined as the normalized test statistic

$$D_M := \frac{X_{(M)}}{Y_{(M)}} = \frac{\sum_{i=1}^{M} (\hat{x}_1^i - \hat{x}_2^i)(\hat{x}_3^i - \hat{x}_4^i)}{\sum_{i=1}^{M} (\hat{x}_1^i + \hat{x}_2^i - 2x_1^i x_2^i)(\hat{x}_3^i + \hat{x}_4^i - 2x_3^i x_4^i)}, \tag{6}$$

with $X_{(M)}$ and $Y_{(M)}$ denoted as the numerator and the denominator of the D-statistic, respectively. A way of interpreting D_M is to see it as the difference between the probabilities of having an ABBA and BABA pattern of alleles over all loci, conditionally on observing only ABBA or BABA patterns.

In Manuscript 1 it is proven that as $M \to +\infty$, D_M converges in distribution to a standard normal variable under a specific set of conditions. This result makes it possible to use D_M as a test-statistic for a standard normal test in order to verify the null $\mathcal{H}_0$.

Extended D-statistic

In the extended D-statistic the issues arising in the current methods, intrinsically related to NGS data, are addressed. The improvements implemented in the extended D-statistics do not alter the unbiasedness of the frequency estimators used to calculate it.

To avoid calling procedures the sampling method is not applied in the extended D-statistic. Instead we use all aligned bases in multiple individuals per population, in order to estimate the population frequency at each locus. At each locus i, such an estimator is a weighted combination of the estimated allele frequency of each individual in the j-th population of interest, that is,

$$\hat{q}_j^i := \sum_{\ell=1}^{N_j} w_{j,\ell}^i \cdot \hat{x}_{j,\ell}^i,$$

where each weight $w_{j,\ell}^i$ is the linear coefficient of the ℓ-th individual within population j. The weights are determined in order to minimize the variance of $\hat{q}_j^i$ within respect to the weights. Further, they make it possible to consider datasets with a wide range of coverages within the same population. The obtained frequency estimator can be proven to be unbiased for the population frequency and has been first applied to reveal signatures of natural selection [58].

Since the aim of the D-statistic is the application in studies often involving ancient data, the type-specific error, that is the probability $\boldsymbol{e}(a, b)$ of observing base b when the true base is a, is calculated for every pair of bases and for each individual. The estimated type-specific errors of all individuals are organized into an error matrix $\boldsymbol{e}$ for the four-population tree, where each entry corresponds to the probability that a combination of four alleles is mistakenly observed instead of another combination. Applying the product of this matrix to the vector of observed combinations, it is possible to obtain true (error-corrected) combinations of alleles.

The rejection of the null hypothesis can arise when there is gene flow between an external population H_5 and one between H_1, H_2, H_3 and H_4. Let $p_{1:4}$ and p_{out} be the probabilities of allele patterns in the four-population tree where H_5 is removed and substitutes the population affected by introgression, respectively. If the admixture rate α is known, it is possible to calculate the probability of allele patterns p_{un} where the portion of introgressed genome from H_5 is removed:

$$p_{un} = \frac{1}{1-\alpha}(p_{1:4} - \alpha p_{out}).$$

If the admixture rate is unknown, it is possible to indirectly estimate it as the value of α for which p_{un} makes $E[D_M] = 0$. Note that the source of gene flow must be always known in order to determine the probabilities p_{out}, and that this model assumes recent admixture (ideally no drift after the admixture pulse).

Results and perspectives

The extended D-statistic has been tested on both simulated and real-data scenarios to study the effectiveness of the implemented improvements.

The use of all bases in multiple individuals per populations shows a higher sensitivity to introgression, because the estimated frequencies are less biased when using all available information. In our simulations, the power of the test based on the extended D-statistic with five individuals per populations is almost as high as performing the four-population test with the true genotypes at depth $2X$, and it greatly outperforms the power of the sampling approach. The test on a real-data scenario representing the admixture of southwestern Europeans into Native Americans [37] provides a more significant rejection with the extended D-statistic, when compared to the sampling approach. Moreover, the standard deviation of D is reduced by increasing the available individuals.

Using a simulated four-population tree in absence of admixture, where the type-specific errors affect populations H_1 and H_3, the scenario is rejected because the number of ABBA and BABA patterns is biased by the errors. This problem is solved using the error correction method implemented for the extended D-statistic. In the real-data scenario represented by the tree (((Saqqaq,Dorset)French)Chimpanzee), the ancient genomes of Saqqaq and Dorset - that are known having a common ancestral population [39] - are heavily affected by deamination and the scenario is rejected with high significance. Using the estimated type-specific errors for each population, we are able to restore acceptance of the configuration.

The extended D-statistic proves to be effective in simulated scenarios after removal of the bias in the number of ABBA and BABA patterns due to introgression from an external population. We successfully obtained the acceptance of a simulated scenario comparable to the tree (((Han Chinese,Dinka)Yoruban)Chimpanzee), with introgression from the Neandertal into the Han chinese population, representing the admixture between Neandertal and out-of-Africa populations [15, 38]. In the application with real data, it was possible to estimate almost the same admixture proportion with similar uncertainty compared to the one calculated in [38], in relation to the introgression into the Han Chinese population. Furthermore the correction for introgression seems not to be affected by drift in the simulated scenarios aforementioned, and is performed correctly with the time of admixture being 4000 generations in the past, due to a split that happened 8000 generations ago.

The extended D-statistic is therefore well-suited to detect gene-flow from low-depth sequencing data with high sensitivity. Such a property can result in the drawback of interpreting results in the wrong way when the underlying structure between populations is more complex, for example with multiple admixture pulses between H_3 and both H_1 and H_2, drift after the moment of admixture, or introgression from a distant ancestor of the population whose data is available. Note that the presence of structure between H_1 and H_2 does not influence the power of the test, since it changes only the sign of the numerator of D_M. One of the possible alternative scenarios in case of rejection has been explored when correcting for introgression from an external populations, but there are of course unnumbered complex scenarios that can occur.

An example of bias introduced by post-admixture drift and introgression from an ancestor of the available data (that again, introduces bias through drift), is given by the correction for introgression of the Mal'ta population into the Peruvian group for the tree (((Peruvian,Han Chinese)Central European)Yoruban), representing the fact that the Mal'ta population is strictly related to the modern Native Americans, but has no affinity to Eastern Asians [37] (see Figure 11B). More precisely, the relatedness with the Mal'ta population is due to admixture between its ancestor and the ancestral population of the Native Americans [16, 17]. Here the adjustment for external introgression reveals both higher admixture rate and uncertainty when compared to the results in [16, 17], where the effect of the ancestry and the post-admixture drift have been considered. In fact, the correction for introgression through the D-statistic assumes recent admixture with an instantaneous pulse, that is, ideally no drift before and after the admixture.

Another drawback of the extended D-statistic resides in the error correction method. In theory the application of the error correction works when applied locus-by-locus to each individual of interest, for which the type-specific errors have been previously estimated. This generates in practice two problems: the first consists in a great computational cost that makes the estimation of the D-statistic unacceptably slow, the second consists in over correcting the observed allele frequencies into negative values on loci not affected by error.
Solutions to the latter issue could be approached in different ways. An easy fix is to set a lower threshold of 0 when overcorrection happens, but this would probably generate a bias in the D-statistic. Another possibility is to check on which loci allele frequencies become negative after correction, and therefore avoid correcting on such loci. The approach that seems more reasonable and suitable for future development of the error correction is to study a weighted error matrix for each population. The weights should be related to the ones of the population frequency estimator for the extended D-statistic.

The proposals above do not solve the problem of computational speed. In the extended D-statistic the problem has been tackled by correcting the unobserved pattern frequencies over blocks of loci, and without considering the depth of each individual when building the error matrix. This means that the individuals with low depth might undergo an excessive error correction and bias the numerator of the D-statistic.

This effect will likely increase with growing estimates of type-specific errors and/or variability in the depth

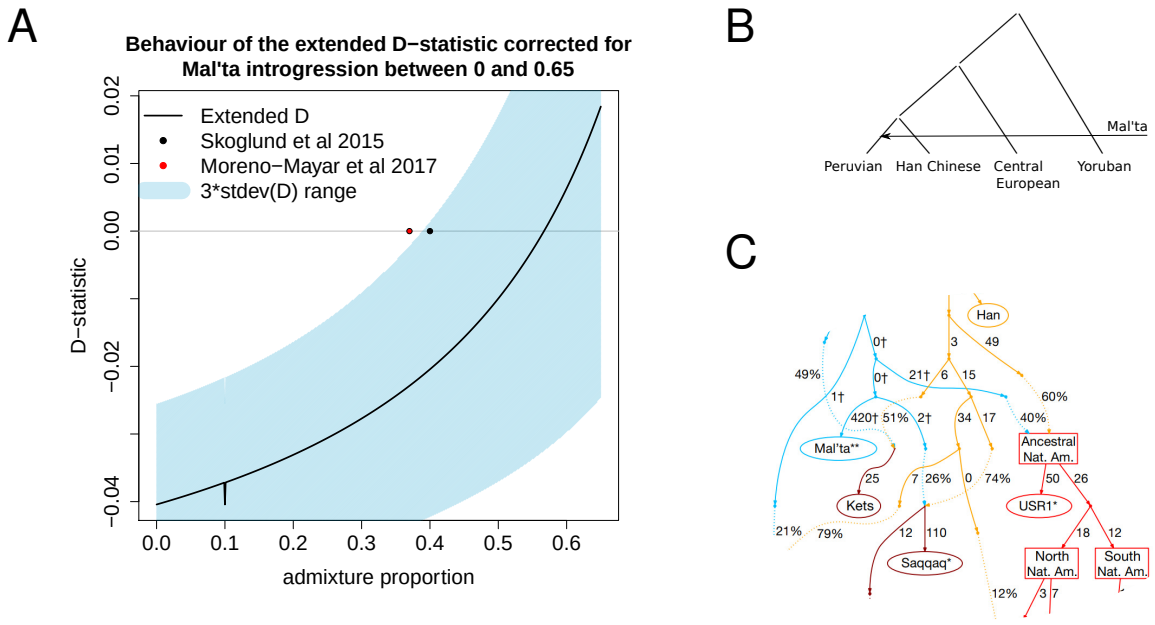


Figure 11: **Inference of Mal'ta gene flow into a Native American population.** (A) Inference of admixture rate from the Mal'ta population into the Peruvian (Native American), based on when $E[D] = 0$ (admixture proportion 0.56). The value 0.40 inferred in Skoglund's paper is in the uncertainty range of the plotted D-statistic, but not the proportion 0.37 from Moreno-Mayar's study. (B) Configuration for the ABBA-BABA test affected by external introgression of the Mal'ta population into Native Americans. (C) Detail of the configuration from [17], where the drift after admixture and the ancestry of the Mal'ta and Native American populations are taken into account.

of the different individuals within a group. Analogously as suggested before, a future development could be a weighting system for the type-specific errors based on the linear coefficients of each population frequency in the whole block of loci.

Theory for Gene Flow Inference in Model-based Methods

The evolution of populations is often characterized by many factors in its history, such as gene flow due to migrations, admixtures and splits. Such a complexity has always been a challenge for population geneticists. Many traditional analysis in population genetics have been based on statistics calculated from genetic data, e.g. heterozygosity, and then compared to their expectation under a specific setup of demography and mutations, allowing for parameter inference [59].

The mathematical development of the coalescent theory [60, 61] lead to an increasing focus on methods to infer populations' history and mutation rates from molecular data, e.g. with MCMC techniques using simulated genetic data [62] or likelihood-based methods [63, 64]. An example is the study of population size variations to deduce past genealogical events [62]. However, those methods are computationally intensive and not useful in cases of complex evolutionary histories.

However, the techniques mentioned above are considerably demanding in terms of computation complexity even when inference happens on a single locus [63]. Further, those techniques also rely on mutation rates and are hardly applicable when short time scales and low per-locus mutation rates are considered. A low mutation rate implies that there might not be enough mutations to trace coalescent events back in time.

Another genetic characteristic, the allele frequency, have instead become of increasing interest in studying genetic relationship between populations. Compared to mutations, the allele frequency is more informative of evolutionary changes in different populations. In fact, mutations contributes very little to variations in allele

frequencies over relatively short periods of time. But allele frequencies change remarkably as a consequence of genetic drift and other factors (e.g. selection, population structure, gene flow and admixture) [10].

With the advent of high-throughput genetic data, such as NGS data [22, 24, 29], a large amount of data has become available to scientists. Computationally demanding methods to study populations' history have therefore become ineluctably unusable. In recent years, an explicit model illustrating past gene flow between populations, the admixture graph [13, 14, 65, 66], have been proposed as a generalization of the phylogenetic tree [67, 68].

The phylogenetic tree is an attempt to model genetic relationships between populations through a graph that admits only the split of a node that gives rise to two descendant populations. This model has generated many applications. For example the Neighbor Joining Tree [69] method infers a tree using a measure of genetic distance as clustering metric (see Figure 12). This method has become popular already with microsatellite data, and it is still used in modern tools to infer a tree, to which mixtures are added in other steps [21].

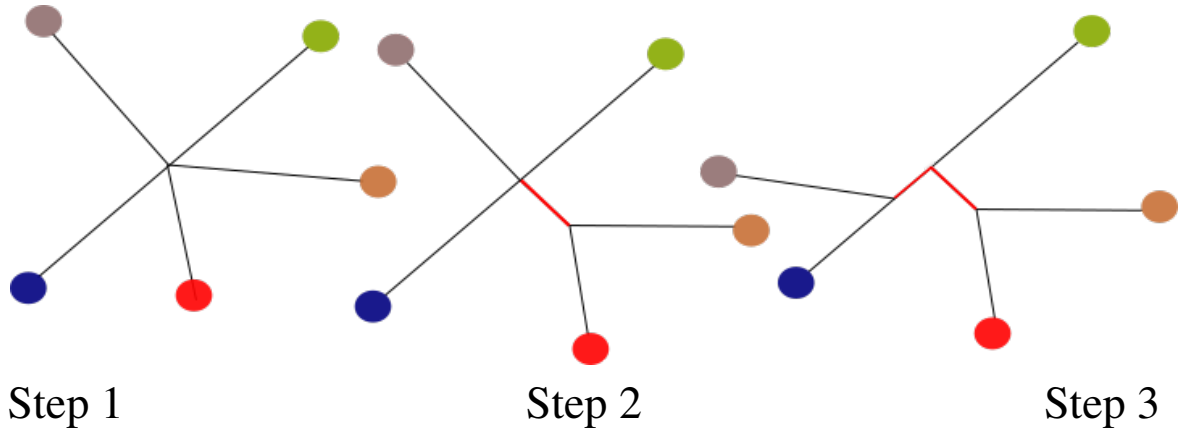


Figure 12: **Neighbor Joining Tree method: toy example.** The Neighbor Joining Tree method starts from a star-shaped unrooted tree and pulls out a split that is optimal in term of a measure of genetic distance calculated from the available data. The added branches are highlighted in red

An admixture graph admits a more elaborated evolutionary history. Here the model's formulation includes gene flow between populations, so that populations can merge and generate new lineages. At each locus, the alleles of a sequence have frequency given by a linear combination of the allele frequencies of the admixing populations [13, 14]. The linear coefficients of this combination are called admixture rates.

Some simple examples of admixture graphs are the ones used for the four-populations test in Figure 10B and Figure 11B, where only one admixture and a limited amount of populations is involved. The admixture graph in Figure 2 shows many populations and four different admixture events.

A tool that works on admixture graphs is `qpgraph` [70], where the authors use a heuristic method to exclude unlikely edges, by building specific subgraphs denoted as qp-graphs. More recent methods based on admixture graphs use moment statistics, called F-statistics, between populations.

The F-statistics, namely F_2, F_3 and F_4, are based on allele frequencies. Their formulation allows for a greater computational efficiency when compared to earlier studies based on computationally intensive likelihood optimizations [65, 66]. The F-statistics are a particularly successful tool in population genetics not only because of their applicability on genome-scale data, e.g. NGS data, but also because of their properties. In fact, the F_2-statistic, defined between two nodes, is interpretable in different ways and allow for a deeper understanding and easier applicability to computational methods.

For example, F_2 can be expressed in terms of variances on each independent lineage, and F_2s are additive on adjacent lineages in absence of admixture [13, 21, 71]. Moreover, the F_3- and F_4-statistic can be written as combination of F_2-statistics [12], and are the base of testing for admixtures in the three-population test and the ABBA-BABA test [12, 13].

Amongst methods for admixture graphs based on the F-statistics there are `AdmixTools` [13], `TreeMix` [20] and `MixMapper` [21]. The first is used to infer admixture rates from a hypothesized graph. The latters essentially build a tree, and then add admixtures in steps evaluating the fitness to the data. All those methods use the fact that the additivity of F_2 on lineages is not possible in case of admixtures [13, 21], but such sum involves further terms and admixture proportions. This allows for inference on the admixture proportions based on the topology of the graph, by equating the F_2-statistics calculated from data to their theoretical value.

In this thesis a formalization of the admixture graphs and their properties are analyzed. The definition of a stochastic structure on the graphs allows to study in depth the F-statistics and to find useful results, amongst which the most important are a general formula to express the F_2-statistics, and their properties in terms of linear independence. The results of this theory can be related to the population genetics framework used in recent computational tools and provide a solid background to future studies of the admixture graphs.

Admixture graphs and stochastic structure

An admixture graph is formalized as an acyclic directed graph with multiple roots. The use of multiple roots describes a situation in which there is no hypothesis on the relation between the corresponding populations. Therefore edges between each pair of roots are undirected to avoid making assumptions on which population is ancestor of the other. The edges of such graph have labels that correspond to admixture proportions. Each edge can be seen as a lineage.

To describe the genetic relationship between two populations in terms of their common ancestry, admixture paths between them are defined. Each path is the composition of the connection between each of the two populations and one of their common ancestors (there can be more than one ancestor because of admixtures). Overlapping lineages are not considered, because those are verified when a common ancestor has been already reached. Each path is assigned a label. This is the product of edges' labels encountered by such path.

Figure 13 shows an example of admixture graph and the paths between nodes 4 and 5. Paths have a direction identified by the starting node. Consider for example the paths starting in 4 and ending in 5 in Figure 13. One can simply write them as sequence of connected nodes. For example the green path can be written as $(4, 3, 5)$.

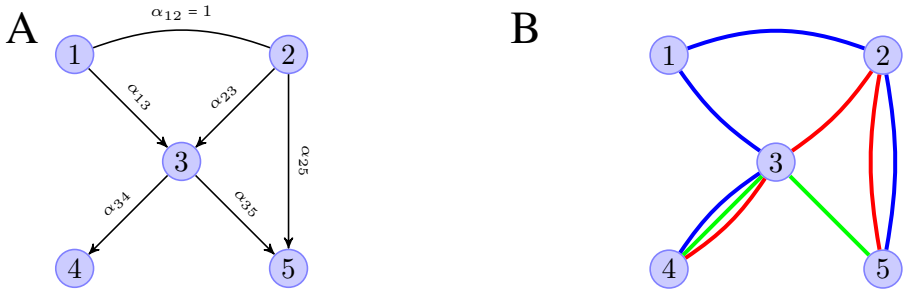


Figure 13: **Example of admixture graph and admixture paths.** (A) Admixture graph where nodes $1, 2$ are roots, 3 is an admixed node (population) and $4, 5$ are leaves that have only one parent population. (B) Paths between populations 4 and 5 traced in different colours. The edges of the graph are not represented to avoid confusion; note that while going backward from either node 4 or 5 to a common ancestor, all the edges of the graph are met from the end to the beginning of the arrow. This corresponds to the backward point of view when considering lineages in the Wright-Fisher model.

A stochastic structure for admixture graphs is formalized using assumption that can be matched with properties of the allele frequencies in population genetics. Each node i of the admixture graph has attached a random variable V_i, modeling for example the allele frequencies. Moreover, one would like to take into account changes in allele frequencies along a lineage. For each node i and each branch going to a population j, we introduce an additional variable C_{ij}. In terms of population genetics, C_{ij} describes how the frequency of V_i has changed at the time in which population j is generated.

The random variables are characterized by further assumptions to fit into the population genetics framework.

For example, any variable V_j whose node has parents is given by $\sum_{i \in par(j)} \alpha_{ij} C_{ij}$, where $par(j)$ are the parents of node j in the graph. This corresponds to the model of admixture pulse adopted, for example, in the four-population test, and commonly applied in population genetics, where an admixed population inherits a fraction of alleles from each ancestor [13, 14]. Another characteristic of methods based on allele frequencies is that mutation are often not considered since they have a negligible effect compared to other factors affecting the frequencies. In accordance with the property of frequency in (2) for the Wright-Fisher model without mutations, it is here assumed that $E(C_{ij}|V_i) = V_i$ for each node i.

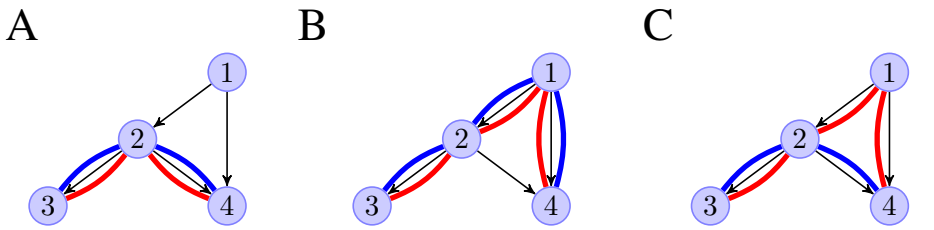
Results and perspectives

Using the stochastic structure introduced above, it is possible to describe the drift between two nodes. In terms of allele frequencies, this can be seen as the difference in frequencies between two populations. Intuitively, each frequency depends on a combination of admixture rates and parents according to the model of admixture pulse.

Therefore the drift between two nodes should depend on labels and nodes found recursively in paths from the two nodes to some common parents. It is shown that a drift can be characterized using admixture paths and labels. Each path contributes to the drift with an additive term proportional to the labels on the edges of the path. This matches the intuition based on the model for admixtures adopted in population genetics.

The drift proves to be fundamental in analyzing the F-statistics. Amongst those, we consider mainly the F_2-statistic, because F_3 and F_4 can be written as linear combination of F_2s [12]. The F_2-statistic is defined between two nodes i, j by $E[(V_j - V_i)^2]$; its objective in population genetics is to measure how different two populations are in terms of allele frequencies [12].

In manuscript 2 it is proven, using the properties of drifts, that the F_2-statistic between two nodes can be decomposed using admixture paths. One needs to consider the edges where at least a pair of paths between the two nodes overlaps, and the labels involved in all those pairs of paths. Each edge $k \to \ell$ of the graph contributes with an additive term to the F_2-statistic with the product of the squared expectation $E\big(d^2_{k\ell}\big) := E\big((C_{k\ell} - V_k)^2\big)$, properly scaled by paths' labels. In other words, the F_2-statistic highlights shared changes in allele frequencies on the possible lineages to common ancestors (see Figure 14).

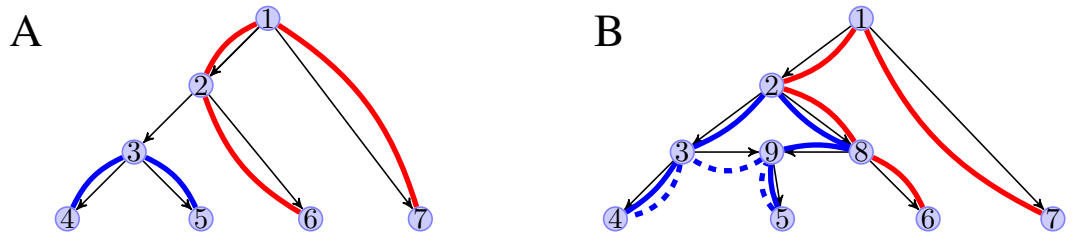


(A) $\alpha^2_{24} E(d^2_{23} + d^2_{24})$ (B) $\alpha^2_{14} E(d^2_{23} + d^2_{12} + d^2_{14})$ (C) $2\alpha_{14}\alpha_{24} E(d^2_{23})$

Figure 14: **Decomposition of the F_2-statistic.** Possible pairs of paths between nodes 3 and 4 used to interpret the decomposition of the statistic $F_2(3, 4)$. In (A) and (B) the two paths overlap on all edges, while in (C) they overlap only between nodes 2 and 3. Below the graphs are written the additive terms that contributes to the F_2-statistic in each of the three pairs of paths.

The decomposition matches the graphical method proposed in [13, 14] on admixture graphs, where the authors propose it to take into account that lineages are not independent in presence of admixture. In absence of admixtures, the F_2 is the sum of F_2-statistics between adjacent nodes on the unique path between two nodes.

In a similar way one can interpret the F_3- and F_4-statistic. The former is interpreted as the amount of shared frequency change between pairs of paths starting in the same node and ending in two different nodes. The latter highlights the amount of shared drift between two different pairs of nodes. The F_4-statistic is used in the first application of the ABBA-BABA test [12], and is the numerator of the D-statistic. The idea behind the ABBA-BABA test when using F_4 is that, in absence of admixture as in Figure 15A, the paths between $4, 5$ and $6, 7$ do not have shared drift, and therefore the F_4-statistic has value zero (see Figure 15).



(A) $F_4(4,5;6,7) = 0$ $F_4(4,5;6,7) = -\alpha_{89}E(d_{28}^2)$

Figure 15: **Relationship between F_4 and the ABBA-BABA test.** (A) Possible paths on the four-population tree used for the ABBA-BABA test in absence of admixture. Here the paths between nodes $4,5$ (blue) and $6,7$ (red) do not overlap and therefore the F_4-statistic between the two pairs is equal to zero. (B) Here the two possible paths between $4,5$ (blue) overlap in one case (solid blue line) with the red path between $6,7$ on edge $2 \to 8$. Therefore $F_4(4,5;6,7)$ is not equal to zero in this case.

Using the canonical decomposition for the F_2-statistic and the fact that the F_2 between adjacent nodes can be expressed as a difference of variances [71], one could in some cases express F_2 as a sum of variances along paths. An application that is possible to explore is the existence of a variance decomposition. In some specific cases, such as in [72], a variance decomposition has been studied for undirected gaussian graphical models. Gaussian variables are often used to describe frequencies in the F-statistics-based computational tools for admixture graphs [20, 21].

A natural question that arises when studying the F-statistic is the possibility of defining F_k-statistics with $k > 4$. For example, consider $k = 5$. In a definition of F-statistic involving five nodes where drifts are multiplied, one would expect the product of at least three drifts, so that all nodes are considered. This would make it impossible to express an F_5-statistic as combination of F_2-statistics (as it happens for F_3 and F_4). In fact F_2 has terms at most quadratic in the partial drifts, while F_5 would contain terms cubic in the partial drifts. However, it is a possible development to research more into this aspect of the topic to understand if it possible to find further F-statistics - or to prove for example that this is not possible.

The F_2-statistic can also be interpreted, in some cases, as a metric between two nodes. This property is strictly related to the F_3-statistic and to the topology of the admixture graph. A result giving a condition for verifying if the F_2-statistic is a metric, based on admixture paths, has been deduced and discussed.

The conditions for verifying that F_2 is a metric between a pair of nodes i, j are relatively complicated, but possible to implement. In few words, it is necessary to find in which pairs of paths between i and j there are coincident nodes in opposite order, so that F_2 does not fulfill the triangular inequality for metrics. The problem is therefore redirected to listing all paths between two nodes.

A further development related to the F_2-statistic as a metric is to study if there is a relationship between the F_2-statistic and the split decomposition [73, 74]. Here, a metric is decomposed as the sum of weighted metrics on elementary subgraphs called splits and a non-metric residual. This proves to be extremely complicated even on elementary examples of admixture graphs, primarily due to the difficulty of understanding when the decomposition of the F_2-statistic allows to define splits.

An important part of the theory of admixture graphs here proposed is the analysis of the linear independence of the F_2-statistic in a set of nodes. In fact, this property is fundamental when a linear system of decomposed F_2-statistics is considered in applications for admixture graphs [13, 20, 21].

The study of the linear independence involves all the elements introduced in the admixture graph theory, such as the graph topology, the admixture paths, the decomposition of the F_2-statistics and their additivity property.

Firstly, it is proven the additivity is verified only under some specific conditions. Already in [13, 14] it has been pointed out that in presence of admixtures, the additivity does not hold. In [13] the renowned graphical method to determine the F_2-decomposition has been proposed, and an insight to its proof has been given in [21]. Here the graphical method is a consequence of the F_2-decomposition along admixture paths.

Lastly, a theorem giving conditions under which the F_2-statistics are linearly independent is proven. This result sets a relationship between admixture rates, additivity property, decomposition of the F_2s and admixture paths. Specifically, the linear independence is explored using the system of equations of decomposed F_2s from pairs of populations to formulate another system of equations. Each equation of this system has terms based shared edges of the decomposed F_2s. When some shared edges appear on the same decompositions with the same coefficients, the linear independence might be broken.

The theorem for the linear independence of the F_2-statistics holds only in cases where there are at most two roots in the admixture graph of interest. It is still to be proved if there are conditions under which the F_2-statistics on a subset of nodes of a graph with an arbitrary number of roots fulfill the linear independence.

Inference of Ploidy Numbers from NGS Data

The ploidy number (or ploidy) is the number of sets of chromosomes that are found in a cell. If the chromosomes are grouped one by one, then an organism is said to be haploid. Chromosomes that go in pairs are found in diploid organisms. Organisms with higher number of chromosome copies grouped together (triploid, tetraploid, pentaploid, etc. organisms) are said to be polyploidy.

Humans are known to be diploid, as it is often the case for animals, but other species are often characterized by a different ploidy. Especially plants and fungi are known to have many polyploidy species, eventually with different ploidy within different chromosome sets in the same individual [75, 76]. The polyploidy state is often the consequence of hybridization or whole genome duplications (see Figure 16) and is mostly observed in plants and fungi [77]. The genus of the *Spartina*, a perennial, has split into triploid, hexaploid and dodecaploid species [78, 79].

The changes in ploidy are considered to be playing an essential role in the evolution of plants in natural populations [76] and is probably the most important factor concurring in speciation of plants [80]. Moreover, polyploidy can be an advantage for adapting to environmental factors when it causes alterations of the morphology and phenology of the organisms [81]. Those alterations can happen even as fast as in one generation [82]. Polyploidy events have been detected in the ancestry of some types of crops and tomatoes [75], in lineages of the maize [83, 84], in the common ancestry of cotton types [85, 86] and soybeans [75], and in fungi [87, 88].

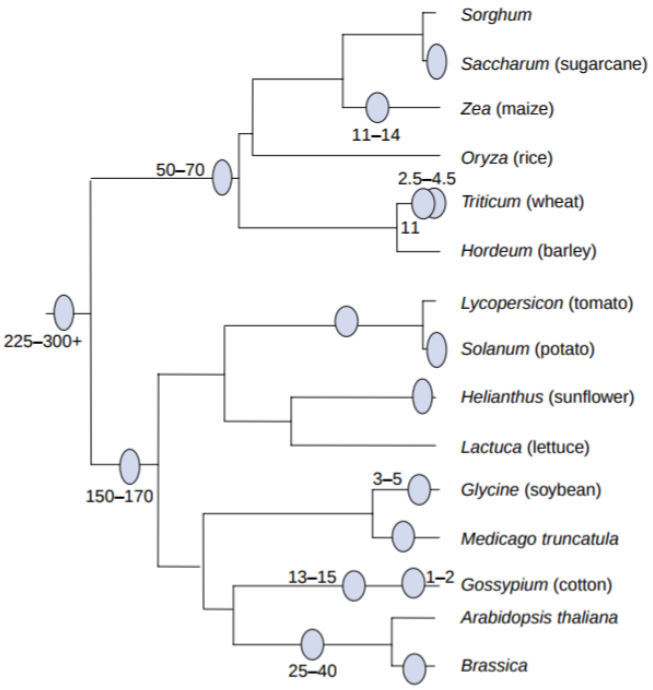


Figure 16: **Inferred times of whole-genome duplications in the past evolutionary history of angiosperms.** Tree representing the inferred times (in million of years in the past) at which whole-genome duplications creating new polyploidy states happened for the evolutionary history of angiosperm [86]. Source: [75].

An experimental method to detect ploidy numbers in a genome is the flow cytometry procedure [89]. Flow cytometry is a high-throughput technique with which scientists are able to obtain a quantification of optical properties, such as fluorescence, from particles floating in a special fluid. When flow cytometry is applied to a cell, it is possible to determine very precisely the amount of genetic material in the nucleus, and estimate the ploidy. Modern flow cytometry instruments are very sensible and reliable, but their cost is high [90, 91] and not feasible if the only focus of an analysis is on the detection of ploidies.

The advances in high-throughput sequencing techniques of the recent years, such as Next Generation Sequencing (NGS) [23, 29], have rapidly resulted in a vast amount of cost-effective high-throughput data available for a wide range of genetic studies. The available NGS protocols [23, 24, 29] essentially result into an output that consists of short reads whose length is in the order of hundreds of bases, that are further aligned to a reference genome or denovo assembled in scaffolds. It is often the norm that studies based on NGS data rely on low-depth sequencing ($< 10X$) because of cost-efficiency and/or degradation of the samples; moreover NGS data is usually affected by a relatively high sequencing error, if compared for example to the traditional Sanger sequencing [32, 33].

This results in potentially unreliable estimates of allele frequencies in the data, and consequently a bad estimation of genotypes. Moreover, note that allele frequencies themselves can be misleading in revealing ploidy numbers through genotypes. For example, consider the simple setup of having sequenced a set of di-allelic chromosomes that have same ploidy, so that the sequencing depth is not necessarily informative. Assume tha,t at a locus, allele C has been observed with a proportion of $2/3$, and allele T with a proportion $1/3$. This might point to genotype CCT and ploidy equal to three, but also to genotype $CCCCTT$ and ploidy six, and so on with the ploidies multiple of three (see Figure 17).

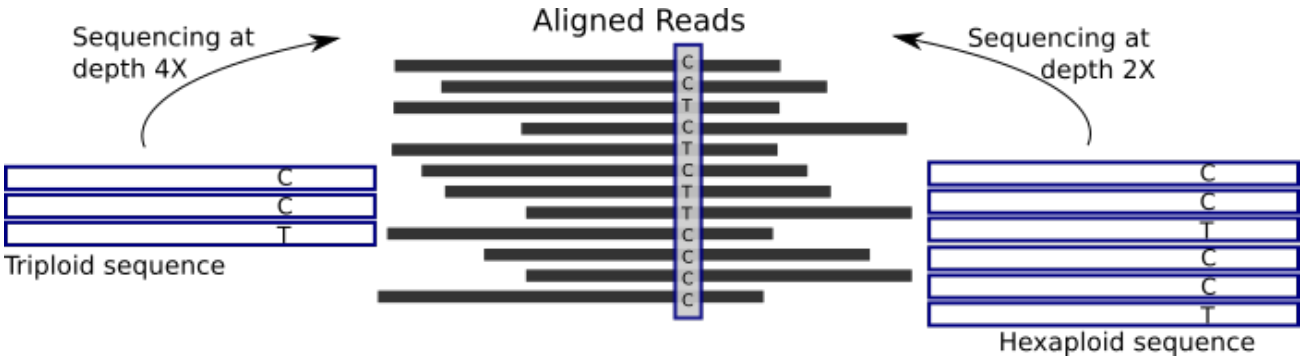


Figure 17: **Misleading ploidy inference from allele frequencies.** Representation of the case in which triploid and hexaploid sequence have the same proportions of alleles at a locus. It is not possible to say if the aligned reads are due to sequencing of the triploid individual at depth 4X, or of the hexaploid individual at depth 2X. The word depth is referred to the haploid depth, that is, the number of reads expected for each copy of the chromosome.

Many of the current methods for the estimation of ploidy numbers in NGS data are based on loci's depth and allele frequencies. For example `conPade` [92] detects the ploidy of a given contig/scaffold using allele frequencies. The tool `ploidyNGS` [93] estimates allele frequencies and provides a visualization tool through which ploidy can be estimated visually. The visual approach is very commonly used to empirically estimate the ploidy [94]. `AbsCN-seq` [95] combines the information on depth and allele counts to estimate, amongst other parameters related to tumor-specific applications, the ploidies from NGS data. Analogous information are applied to cancer cells' data with in the package `sequenza` [96].

Changes in ploidy numbers can also be detected because of Copy Number Variations (CNV) When sequenced reads are aligned, the ones from the copied segments will be mapped to the same region of the reference genome. This results in a multiplying factor for the sequencing depth, that is therefore detected as a change in ploidy. Studies have reported that CNVs are present in humans [97, 98] and can be connected to the possibility of developing diseases [99, 100].

We propose a method called `hiddenMarkovPloidy`, dedicated to infer ploidy numbers from NGS data. The method builds a Hidden Markov Model [101, 102] with a double set of observations that consists of sequencing depths and observed reads. The formers are used to detect changes in ploidy, while the latters are based on the so-called genotype likelihoods [35], and contribute in assigning each hidden state to its corresponding ploidy.

Results and perspectives

Preliminary results show that the implemented model is able to recognize ploidy numbers from one to five at low-depth (2X). However, at lower depth (0.5X) the ploidy number five is almost completely missed. In fact, the drawback of this model is that many individuals are needed to estimate minor allele frequencies. If there are not enough individuals, then the estimates might be biased, and consequently genotype probabilities would be of little use for high ploidies. However, it is possible to calculate the expected frequency over all sites when only one individual is available.

Moreover, high ploidies need the inference of the correct genotypes. If depth is too low, there will not be enough reads to estimate genotypes on poliploidies. Those depth scenarios are quite extreme and not really expected when using real data.

The performances on a strain of Bd fungi shows promising results for applications on real data. It is possible to match the ploidy numbers that can be deduced by looking at the sequencing depth.

Further directions in the development of the model are being taken. An idea is to apply the Hidden Markov Model to the detection of CNVs. The idea behind it is to proceed in two steps. Firstly, hidden states are detected according only to the sequencing depth. Secondly, if the change of state is not followed by a new ploidy that maximizes the probability of sequenced data, then the state is marked as CNV.

Moreover, the Hidden Markov Model uses only the depth of a single individual, even though genotypes likelihoods come from all genomes. The EM algorithm for multiple observations, assuming same ploidy numbers on each window, and same haploid depth, is under development. The Hidden Markov Model could be applied only on a subset of individuals with the same ploidy and haploid depth to develop a test for aneuploidy based on the likelihood of the model for other individuals.

The window size used in this application is predefined by the user. Therefore, a window could overlap a change of ploidy number. A further improvement would be developing automatic windows that do not need to be predefined in input, but follow a criteria to perform unsupervised selection of the window size.

It is planned to use the data from more than 200 fungi to detect their ploidy numbers. The focus is on the chytrid fungus *Batrachochytrium dendrobatidis* (Bd), whose spreading has become worrying, since it causes the losses of amphybians worldwide. By performing a mapping of the ploidy numbers at different lineages, it might be possible to understand the genetic mechanisms at the bases of the worrisome spreading trend of the Bd.

Manuscript 1

Powerful Inference with the D-statistic on Low-Coverage Whole-Genome Data

Samuele Soraggi, Carsten Wiuf, Anders Albrechtsen

Status: Published in G3: Genes, Genomes, Genetics. Volume 8, Issue 2, February 2018.

Contribution

This paper contributes with a new implementation of the D-statistic and a thorough statistical analysis that motivates its application. The proposed implementation is especially aimed at low-depth high-throughput data. In fact, it is possible to use multiple genomes with different sequencing depths for each population. Moreover, issues related to ancient DNA are solved thanks to the possibility of applying type-specific error correction. Finally, we avoid calling procedures by using all reads available at each locus, instead of applying a sampling approach.

As a results, the power of the newly implemented test is as high as the power of the D-statistic for known genotype when the depth is 2X. The error correction performs well when ancient data with high error rates is involved in the analysis, and it is possible to estimate admixture rates within reasonable intervals of uncertainty.

The D-statistic is implemented in `C++` for the tool `ANGSD` [35] and illustrated at the address `http://popgen.dk/angsd/index.php/Abbababa2`.

Future perspectives

Many possible developments for the implemented method are explorable. First of all there is the necessity of building a better model to correct for type-specific errors. Here we correct for errors on blocks of loci, without weighting the correction factors by the sequencing depth of the individuals. There is therefore need for a model that is still computationally fast, but considers the sequencing depth of each genome involved in the estimation of error rates.

Another possible extension of this method is to follow the idea of [103], where allele combinations are used to detect the polarization of gene flow, considering that a fifth population must be available.

Further, one could look into developing a way to calculate the F_2- and F_3- statistic with a similar approach, that is, by including multiple individuals and considering the allele counts at each locus. Another possible framework could be determining the F_2-, F_3 and D-statistic by using the genotype likelihoods, so that the uncertainty of each read could be taken into account.

The D-statistic is often applied on many combinations of four populations to detect gene flows. Using this results as a starting point to roughly build the past genetic interactions of those populations, the reticulate of gene flows is inferred or tested with tools working on admixture graphs and considering the effect of drift. However, it could be interesting to implement an ABBA-BABA test that considers the effect of drift and can aid in better estimates of admixture proportions and more precise values of the D-statistic.

G3: Genes|Genomes|Genetics Early Online, published on December 1, 2017 as doi:10.1534/g3.117.300192

INVESTIGATIONS

G3 Genes | Genomes | Genetics

Powerful Inference with the D-statistic on Low-Coverage Whole-Genome Data

Samuele Soraggi[*,1], **Carsten Wiuf**[†] **and Anders Albrechtsen**[‡]

[*]Department of Mathematical Sciences, Faculty of Science, University of Copenhagen, Copenhagen, Denmark, [†]Department of Mathematical Sciences, Faculty of Science, University of Copenhagen, Copenhagen, Denmark, [‡]Center for Bioinformatics, Faculty of Science, University of Copenhagen, Copenhagen, Denmark

ABSTRACT The detection of ancient gene flow between human populations is an important issue in population genetics. A common tool for detecting ancient admixture events is the D-statistic. The D-statistic is based on the hypothesis of a genetic relationship that involves four populations, whose correctness is assessed by evaluating specific coincidences of alleles between the groups.

When working with high throughput sequencing data calling genotypes accurately is not always possible, therefore the D-statistic currently samples a single base from the reads of one individual per population. This implies ignoring much of the information in the data, an issue especially striking in the case of ancient genomes.

We provide a significant improvement to overcome the problems of the D-statistic by considering all reads from multiple individuals in each population. We also apply type-specific error correction to combat the problems of sequencing errors and show a way to correct for introgression from an external population that is not part of the supposed genetic relationship, and how this leads to an estimate of the admixture rate.

We prove that the D-statistic is approximated by a standard normal. Furthermore we show that our method outperforms the traditional D-statistic in detecting admixtures. The power gain is most pronounced for low/medium sequencing depth (1-10X) and performances are as good as with perfectly called genotypes at a sequencing depth of 2X. We show the reliability of error correction on scenarios with simulated errors and ancient data, and correct for introgression in known scenarios to estimate the admixture rates.

KEYWORDS
Admixture
Gene flow
Introgression
D-statistic
ABBA-BABA test
Tree test
Four-population test
ANGSD
NGS data
Low depth

INTRODUCTION

An important part in the understanding of a population's history and its genetic variability is past contacts with other populations. Such contacts could result in gene flow and admixture between populations and leave traces of a population's history in genomic data. In fact, the study of gene flow between populations has been the basis to uncover demographic histories of many species, including human and archaic human populations (Patterson *et al.* 2012; Raghavan *et al.* 2013; Green *et al.* 2010; Reich *et al.* 2009; Wall *et al.* 2013; Raghavan *et al.* 2015; Rasmussen *et al.* 2010, 2014; Reich *et al.* 2010, 2011; Lalueza-Fox and Gilbert 2011; Skoglund *et al.* 2015).

The study of the history of human populations using both modern and ancient human genomes has become increasingly topical with the recent availability of new high-throughput sequencing technologies (Stoneking and Krause 2011), such as Next Generation Sequencing (NGS) technologies (Black *et al.* 2015). These technologies have made it possible to obtain massive quantities of sequenced DNA data even from ancient individuals, such as an Anzick-Clovis individual from the Late Pleistocene (Rasmussen *et al.* 2014), a Neandertal individual (Green *et al.* 2010) and a Paleoamerican individual (Chatters 2000).

There are many different methods for inferring and analyzing admixture events using genome-scale data. Popular methods such

Manuscript compiled: Friday 1st December, 2017
[1]Corresponding author: samuele@math.ku.dk

as STRUCTURE (Pritchard *et al.* 2000) and ADMIXTURE (Alexander *et al.* 2009) estimate how much a sampled individual belongs to K clusters that often can be interpreted as the individual's admixture proportion to the K populations. However, these approaches are not appropriate to detect ancient gene flow and do not work well with a limited number of individuals per population.

A recent alternative to the above methods is the D-statistic. The D-statistic is based on the di-allelic patterns of alleles between four groups of individuals, and provides a way to test the correctness of a hypothetical genetic relationship between the four groups (see Figure 1). A variant of the D-statistic (called the F_4-statistic) was first used in (Reich *et al.* 2009) to identify that subgroups of the Indian Cline group are related to external populations in term of gene flow. Also the amount of gene flow might be estimated using the F_4-statistic (Wall *et al.* 2013).

In the pivotal study (Green *et al.* 2010) the D-statistic was used to show that 3 non-African individuals are more genetically similar to the Neandertal sequence than African San and Yoruban individuals are. Moreover, it has been shown that the Eastern Asian populations have a higher amount of Neandertal shared genetic material (Wall *et al.* 2013).
Using the D-statistic on many Old World and Native Americans it has been suggested gene flow into some Native American populations, such as evidence of admixture from Australasian populations into New World Populations (Raghavan *et al.* 2015; Skoglund *et al.* 2015).
In another study the affinity between the Anzick genome and the Native Americans genome was analyzed with the D-statistic to compare different hypotheses regarding their ancestry (Rasmussen *et al.* 2014). Using the D-statistic, it has been reported that the remains of an individual from the Mal'ta population in south-central Siberia have contributed to the gene pool of modern-day Native Americans, with no close affinity to east Asians (Raghavan *et al.* 2013).

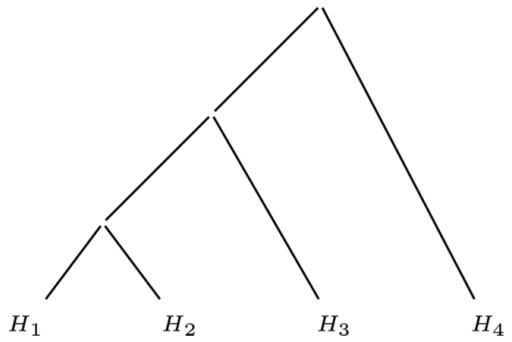


Figure 1 Tree topology for the D-statistic. Hypothesis of genetic relationship between four populations H_1, H_2, H_3, H_4.

The first use of the D-statistic was based on a sampling approach that allowed to perform the test without the need to call SNPs or genotypes (Green *et al.* 2010). This approach is still widely used, and amongst the available computational tools implementing this approach is the doAbbababa program of ANGSD (Nielsen *et al.* 2011) (supporting low depth NGS data) or the fourpop program of TreeMix (Pickrell and Pritchard 2012) (supporting di-allelic genotype data and microsatellite data). The program qpDstat of ADMIXTOOLS (Patterson *et al.* 2012) computes the D-statistic from populations with multiple individuals from di-allelic genotype data. The program doAbbababa relies on sampling one base from every locus, using the sequenced reads to define the sampling probabilities.
The D-statistic is often applied to scenarios involving ancient individuals, that are commonly affected by deamination, i.e. the natural degradation of DNA after death of the organism that leads to there being few molecules remaining in ancient specimens and often results in a low sequencing depth. Furthermore, deamination can cause high frequency of specific transitions of the bases, low quality of the SNPs and very low depth of the data. The current methods for the D-statistic can be very ineffective and unreliable when applied to ancient data, since both sampling and genotype calling procedures are subject to high uncertainty.

The focus of this paper is to address the problems stated above. We propose a D-statistic - implemented in the program doAbbababa2 of ANGSD - that supports low depth NGS data and is calculated using all reads of the genomes, and therefore allows for the use of more than one individual per group. We prove that the improved D-statistic is approximated by a standard normal distribution, and using both simulated and real data we show how this approach greatly increases the sensitivity of gene-flow detection and thus improves the reliability of the method, in comparison to sampling a single read. We also illustrate that it is possible to correct for type-specific error rates in the data, so that the reads used to calculate the D-statistic will not bias the result due to type-specific errors. Moreover, our improved D-statistic can remove the effect of known introgression from an external population into either H_1, H_2 or H_3, and indirectly estimates the admixture rate.

MATERIALS AND METHODS

This section introduces the traditional D-statistic and the theory that leads to its approximation as a normal distribution. Thereafter we explain how to extend the D-statistic to use multiple individuals per population, without genotype calling and still preserving the same approximation property of the D-statistic. Lastly, we will show how to deal with type-specific errors and introgression from a population external to the tree topology.

Standard D-statistic

The objective of the D-statistic is to assess whether the tree of Figure 1 that relates four present-day populations H_1, H_2, H_3, H_4, is correct. When H_4 is an outgroup, the correctness of the tree corresponds to the absence of gene-flow between H_3 and either H_2 or H_1. This objective is achieved by developing a statistical test based on the allele frequencies and a null hypothesis $\mathcal{H}_0$ saying that the tree is correct and without gene flow. We limit the explanation to a di-allelic model with alleles A and B to keep the notation uncluttered; the extension to a 4-allelic model is fairly straightforward. We do not make assumption on which allele is derived, but we assume that B is the non-outgroup allele. Population H_4 is an outgroup, that splits off at the root of the tree from the other branches. For each population H_j, $j = 1, 2, 3, 4$, in the tree, we consider the related allele frequencies x_j.

For each population H_j, the observed data consists of a certain number of individuals sequenced without error. At every locus i there are n_j^i sequenced bases observed from aligned reads. We consider only the M loci for which there is at least one

sequenced base from aligned reads in all four groups. Moreover, in this theoretical treatment we allow the number M of loci to grow to infinity. Assume that at a locus i the allele frequencies in the four groups of individuals $\boldsymbol{x}_i := (x_1^i, x_2^i, x_3^i, x_4^i)$ and let $\hat{\boldsymbol{x}}_i := (\hat{x}_1^i, \hat{x}_2^i, \hat{x}_3^i, \hat{x}_4^i)$ be an unbiased estimator of $\boldsymbol{x}_i$, such as the relative frequencies of the allele A in each population.

The D-statistic focuses on di-allelic sites where the differences are observed within the pairs (H_1, H_2) and (H_3, H_4). Consider a random allele drawn from each of the four groups of genomes and the resulting combination of the four alleles. We are interested in two patterns:

- ABBA, meaning that we have the same allele in populations H_1 and H_4 and another allele from the individuals in populations H_2 and H_3;
- BABA, where an allele is shared by individuals in populations H_1 and H_3 and the other allele by individuals in populations H_2 and H_4.

The tree of Figure 1 is subject to independent genetic drifts of the allele frequencies along each of its branches. Consequently the probabilities of ABBA and BABA patterns conditionally to population frequencies would rarely be same. Therefore it is interesting to focus on their expected values with respect to the frequency distribution:

$$\mathbb{P}(ABBA_i) = \mathbb{E}[x_1^i x_4^i (1-x_2^i)(1-x_3^i) + (1-x_1^i)(1-x_4^i)x_2^i x_3^i] \quad (1)$$

$$\mathbb{P}(BABA_i) = \mathbb{E}[(1-x_1^i)x_2^i(1-x_3^i)x_4^i + x_1^i(1-x_2^i)x_3^i(1-x_4^i)]. \quad (2)$$

To verify that allele A is shared between genomes in H_1, H_3 as often as it happens between genomes in H_2, H_3, we require as null hypothesis that at each i-th locus the probability (1) equals the probability (2). This condition can be written as

$$\mathcal{H}_0 : \mathbb{E}[(x_1^i - x_2^i)(x_3^i - x_4^i)] = 0 \text{ for } i = 1, \ldots, M,$$

where the expectation is the difference between 1 and 2.
Using the empirical frequencies of the alleles as unbiased estimators for the population frequencies, we define the D-statistic as the following normalized test statistic

$$D_M := \frac{X_{(M)}}{Y_{(M)}} = \frac{\sum_{i=1}^M (\hat{x}_1^i - \hat{x}_2^i)(\hat{x}_3^i - \hat{x}_4^i)}{\sum_{i=1}^M (\hat{x}_1^i + \hat{x}_2^i - 2x_1^i x_2^i)(\hat{x}_3^i + \hat{x}_4^i - 2x_3^i x_4^i)}. \quad (3)$$

The values $X_{(M)}$ and $Y_{(M)}$ are the numerator and denominator, respectively. Using $Y_{(M)}$ to normalize the numerator leads to the interpretation of D_M as difference over all loci of the probabilities of having an ABBA or a BABA events, conditional to the event that only ABBA or BABA events are possible.
Appendix 1 shows that, under the hypothesis $\mathcal{H}_0$, the test statistic can be approximated by a standard normal variable. Specifically, the approximation holds with a proper rescaling, since D_M would narrow the peak of the Gaussian around zero for large M (note that this rescaling is an embedded factor in the estimation of the variance of D_M using the block jackknife method (Busing *et al.* 1999) in the software implementation of ANGSD). More generally the treatment could be extended to blockwise independence of the allele counts to take into account linkage disequilibrium.

The convergence results of Appendix 1 apply to the following special cases of the D-statistic:

1. the original D-statistic D_M calculated by sampling a single base from the available reads (Green *et al.* 2010) to estimate the sampling probabilities,
2. the D-statistic D_M evaluated by substituting the frequencies $\hat{x}_j^i$ with the estimated population frequencies $\hat{q}_j^i$ defined in eq 4 for multiple individuals (see Appendix 2).
3. the D-statistic D_M evaluated only over loci where the outgroup is mono-allelic, such as when the Chimpanzee is set as an outgroup to test for gene flow from the Neandertal population into modern out-of-Africa populations (Green *et al.* 2010).

Multiple individuals per group

The D-statistic defined in equation 3 is calculated using population frequencies. In case only one individual per population is chosen, it is easy to get an estimator of the populations' frequencies by simply counting observed bases. In what follows we are interested in getting a meaningful estimate of the frequencies in the case we want to use all the available sequenced individuals without calling genotypes.

This is done using a weighted sum of the estimated allele frequencies for each individual in every group. Assume that given the allele frequency x_j^i, $j = 1, 2, 3, 4$, at locus i for the jth population, we model the observed data as independent binomial trials with parameters n_j^i and x_j^i, where n_j^i is the number of trials. We take the frequency of allele A in the reads of each jth population as an unbiased estimator of the population frequency. Let N_j be the number of individuals in population j. For the ℓth individual within the jth population, let $x_{j,\ell}^i$ be the frequency of allele A at locus i, with estimator $\hat{x}_{j,\ell}^i$ the frequency of allele A for $\ell = 1, \ldots, N_j$. Define $\hat{q}_j^i$ as the weighted sum

$$\hat{q}_j^i := \sum_{\ell=1}^{N_j} w_{j,\ell}^i \cdot \hat{x}_{j,\ell}^i, \quad (4)$$

where each $w_{j,\ell}^i$ is a weight, that is proportional to a quantity depending on $n_{j,\ell}^i$, the number of sequenced bases at locus i for individual ℓ:

$$w_{j,\ell}^i \propto \frac{2n_{j,\ell}^i}{n_{j,\ell}^i + 1}. \quad (5)$$

The estimator $\hat{q}_j^i$ in equation (4) is an estimator for the jth population frequency at locus i with minimal variance (the derivation of the weights as minimizer of the frequency estimator's variance can be found in Appendix 2). Substituting the estimated population frequencies in equation (3) with the weighted estimators determined by formula (4), it is possible to account for multiple individuals per population. Since the weighted estimator is also unbiased, it does not affect the approximation of the D-statistic with a standard normal distribution.
A first application of this method has been the estimation of population frequencies to reveal signatures of natural selection (Li *et al.* 2010). The weights have a strong impact on loci with low number of reads, where they assume a low value, leading to a stronger impact of population frequency estimated from high-depth individuals in each group.

Error estimation and correction

The study of genetic relationships between populations often involves the use of ancient genomes that are subject to high error-rates. We introduce error correction following the idea illustrated in (Orlando *et al.* 2013) to take errors into account and

to obtain a more reliable D-statistic.

The estimation of the type specific error rates is possible using two individuals (one affected by type-specific errors, and one sequenced without errors) and an outgroup, denoted by T, R and O, respectively. Those individuals are considered in the tree ((T,R),O) (see Appendix 3).

After the error matrix is estimated for each individual it is possible to obtain error-adjusted frequencies of alleles in locus i through the following matrix-vector product:

$$\boldsymbol{p}_G^i = \boldsymbol{e}^{-1}\boldsymbol{p}_T^i. \tag{6}$$

where $\boldsymbol{p}_G^i$ and $\boldsymbol{p}_T^i$ are the true and observed vectors of allele frequencies locus i, respectively, and $\boldsymbol{e}$ is the 4×4 type-specific error matrix whose entry $\boldsymbol{e}(a,b)$ is the probability of observing a base of type b when the true base is of type a. Note that estimating $\boldsymbol{e}$ and correcting the allele frequencies is a process best applied before the calculation of weighted allele frequencies for multiple individuals.

Using error-corrected estimators of the population frequencies to calculate the D-statistic does not prevent it to be approximated by a standard normal, because the error-corrected estimators are unbiased for the true population frequency (see Appendix 3).

According to equation (6) one is able to perform the error correction at every locus for every individual. In this way it is possible to build a weighted frequency estimator for each population after the error correction. However the implementation of equation (6) involves the inversion of a matrix and a matrix-vector multiplication at every locus for each individual in all populations. Moreover, as a consequence of the error estimation, there might be negative entries of the inverse $\boldsymbol{e}^{-1}$, which might cause the product of formula (6) to result in negative entries in the vector $\boldsymbol{p}_G^i$.
Consequently we have decided to implement a less precise version of the error correction that is applied to each whole group of individuals instead of every single individual. Assume that the populations' frequencies have been estimated from equation(4), and that it is possible to estimate the probabilities of the 256 alleles combinations AAAA, AAAC, ..., TTTT between the four populations.
In each jth population of individuals, let $\boldsymbol{e}_{(j)}$ be the mean of their error matrices. Then build the error matrix for the four groups, $\boldsymbol{E}$. This has dimension 256×256 and its entry $(a_{1:4}, b_{1:4})$, where $a_{1:4} = (a_1, a_2, a_3, a_4)$ and $b_{1:4} = (b_1, b_2, b_3, b_4)$ are two possible allele patterns of the four populations, is defined as the probability of observing $b_{1:4}$ instead of $a_{1:4}$, assuming independence of the error rates between the four populations:

$$\boldsymbol{E}(a_{1:4}, b_{1:4}) = \boldsymbol{e}_1(a_1, b_1)\cdot\boldsymbol{e}_2(a_2, b_2)\cdot\boldsymbol{e}_3(a_3, b_3)\cdot\boldsymbol{e}_4(a_4, b_4).$$

The equation states that the change from pattern $a_{1:4}$ to $b_{1:4}$ happens with a probability that is the product of the error rates of each population. Note that each error rate is the sum of the error rates of each individual in that population, and so does not take into account how every individual is weighted according to the frequency estimator of formula (4).
Let $\boldsymbol{P}_{error}$ be the vector of length 256 that contains the estimated probabilities of observing allele patterns between the four populations, affected by type-specific errors. Denote by $\boldsymbol{P}_{corr}$ the vector containing the estimated probabilities of patterns not affected by errors. With an approach similar to the one leading to equation 6 it holds that

$$\boldsymbol{P}_{corr} = \boldsymbol{E}^{-1}\boldsymbol{P}_{error}.$$

Using the error-corrected estimated probabilities of combinations of alleles of the type ABBA and BABA it is then possible to calculate numerator and denominator of the D-statistic. This procedure is fast but has the drawback that in every group the error matrix takes into account every individual within a population without its associated weight of equation 5. This means that the portion of alleles related to individuals with lower weights might undergo an excessive error correction.

Correction for introgression from an external population

The improved D-statistic proves to be very sensitive to introgression, but a hypothesized genetic relationship might be rejected because of an admixture involving a population not part of the considered tree. We propose a way to correct this issue and obtain an estimate of the amount of introgression when the source of gene-flow is available.

In this section we analyze the case in which the null hypothesis might be rejected in favor of the alternative hypothesis, but the cause of rejection is not the presence of gene flow between H_3 and either H_1 or H_2, but instead gene flow between an external population H_5 and either H_2 or H_1. Consider the case of Figure S6A, where the null hypothesis might be rejected because of introgression from an external population H_5 into H_2 with rate α. We assume that the external sample for H_5 represents the population that is the source of introgression. Consider H_2 being the population subject to introgression from H_5, and define H_2' the same population when it has not undergone admixture.
The four population subtrees of interest (see Figure S3) are $T_{1:4} = (((H_1, H_2)H_3)H_4)$, which includes the 4-population tree excluding the admixing population, $T_{out} = (((H_1, H_5)H_3)H_4)$, where the population source of introgression replaces the admixed population, and $T_{un} = (H_1(H_2'(H_3, H_4)))$, in which H_2 has not yet undergone admixture and therefore reflects the null hypothesis $\mathcal{H}_0$.
Consider the patterns of four alleles for the three subtrees mentioned above, whose estimated probabilities are respectively denoted as $p_{1:4}$, p_{out} and p_{un}. Using the frequency estimators of equation (4) it is possible to estimate $p_{1:4}$ and p_{out}, but not p_{un} since H_2' is not an observed population.
Assume that testing with the D-statistic on the tree $T_{1:4}$ leads to a rejection of $\mathcal{H}_0$ because the allele frequencies of H_2 are altered by the gene flow from H_5. In fact, any combination of four alleles observed in $T_{1:4}$ has probability

$$p_{1:4} = (1-\alpha)p_{un} + \alpha p_{out}.$$

By solving for p_{un} it follows that

$$p_{un} = \frac{1}{1-\alpha}(p_{1:4} - \alpha p_{out}). \tag{7}$$

Note that if the admixture proportion α is known, then admixture correction is possible. If α is not known and we assume the tree is accepted for $\mathbb{E}[D_{un}] = 0$, where D_{un} is the D-statistic related to the tree T_{un}, then α can be estimated. In this case, p_{un} has to be determined for all values of α, and the correct one will be the value for which $\mathbb{E}[D_{un}] = 0$. In this way an estimate of the admixture rate is obtained for the topology of Figure S3A.

Simulations

Different scenarios have been generated using `msms` (Ewing and Hermisson 2010) to reproduce the trees of Figure 2A, Figure 2B and Figure 2C, in which times are in units of generations. Each topology has been simulated 100 times for a constant population size of $N_e = 10^4$. Mutation and recombination of the simulations are consistent with human data (Ewing and Hermisson 2010). Migrations and admixtures, respectively, for the scenarios of Figure2A and Figure 2C, were simulated with specific options of `msms`. For each simulation we generated 200 regions of size 5MB for each individual and considered only variable sites, except for the case of Figure 2B, where the null hypothesis is affected by type-specific error on some of the individuals. We used a type-specific error of $e_{A\to G} = 0.005$ in populations H_1, H_3. The choice of the region size is compatible with the one estimated for applications with human genomes in (Rasmussen *et al.* 2010). The regions are used by the jackknife estimator (Busing *et al.* 1999) to estimate the standard deviation of the D-statistic accomodating for the non-independence of loci.

As a second step, the simulated genotypes from `msms` were given as input to `msToGlf`, a tool that is provided together with `ANGSD`. Using `msToGlf` it is possible to simulate NGS data from `msms` output files by generating the `pileup` files; that are used as input for `ANGSD`. As parameters for `msToGlf`, we set up the depth as mean of a poisson distribution and we hardcoded the error rates in the program when necessary for the scenario of Figure 2B.

Sequenced human populations

For the real data scenarios of Figure 3A, Figure 3B and Figure 3C we used Illumina sequenced individuals from several human populations. See Table 1 for an overview of the data. The depth of each individual has been calculated using the program `doDepth` of `ANGSD`. The Peruvian individuals used in our study were unadmixed with proportion ≥ 0.95. Estimation of the admixture proportions of these individuals was performed using `ADMIXTURE` (Alexander *et al.* 2009). In every individual, only the autosomal regions of all individuals were taken into consideration and bases were filtered out according to a minimum base quality score of 20 and a mapping quality score of 30. Type-specific error estimates for the Saqqaq, Mi'kmaq and French individuals were performed using the program `doAncError` of `ANGSD`, where the Chimpanzee was used as outgroup and the consensus sequence of human NA12778 as error-free individual (See Figure S4 for the barplot of the estimates of the type-specific error).

Data Availability

The real data used is specified in Table 1. The simulated data has been produced using `msms` (Ewing and Hermisson 2010). The `msms` code for simulations is in the caption of Figure 2. From the output of msms NGS pileup files were simulated with the tool `msToGlf` integrated in `ANGSD` (Nielsen *et al.* 2011). The 1-sample D-statistic and the extended D-statistic implemented in this paper are performed on both real and simulated data with the program `doAbbababa2` of `ANGSD`. `ANGSD` can be downloaded here. A detailed guide including a tutorial for the program `doAbbababa2` is found here.

RESULTS AND DISCUSSION

In the study of our results we compare different implementations of the D-statistic on simulated and real scenarios. We briefly define as D_{ext} the extended D-statistic that we implemented,

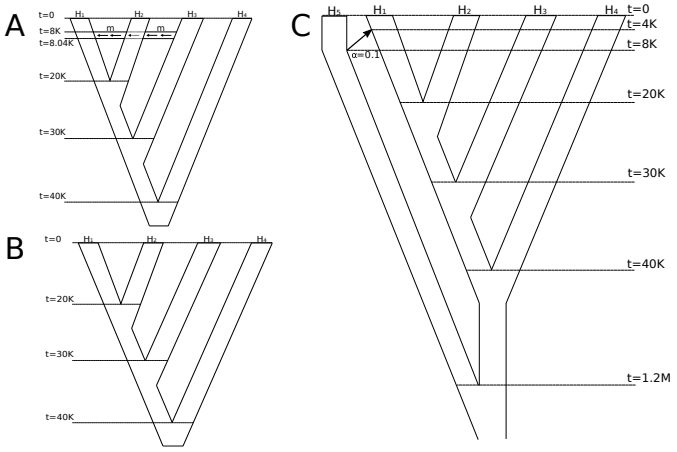


Figure 2 Simulated Scenarios. (A) Simulation of a tree in which migration occurs from population H_3 to H_1. The variable `m` is the (rescaled) migration rate varying between 0, 8, 16, 24, 32, 40 up to 280 with steps of size 20. Expressed in percentage, the migration rate varies between 0%, 0.02%, 0.04%, 0.06%, 0.08%, 0.1% up to 0.7%. Command: `msms -N 10000 -ms 40 200 -I 4 10 10 10 10 0 -t 100 -r 100 1000 -em 0.2 3 1 $m -em 0.201 3 1 0 -ej 0.5 1 2 -ej 0.75 2 3 -ej 1 3 4`. The same command line has been applied with the option `-I 4 40 40 40 40 0` to generate populations of 20 diploid individuals, used to study the power of the method using subsets of $1, 2, 5, 10, 20$ individuals of such populations. (B) Simulation of a tree in which no migration occurs, but type-specific errors on some individuals provide a rejection when testing for correctness of the null hypothesis. Command: `msms -N 10000 -ms 8 200 -I 4 2 2 2 2 0 -t 100 -r 100 1000 -ej 0.5 1 2 -ej 0.75 2 3 -ej 1 3 4`. (C) Simulation of a tree in which H_5 admix with H_1 with an instantaneous unidirectional admixture of rate $\alpha = 0.1$. In this case we expect the null hypothesis to be rejected since H_5 will alter the counts of ABBA and BABA patterns, but the alternative hypothesis does not involve gene flow with H_3. Command: `msms -N 10000 -ms 50 200 -I 5 10 10 10 10 10 0 -t 100 -r 100 1000 -es 0.1 1 0.9 -ej 0.2 6 5 -ej 0.25 1 2 -ej 0.5 2 3 -ej 0.75 3 4 -ej 30 4 5`.

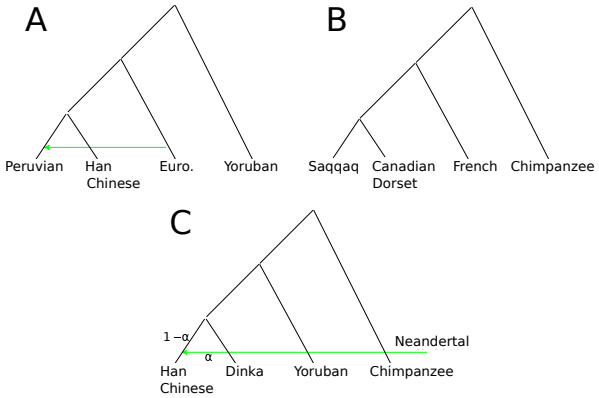


Figure 3 Real Data Scenarios. (A) Tree representing the south-western European migration into the Americas during the European colonization. (B) Tree representing two independent migrations into northwestern Canada and Greenland. (C) Tree representing the presence of Neandertal genome into a modern non-african population, specifically the Han Chinese.

Table 1 List of the Genomes Used in Real Data Scenarios.

Genome Id	Major population division	Depth	Reference study
HG01923	Peruvian (PEL)	6.3X	(Altshuler *et al.* 2010)
HG01974	Peruvian (PEL)	11.9X	(Altshuler *et al.* 2010)
HG02150	Peruvian (PEL)	7.3X	(Altshuler *et al.* 2010)
HG02259	Peruvian (PEL)	6.5X	(Altshuler *et al.* 2010)
HG02266	Peruvian (PEL)	3.8X	(Altshuler *et al.* 2010)
NA18526	Han Chinese (CHB)	6.6X	(Altshuler *et al.* 2010)
NA18532	Han Chinese (CHB)	7.3X	(Altshuler *et al.* 2010)
NA18537	Han Chinese (CHB)	2.9X	(Altshuler *et al.* 2010)
NA18542	Han Chinese (CHB)	7.3X	(Altshuler *et al.* 2010)
NA18545	Han Chinese (CHB)	6.2X	(Altshuler *et al.* 2010)
NA06985	CEPH (CEU)	12.8X	(Altshuler *et al.* 2010)
NA06994	CEPH (CEU)	5.5X	(Altshuler *et al.* 2010)
NA07000	CEPH (CEU)	9.4X	(Altshuler *et al.* 2010)
NA07056	CEPH (CEU)	4.9X	(Altshuler *et al.* 2010)
NA07357	CEPH (CEU)	5.7X	(Altshuler *et al.* 2010)
NA12778	CEPH (CEU)	6.9X	(Altshuler *et al.* 2010)
NA18501	Yoruba (YRI)	6.4X	(Altshuler *et al.* 2010)
NA18502	Yoruba (YRI)	4.9X	(Altshuler *et al.* 2010)
NA18504	Yoruba (YRI)	10.1X	(Altshuler *et al.* 2010)
NA18505	Yoruba (YRI)	6.1X	(Altshuler *et al.* 2010)
NA18507	Yoruba (YRI)	3X	(Altshuler *et al.* 2010)
HGDP00778	Han Chinese (CHB)	23.4X	(Consortium 2003)
DNK02	Dinka	25.8X	(Meyer *et al.* 2012)
HGDP00927	Yoruban (YRI)	28X	(Consortium 2003)
AltaiNea	Neanderthal	44.9X	(Green *et al.* 2010)
pantro2	Chimpanzee	-	(Kent *et al.* 2002)
saqqaq	Saqqaq	15.7X	(Rasmussen *et al.* 2010)
MARC1492	Ancient Canadian Dorset (Mi'kmaq - New England)	1.1X	(Raghavan *et al.* 2014)
HGDP00521	French	23.8X	(Consortium 2003)

D_{1base} the D-statistic calculated by sampling 1 sequenced base per locus (Green *et al.* 2010) and D_{geno} the D-statistic calculated with equation (3) using the allele frequencies estimated from the true genotype (the true genotype is only available in the case of simulated data).

The D-statistic is computed on blocks of 5Mb, to ensure that every block is not subject to linkage disequilibrium from the other blocks, and that the number of loci in each block is large enough to make the D-statistic approach the approximation by a standard normal distribution (see Appendix 1). The use of blocks allows for estimation of a proper normalization constant for the D-statistic using the m-block jack-knife method (Busing *et al.* 1999). The threshold for rejection of the null hypothesis is set to a p-value 0.001, corresponding approximately to the two-tailed acceptance region $[-3, 3]$.

The formula for calculating the D-statistic is given in equation (3) and finds amongst its current implementations, the ones in (Patterson *et al.* 2012) and (Nielsen *et al.* 2011), with sampling of one base per locus from only one individual in each population. Such an implementation is computationally fast but has many drawbacks:

- when genomes are sequenced at low or medium depth (1X-10X), sampling one base might lead to a process with high uncertainty;
- base transition errors might affect the sampling of the base adding more uncertainty;
- only one individual per population is used;
- for a chosen individual chosen from a population, the reads are not used to evaluate the D-statistic, but only to sample one base.

We have proposed a solution to these problems with the extended version of the D-statistic D_{ext} implemented in `ANGSD` and we will show in the following results how all the problems mentioned above are addressed.

Comparison of Power Between the Different Methods

Using simulated and real data we compare the different types of D-statistics to study their sensitivity to gene flow, and illustrate how the improved D-statistic D_{ext} is not affected by the issues faced by the current D-statistic D_{1base}, and even reach the performances of the D-statistic based on true genotype D_{geno} at a rather low sequencing depth.

To evaluate the power of the different methods we first simulated NGS data based on coalescent simulations with mutation and recombination rates consistent with human populations (Ewing and Hermisson 2010). We simulated without sequencing error four populations with a varying amount of migration from H_3 to H_1 (see Figure 2A) and applied the D-statistic based on five individuals from each population for two different sequencing depths. Figure 4A and Figure 4B show the power of the methods for depth 0.2X and 2X. Here power is the rejection rate of the null hypothesis when there is a migration from H_3 to H_1 in the tree $(((H_1, H_2)H_3)H_4)$.

The extended D-statistic proves to be effective in detecting gene flow even when the simulated depth is very low. For the scenario with sequencing depth 0.2X, D_{1base} is not able to detect almost any case of migration from H_3, while D_{ext} reacts with an acceptable rejection rate already for a migration rate as low as $m = 0.15\%$. Of course such a very low depth does not allow the D-statistic to perform as well as D_{geno}. In the case of sequencing depth 2X, D_{1base} does not always detect the alternative hypothesis and has also a considerable delay in terms of the migration rate necessary to do that, when compared to D_{ext}. Furthermore D_{ext} follows almost exactly the behavior of the power related to D_{geno}. This means that with a depth above 2X we can expect the D-statistic D_{ext} to perform as well as knowing the exact genotypes of the data.
A deeper analysis to study the effect of using multiple individuals per group is illustrated in Figure S1. Here we simulated again the scenario with depth 0.2X, and compared the use of 1, 2, 5, 10 and 20 individuals per population. The graph shows that using multiple individuals increases the power of the method and at the same time decreases the standard deviation of D_{ext}.
From 5000 simulations of the null hypothesis at depth 0.2X we produced the QQ-plot of Supplementary Figure 8. Here we can see that, despite we simulated only 200 blocks of 5Mb length for each individual, the D-statistic already shows its asymptotic property of convergence to a standard normal.

The power of D_{ext} and D_{1base} are compared in a real data scenario using Illumina sequenced modern human populations from the 1000 Genomes Project with a varying sequenced depth in the range 3-13X. We specifically used PEL=Peruvian, CEU=European, CHB=Han Chinese and YRI=African Yoruban individuals to form the tree (((PEL,CHB)CEU)YRI) shown in Figure 3A. This scenario represents the southwestern European gene flow into the ancestors of the Native Americans (Raghavan *et al.* 2013). Each of the four populations consists of 5 sequenced individuals when evaluating D_{ext}, and a distinct one of those individuals when evaluating D_{1base} five times (see Figure 4C). The extended D-statistic D_{ext} has much lower standard errors, that corresponds to a smaller p-value than in the case of D_{1base}, and therefore a more significant rejection. See Table S1 for a better comparison of the values of the different D-statistics.
It is worth to underline that the presence of structured populations might lead to false positives because the structure is not considered in the model. If there is structure within H_1, H_2, the properties of the D-statistic are preserved. However, if the population was structured prior to the split of H1 and H2, then it will affect the D-statistics.

Error Impact and Correction

Sequencing or genotyping errors are known to have a large impact on the D-statistic (Orlando *et al.* 2013). Using simulation we show that if the type-specific error rates are known then we can correct the D-statistic accordingly. We simulate the tree under the null hypothesis. However, we add base $A \to G$ error rate of 0.005 in populations H_1 and H_3 in order to alter the observed number of ABBA and BABA combination of alleles, and consequently lead to a possible rejection of the null hypothesis.

In the plot of Figure 5A are represented the estimated distributions of the Z-scores related to D_{ext} before and after error estimation and error correction, for 100 simulations of a tree $(((H_1, H_2)H_3)H_4)$ without any gene flow, where we have also introduced type-specific error for transitions from allele A to another allele for the individuals in H_1, H_2, H_3 at different rates. The test statistic has high values due to the error while all simulations fall in the acceptance interval if we perform error correction.
The uncorrected D-statistic performs poorly because of the errors in the data that cause rejection of the null hypothesis in

all simulations. It is remarkable to observe that D_{ext} has good performances already at depth 0.5X. This means that even small error rates in the data make the D-statistic very sensible to the rejection of $\mathcal{H}_0$. Therefore we require to apply error correction to our data. The result is that the Z-scores fall into the acceptance threshold and the null hypothesis is fulfilled. The distribution of corrected Z-scores is not perfectly centered in 0 because of imperfect error correction.

The most obvious need for error correction in real applications is the use of ancient genomes, which have a large amount of errors, especially transitions. To illustrate the effect of errors in real data and our ability to correct for them we use two ancient genomes which contain a high sequencing error rate due to *post mortem* deamination. The tree (((Saqqaq,Dorset)French)Chimpanzee) of Figure 3B illustrates the migrations to western Canada (Canadian Dorset Mi'kmaq genome) and southwestern Greenland (Saqqaq genome). Due to the effect of deamination prior to sequencing (Rasmussen *et al.* 2010; Raghavan *et al.* 2014), the two ancient genomes have high type-specific error rates as shown in Table S2 and Figure S4. The error rates alter the counts of ABBA and BABA patterns, which bias the uncorrected D-statistic.
We expect the tree to be true under the null since Saqqaq and Dorset have a recent common ancestor (Raghavan *et al.* 2015). In Figure 5B we compare the extended D-statistic D_{ext} in four cases: firstly using observed data, secondly removing all transitions which are related to most of the errors, thirdly applying error correction and lastly combining error correction and transitions removal. Note that the removal of transitions related to the pairs of alleles A,C and G,T is the current standard technique to avoid high error rates when calculating the D-statistic from damaged low-coverage data. The uncorrected D-statistic rejects the null hypothesis whereas correction or transition removal gives a non-significant test. Error correction performs better than transition removal, providing a value of the D-statistic that is closer to 0 and has smaller standard deviation. Table S3 shows the values related to the four D-statistics in this scenario. Supplementary Figure 11 illustrates the effect of increasing and decreasing the removal of error for the base transition $C \to G$ and $C \to T$ for one of the Saqqaq, Dorset and French genomes. This correspond to add a value to the estimated error rate matrix of one of the individuals. Observe that the French individual is less affected by the addition or removal of error than the first two individuals. Moreover all 3 individuals are more sensible to the error rate in the case of transversion $C \to T$.

Correction for External Introgression

We use simulations of a scenario with external introgression to verify the performance of correction for gene-flow in restoring a four-population tree configuration that lead to the acceptance of the null hypothesis $\mathcal{H}_0$. In the simulation case we know the value of α, that is the amount of introgression, therefore correction is possible. Thereafter we use a known genetic relationship involving the Neandertal introgression into out-of-Africa modern individuals in Europe and Asia (Green *et al.* 2010; Wall *et al.* 2013) to correct for the effect of admixture. In addition we show that, if we assume the absence of gene flow in the tree topology, then we can estimate the amount of introgression, and compare it with the estimation involving the original D-statistic tools.

For some species there are introgression events from an external source which can affect the D-statistic when performing test for admixture among the species. We performed 100 simulations of the null hypothesis $(((H_1, H_2)H_3)H_4)$ of Figure 2C, for which an external population H_5 is admixed with H_2 with rate $\alpha = 0.1$. The plot of Figure 6A shows the estimated distribution of the Z-scores related to the observed and admixture-corrected D_{ext}. The observed D-statistic is positive and has Z-scores that reject the null hypothesis. Applying equation (7) we are able to remove the effect of gene flow from H_2. The result of removal of the gene flow's effect is that the estimated probabilities of ABBA and BABA combinations of alleles are altered and the resulting calculated values of the D-statistic lead to acceptance of the null hypothesis $\mathcal{H}_0$.

For human populations it is problematic to use the D-statistics when applied to both African and non-African populations because of ancient gene-flow from other hominids into non-Africans. Therefore, $\mathcal{H}_0$ might not fulfilled for any tree $(((H_1, H_2)H_3)H_4)$ where an ingroup consists of both an African and a non-African population. This leads to rejection of the tree and to the natural conclusion that there is gene flow between H_3, H_2 (resp. H_3, H_1). However, if there is known external admixture from a population H_5, it is possible to correct for admixture from this external contribution.

We illustrate the problem and our ability to correct for it using the tree shown in Figure 3C, which shows introgression of the Neanderthal genome into the ancestors of the Han Chinese population. The correction is performed for the admixture proportion α in the range $[0, 0.05]$ in steps of 0.01. The value of α for which the D_{ext} is closest to 0 might be considered as an estimate of the admixture rate. We choose these populations because we can compare our result with the estimate from previous studies of the same populations (Green *et al.* 2010; Wall *et al.* 2013). The study of (Green *et al.* 2010) estimated α to be in the range $[0.01, 0.04]$, while (Wall *et al.* 2013) estimated it as being $\alpha = 0.0307$ with standard deviation 0.0049. The result is shown in Figure 6B for the tree (((Han Chinese, Dinka) Yoruban) Chimpanzee) for different admixture rates α used to correct for the introgression of the Neandertal population into the Han Chinese population. The red polygon is the interval in which α is estimated to be (Green *et al.* 2010). The black dot coincides with the value of $\alpha = 0.0307$ calculated in (Wall *et al.* 2013). The blue polygon is 3 times the standard deviation of D_{ext}. For almost the whole range of reported admixture proportions, the tree is not rejected after adjustment for admixture, indicating that the uncorrected D-statistic concluded the presence of gene flow. When D_{ext} is 0, we estimate $\alpha = 0.03$ with standard deviation 0.0042, which is similar to previous estimates.

In both the cases of simulated and real data we have thus been able to distinguish the case in which the alternative hypothesis is due to an external introgression and not to admixture from H_3. In our simulations, the admixture correction seems not to suffer from the effect of drift, which is not modeled in the correction. In fact the branch leading to H_5 splits 8000 generations in the past and admixes 4000 generations in the past on the branch leading to H_1. Thus there is a drift affecting gene frequencies of both the admixing and admixed populations.
In the case of real data the exact amount of admixture α is not previously known. Therefore we calculated the D-statistic for the tree (((Han Chinese, Dinka) Yoruban) Chimpanzee) using

admixture-corrected values of the probabilities of allele patterns, considering values of the admixture rate falling in the interval estimated in (Green *et al.* 2010). Without admixture correction, the obvious conclusion would have been that for the tree (((Han Chinese,Dinka)Yoruban)Chimpanzee) there is gene flow between the Yoruban and Dinka populations.

Conclusions

In summary we have implemented a different D-statistic that address the drawbacks of the current implementations of the D-statistic, but still preserve the approximation as a standard normal distribution (see Appendix 1) that allows for a statistical test. The extended D-statistic D_{ext} allows for multiple individuals per population and instead of sampling one base according to the estimated allele frequencies, uses all the available sequenced bases.

Using both simulations and real data we have shown that

1) the extended D-statistic D_{ext} has more power than the alternative methods, with an increased sensibility to admixture events. Moreover, even without a large amount of data, the extended D-statistic shows a good asymptotic convergence and therefore a low false positive rate;

2) the performance of the extended D-statistic is the same as when true genotype is known for a depth of at least 2X,

3) we can accomodate type-specific errors to prevent that en eventually wrong acceptance or rejection of the null hypothesis is caused by error-affected allele frequencies. The error estimation and correction reveal to be especially suited in the case of ancient genomes, where error rates might be high due to chemical treatments prior to sequencing and degradation over time;

4) we can calculate the D-statistic after correcting for admixture from an external known population, such as in the case of Neandertal gene flow into the Han Chinese population.

The extended D-statistic D_{ext} is especially effective compared to the standard D-statistic D_{1base} when applied to data with low/variable depth, multiple individuals and ancient DNA.

APPENDICES

The setup of the theoretical treatment consists of four sampled genomes representing four populations H_1, H_2, H_3, H_4, for which we assume the relationship illustrated in Figure 1. Each genome is considered to have M di-allelic loci. We will consider the situation in which M grows to infinity. Each locus i consists of a certain number n_j^i of alleles A and B, where $j = 1,2,3,4$, is the index of the jth genome. Moreover we assume independence between the loci. Assume that at a locus i the allele frequencies in the four groups of individuals $\boldsymbol{x}_i := (x_1^i, x_2^i, x_3^i, x_4^i)$ follow a locus-dependent distribution $F_i(\boldsymbol{x})$, $i = 1,\dots,M$ and let $\hat{\boldsymbol{x}}_i := (\hat{x}_1^i, \hat{x}_2^i, \hat{x}_3^i, \hat{x}_4^i)$ be an unbiased estimator of $\boldsymbol{x}_i$ at locus i, such as the relative frequencies of the allele A in each population. The populations' frequencies are considered to be a martingale process.

The null hypothesis that the tree of Figure 1 is correct can be rewritten as follow:

$$\mathcal{H}_0 : \mathbb{E}[(x_1^i - x_2^i)(x_3^i - x_4^i)] = 0 \ \text{ for } \ i = 1,\dots,M,$$

where the expectation is done on the difference between the probabilities of ABBA and BABA events deduced in equations (1) and 2. Using the empirical frequencies as proxies for the expected values,

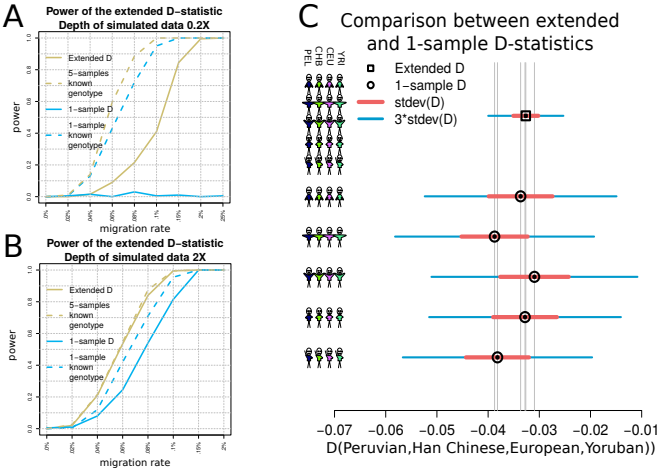


Figure 4 Detection of Admixture and Migration. (A,B) Rejection rate of the null hypothesis as a function of the migration rate in the tree $(((H_1, H_2)H_3)H_4)$, where a migration from H_3 to H_1 occurs. The yellow and blue solid lines represent respectively the power of the method related to D_{ext} and D_{1base}. The yellow dashed line represents the rejection rate when the genotypes of the 5 individuals in each population are known and thus equation (3) can be applied. The blue dashed line illustrates the power of the method when only one genome per population has known genotypes. D_{ext} performs almost as well as knowing the true genotypes already with depth 2X. (C) Value of D_{ext} (black square) and values of D_{1base} (black circles) using respectively 5 genomes per population and one of them from each population. Each D statistic shows its associated standard deviation multiplied by 1 and 3. On the left side of the graph, the stickmen represent for each column the composition of the group by number of individuals.

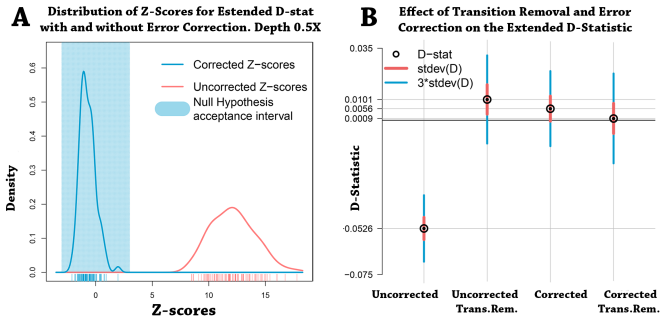


Figure 5 Effect of Error Estimation and Correction. (A) Estimated distributions of the Z-scores related to D_{ext} for the null hypothesis $(((H_1, H_2)H_3)H_4)$ in which H_1, H_3 and H_2 has probability 0.005 and 0.01 of transition from base A, respectively. The blue polygon represents the interval where a Z-score would accept the null hypothesis. The red line represents the distribution of Z-scores before type-specific errors are corrected. In blue we have the Z-scores after correction. (B) Values of D_{ext} in four different cases for the tree (((Saqqaq,Dorset)French)Chimpanzee). The black circles are the values of the uncorrected D-statistic, removal of ancient transitions, error correction, error correction and ancient transitions removal. The red and blue lines represent the standard deviations and the value they need to reach the threshold of $|Z| = 3$, respectively.

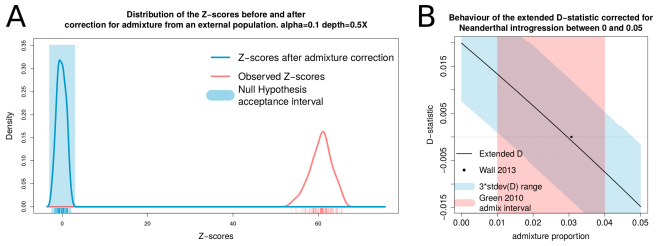


Figure 6 Effect of Correction from External Introgression. (A) Estimated distribution of the Z-scores related to D_{ext} from the 100 simulations of the null hypothesis $(((H_1, H_2)H_3)H_4)$ with introgression of rate $\alpha = 0.1$ from an external population H_5 into H_2. The Z-scores of the observed tree are far off the acceptance interval because of the admixture from H_5. Once the portion of genome from the external population is removed from H_2, the tree fulfills the null hypothesis and the Z-scores all fall in the acceptance interval defined by $|Z| \leq 3$. (B) Behavior of the D_{ext} of the tree (((Han Chinese,Dinka)Yoruban)Chimpanzee) as a function of the admixture rate α used to correct for the introgression of the Neandertal population into the Han Chinese population. The red polygon is the interval in which (Green *et al.* 2010) estimates α to fall in. The black dot coincides with the value of $\alpha = 0.0307$ calculated by (Wall *et al.* 2013) using the tree (((Han Chinese,Yoruban)Neandertal)Chimpanzee), with standard deviation 0.0049. The blue polygon is 3 times the standard deviation of D_{ext}. When D_{ext} is 0, we estimate $\alpha = 0.03$ with standard deviation 0.0042.

we build the following normalized test statistic, also known as D-statistic:

$$D_M = \frac{\sum_{i=1}^{M}(\hat{x}_1^i - \hat{x}_2^i)(\hat{x}_3^i - \hat{x}_4^i)}{\sum_{i=1}^{M}(\hat{x}_1^i + \hat{x}_2^i - 2x_1^i x_2^i)(\hat{x}_3^i + \hat{x}_4^i - 2x_3^i x_4^i)},$$

where the values

$$X_{(M)} = \sum_{i=1}^{M}(\hat{x}_1^i - \hat{x}_2^i)(\hat{x}_3^i - \hat{x}_4^i),$$
$$Y_{(M)} = \sum_{i=1}^{M}(\hat{x}_1^i + \hat{x}_2^i - 2x_1^i x_2^i)(\hat{x}_3^i + \hat{x}_4^i - 2x_3^i x_4^i)$$

are the numerator and denominator of the D-statistic, respectively.

Appendix 1

Convergence of the D-Statistic. In this paragraph we prove that the D-statistic defined as

$$D_M = \frac{X_{(M)}}{Y_{(M)}}$$

converges in distribution to a standard normal variable up to a constant.
Rewrite the numerator and denominator as

$$X_{(M)} = \sum_{i=1}^{M} X_i$$
$$Y_{(M)} = \sum_{i=1}^{M} Y_i,$$

where the values X_i and Y_i are defined for each $i = 1, \ldots, M$ by

$$X_i = (\hat{x}_1^i - \hat{x}_2^i)(\hat{x}_3^i - \hat{x}_4^i),$$
$$Y_i = (\hat{x}_1^i + \hat{x}_2^i - 2x_1^i x_2^i)(\hat{x}_3^i + \hat{x}_4^i - 2x_3^i x_4^i).$$

Consider the series of independent variables X_i in the numerator of D_M, having means μ_i. Every term X_i of the numerator is an unbiased estimate of $(x_1^i - x_2^i)(x_3^i - x_4^i)$, assuming the observed allele counts are binomially distributed (Reich *et al.* 2009). We show in the following proposition that every term of the numerator of the D-statistic has expectation $\mu_i = 0$ for $i = 1, \ldots, M$ by calculating the expectation of $(x_1^i - x_2^i)(x_3^i - x_4^i)$.

Theorem 1. *Given the tree topology of Figure 1, it holds that* $\mathbb{E}[(x_1^1 - x_2^i)(x_3^1 - x_4^i)] = 0$ *for* $i = 1, \ldots, M$.

Proof. Let $x_{1:2}^i$, $x_{1:3}^i$ and $x_{1:4}^i$ be the frequencies of the ancestral populations of (x_1^i, x_2^i), (x_1^i, x_2^i, x_3^i) and the root of the tree, respectively, as illustrated in Figure 1. Let $\mathcal{X}$ be the set of those three frequencies. Using the martingale properties of the frequencies it follows that

$$\begin{aligned}\mathbb{E}[(x_1^i - x_2^i)(x_3^i - x_4^i)] &= \mathbb{E}\Big[\mathbb{E}[(x_1^i - x_2^i)(x_3^i - x_4^i)|\mathcal{X}]\Big] \\ &= \mathbb{E}\Big[\mathbb{E}[x_1^i - x_2^i|\mathcal{X}]\mathbb{E}[x_3^i - x_4^i|\mathcal{X}]\Big] \\ &= \mathbb{E}\Big[\mathbb{E}[x_1^i - x_2^i|x_{1:2}]\mathbb{E}[x_3^i - x_4^i|\mathcal{X}]\Big] \\ &= \mathbb{E}\Big[0 \cdot \mathbb{E}[x_3^i - x_4^i|\mathcal{X}]\Big] = 0\end{aligned} \tag{8}$$

□

Therefore X_i has mean 0 for all $i = 1, \ldots, M$.

To prove convergence of the D-statistic for large M we assume the following:

1. Let σ_i^2 be the variance of every term X_i. Denote with v_M the sum $\sum_{i=1}^{M} \sigma_i^2$, then
$$v_M \to \infty \quad \text{for} \quad M \to \infty. \tag{9}$$
2. Let Y_i, $i = 1, \ldots, M$, be the series of independent variables in the denominator of D_M, having means γ_i. Then
$$\frac{1}{M}\sum_{i=1}^{M} \gamma_i \to \gamma \quad \text{for} \quad M \to \infty. \tag{10}$$
3. Denote with τ_i^2 the variance of Y_i. Then
$$\frac{1}{M^2}\sum_{i=1}^{M} \tau_i^2 \to 0 \quad \text{for } M \to \infty. \tag{11}$$

If the numerator and denominator are sums of iid variables, conditions (9), (10) and (11) are fulfilled. In fact, if every term X_i has variance σ^2, the sum of variances is $v_M = M\sigma^2$ and (9) holds. If every term Y_i has mean and variance γ and τ^2, respectively, equation (10) is still valid because the arithmetic mean is done on identical values. Moreover, equation (11) holds because

$$\frac{1}{M^2}\sum_{i=1}^{M} \tau^2 = \frac{1}{M}\tau^2,$$

that converges to zero for $M \to \infty$.

The convergence of the D-statistic D_M is proved in steps, analyzing separately the numerator and the denominator. We begin by stating all the necessary theorems. Firstly, we consider an extension of the central limit theorem (CLT) (Johnson 2004), that will be applied to the numerator $X_{(M)}$. Subsequently we state the law of large number (LLN) (Lamperti 1996) for not i.i.d. variables that is used for the denominator $Y_{(M)}$ of the D-statistic. Thereafter we enunciate one of the consequences of Slutsky's theorem (Slutsky 1925; Pesaran 2015). The last step is a theorem for the convergence of the D-statistic, proved by invoking all the previous statements, applied to the specific case of the D-statistic.

Theorem 2 (CLT for independent and not identically distributed variables). *Let $\{X_i\}_{i=1}^M$ be a sequence of independent (but not necessarily identically distributed) variables with zero mean and variances σ_i^2. Define v_M as $\sum_{i=1}^M \sigma_i^2$. Consider the following quantity*

$$\Lambda_\epsilon(M) := \sum_{i=1}^M \mathbb{E}\left[\left(\frac{X_i}{\sqrt{v_M}}\right)^2 \mathbb{I}\left(\left|\frac{X_i}{\sqrt{v_M}}\right| \geq \epsilon\right)\right],$$

where $\mathbb{I}(\cdot)$ defines the indicator function. If for any $\epsilon > 0$ it holds that $\lim_{M\to\infty} \Lambda_\epsilon(M) = 0$, then the normalized sum $U_M = \sum_{i=1}^M X_i/\sqrt{v_M}$ converges in distribution to a standard normal $\mathcal{N}(0,1)$.

Theorem 3 (LLN for independent and not identically distributed variables). *Let $\{Y_i\}_{i=1}^M$ be a sequence of uncorrelated random variables. Define $\bar{Y}_M$ as the empirical average $\frac{1}{M}\sum_{i=1}^M Y_i$. Denote with γ_i and τ_i^2 the expectation and variance of each variable. If conditions (10) and (11) are fulfilled, then for each $\epsilon > 0$*

$$\lim_{M\to\infty} \mathbb{P}\left(|\bar{Y}_M - \frac{1}{M}\sum_{i=1}^M \gamma_i| \geq \epsilon\right) = 0.$$

Equivalently the empirical average $\bar{Y}_M$ converges in probability to $\lim_{M\to\infty} \frac{1}{M}\sum_{i=1}^M \gamma_i = \gamma$.

Theorem 4 (Slutsky's Theorem). *Let $X_{(M)}$ and $Y_{(M)}$ be two sums of not iid random variables. If the former converges in distribution to X and the latter converges in probability to a constant γ for $M \to \infty$, then the ratio $X_{(M)}/Y_{(M)}$ converges in distribution to X/γ.*

The last step is a theorem for the convergence of the D-statistic, proved by invoking all the previous statements, applied to the specific case of the D-statistic.

Theorem 5 (Convergence in distribution of the D-statistic). *Consider the D-statistic defined by*

$$D_n = \frac{X_{(M)}}{Y_{(M)}} = \frac{\sum_{i=1}^M X_i}{\sum_{i=1}^M Y_i} \quad \in [-1, +1],$$

where numerator and denominator are sum of independent (but not necessarily identically distributed) variables. Under the assumptions of (9), (10) and (11), the D-statistic converges in distribution to a standard normal if rescaled by the constant:

$$c_M D_M \xrightarrow{d} \mathcal{N}(0,1) \quad \text{for } M \to \infty.$$

The arrow denotes the convergence in distribution and c_M is defined as

$$c_M := \gamma\frac{M}{\sqrt{v_M}}.$$

Here v_M is the sum of the variances of the first M terms of the numerator, and γ is the convergence value of thee arithmetic mean of the denominator's expectations for $M \to \infty$.

Proof. First consider Theorem 2 applied to the rescaled numerator $U_M = X_{(M)}/\sqrt{v_M}$. It is necessary to prove that for any $\epsilon > 0$ it holds that $\lim_{M\to\infty} \Lambda_\epsilon(M) = 0$ to ensure the convergence in distribution. First observe that $|X_i| \leq 1$ for any index i. Consequently we have the inequality

$$\begin{aligned}\Lambda_\epsilon(M) &\leq \left(\frac{1}{\sqrt{v_M}}\right)^2 \sum_{i=1}^M \mathbb{E}\left[\mathbb{I}\left(\left|\frac{1}{\sqrt{v_M}}\right| \geq \epsilon\right)\right] \\ &= \frac{1}{v_M}\mathbb{P}\left(|X_i| \geq \epsilon\sqrt{v_M}\right) \leq \frac{1}{v_M}\frac{\mathbb{E}[X_i]}{\epsilon\sqrt{v_M}} \leq \frac{1}{v_M}\frac{1}{\epsilon\sqrt{v_M}},\end{aligned}$$

where Markov's inequality is applied to the last line of the equation. Thus U_M converges in distribution to a standard normal $\mathcal{N}(0,1)$

Since conditions (10) and (11) are fulfilled by assumption, it is possible to invoke Theorem 3 to state that the empirical average of the denominator $Y_{(M)}/M$ converges in probability to a constant γ, which is positive since every term of the denominator is positive.

Finally, we apply Theorem 4 using the proper constants that follows from Theorems 2 and 3 applied to the numerator and denominator, respectively. We proved that the sum $X_{(M)}/\sqrt{v_M}$ converges in distribution to a standard normal $\mathcal{N}(0,1)$ and $Y_{(M)}/M$ converges in probability to the constant γ, that is the limit of the arithmetic mean of equation 10. Thus the ratio

$$\frac{M}{\sqrt{v_M}}\frac{X_{(M)}}{Y_{(M)}}$$

converges in distribution to a gaussian $\mathcal{N}(0, \sqrt{\gamma}^{-1})$. The convergence in distribution of D_M to a standard normal variable is accomplished by rescaling by the following multiplicative constant

$$c_M = \gamma\frac{\sqrt{v_M}}{M}.$$

□

The results of this proof apply also in the following cases of the D-statistic:

1. the original D-statistic D_M calculated by sampling a single base at each site from the available reads (Green *et al.* 2010) to estimate the sampling probabilities. In this case every term on the numerator has possible values −1, 0, +1. Each population frequency x_j^i is parameter of a binomial distribution $Bin(1, x_j^i)$, and is estimated by the frequency of the observed base A at locus i in population j,

2. the D-statistic is evaluated using the estimated population frequencies $\hat{q}_j^i$ defined in equation 4 for multiple individuals in a population (see Appendix 2). In fact, the estimator for multiple individuals is still an unbiased estimate for the population frequency (Li *et al.* 2010), therefore every term of the numerator is still an unbiased estimate for the difference between the probabilities of ABBA and BABA events.

3. the D-statistic is evaluated only over loci with allele frequency $x_4 = 1$ for population H_4. This special case of D-statistic has been used, for example, to assess the presence of gene flow from the Neandertal population into modern out-of-Africa individuals, setting a Chimpanzee as outgroup, and considering only loci where the outgroup showed uniquely allele A (Green *et al.* 2010). in fact, Theorem 1 still holds because in equation (8) the term $E[x_1^i - x_2^i|x_{1:2}]$ is zero, independently of which values x_4^i assumes.

Appendix 2

Multiple Genomes. We assume a di-allelic model with alleles A and B and the four populations H_1, H_2, H_3, H_4 that consist each of a number of distinct individuals N_j, $j = 1,2,3,4$, where j indexes the populations. Given the allele frequency x_j^i, $j = 1,2,3,4$, at locus i, we model the observed data as independent binomial trials with parameters n_j^i and x_j^i for $j = 1,2,3,4$, where n_j^i is the number of

trials. One possible unbiased estimator of the population frequency is

$$\hat{x}_j^i := \frac{n_j^{i,A}}{n_j^i},$$

where $n_j^{i,A}$ is the total number of As and n_j^i the total number of bases observed for the selected population and locus.
For locus i denote the allele frequency of individual ℓ in population j as $x_{j,\ell}^i$. We use as its unbiased estimator

$$\hat{x}_\ell^i := \frac{n_{j,\ell}^{i,A}}{n_{j,\ell}^i},$$

namely the ratio between the number of observed As and the total number of observed alleles at locus i in genome ℓ. The idea is to condense all the quantities $\hat{x}_\ell^i$ into a single value $\hat{q}_j^i$ that minimizes the variance of the sum of the estimated individuals' frequencies w.r.t. a set of normalized weights

$$\{w_{j,\ell}^i\}_{\ell=1}^{N_h}, \quad \sum_{\ell=1}^{N_h} w_{j,\ell}^i = 1$$

such that

$$\hat{q}_j^i := \sum_{\ell=1}^{N_h} w_{j,\ell}^i \cdot \hat{x}_{j,\ell}^i.$$

The estimated population frequency $\hat{q}_j^i$ is an unbiased estimator of the frequency of population j at the ith locus (Li *et al.* 2010). The aim of the weight estimate is to determine the set of weights that minimizes the variance of $\hat{q}_j^i$. To do this, we first determine the variance of each individual's frequency.
Consider a genome ℓ in population j. We approximate the frequency estimator of genome ℓ in population j, namely $\hat{x}_{j,\ell}^i$, defining

$$Y_{j,\ell}^i := \frac{\sum_{m=1}^{n_{j,\ell}^i} I_m}{n_{j,\ell}^i},$$

where $n_{j,\ell}^i$ is the total number of reads for individual ℓ and $I_m \backsim \mathcal{B}in(1, x_j^i)$ for $m = 1, \ldots, n_{j,\ell}^i$. Note that the Binomial variables are parametrized by x_j^i and not by $x_{j,\ell}^i$. The variance of $Y_{j,\ell}^i$ is

$$\mathbb{V}[Y_{j,\ell}^i] = \frac{1}{(n_{j,\ell}^i)^2}\left(\sum_{m=1}^{n_{j,\ell}^i} \mathbb{V}[I_m] + 2\sum_{r<t}^{n_{j,\ell}^i} Cov[I_r, I_t]\right). \qquad (12)$$

The variance of the indicator function I_m

$$\mathbb{V}[I_m] = x_j^i(1 - x_j^i).$$

It remains to find the covariance

$$Cov[I_r, I_t] = \mathbb{E}[I_r I_t] - \mathbb{E}[I_r]\mathbb{E}[I_t] = \mathbb{E}[I_r I_t] - {x_j^i}^2,$$

where, marginalizing on the underlying genotype G and assuming HWE, it follows that

$$\begin{aligned}\mathbb{E}[I_r I_t] &= \sum_{g\in\{AA,AB,BB\}} \mathbb{P}(I_r I_t = 1, G = g)\\ &= \mathbb{P}(I_r I_t = 1|G = AA)\mathbb{P}(G = AA)\\ &+ 2\mathbb{P}(I_r I_t = 1|G = AB)\mathbb{P}(G = AB)\\ &+ \mathbb{P}(I_r I_t = 1|G = BB)\mathbb{P}(G = BB)\\ &= 0 + \frac{1}{2}\cdot\frac{1}{2}\cdot 2x_j^i(1 - x_j^i) + 1\cdot {x_j^i}^2 = \frac{1}{2}x_j^i(1 - x_j^i) + {x_j^i}^2.\end{aligned}$$

Considering that the sum over $r < t$ in equation (12) is made over $\frac{1}{2}n_{j,\ell}^i(n_{j,\ell}^i - 1)$ equal expectations, we can write

$$\begin{aligned}\mathbb{V}[Y_{j,\ell}^i] &= \frac{1}{(n_{j,\ell}^i)^2}[n_{j,\ell}^i x(1-x) + 2\frac{n_{j,\ell}^i(n_{j,\ell}^i - 1)}{2}\frac{1}{2}x_j^i(1 - x_j^i)]\\ &= \frac{1}{(n_{j,\ell}^i)^2}[n_{j,\ell}^i x_j^i(1-x_j^i) + 2\frac{n_{j,\ell}^i(n_{j,\ell}^i - 1)}{2}\frac{1}{2}x_j^i(1 - x_j^i)]\\ &= \frac{n_{j,\ell}^i + 1}{2n_{j,\ell}^i}x_j^i(1 - x_j^i) = R_{j,\ell}^i x_j^i(1 - x_j^i),\end{aligned}$$

where for practical purposes we have defined, for each ℓth individual, $R_{j,\ell}^i$ as the ratio

$$\frac{n_{j,\ell}^i + 1}{2n_{j,\ell}^i}.$$

Consider at this point the approximation of the variance of the weighted "pseudo-individual", having estimated frequency $\hat{q}_j^i := \sum_{\ell=1}^{N_j} w_{j,\ell}^i \cdot \hat{x}_{j,\ell}^i$.

$$\mathbb{V}[\hat{x}_j^i] = \sum_{\ell=1}^{N_j}(w_{j,\ell}^i)^2\mathbb{V}[\hat{x}_{j,\ell}^i] \approx \sum_{\ell=1}^{N_j}(w_{j,\ell}^i)^2\mathbb{V}[Y_{j,\ell}^i]. \qquad (13)$$

Our objective is to perform a Lagrange-constrained optimization w.r.t. the weights, being sure to find a minimum since equation (13), as function of the weights, is convex. This is easily done since the Lagrange-parametrized function is

$$\mathcal{L}(w_{j,1:N_j}^i, \lambda) = \sum_{\ell=1}^{N_j}(w_{j,\ell}^i)^2 x_j^i(1 - x_j^i)R_{j,\ell}^i - \lambda(\sum_{\ell=1}^{N_j} w_{j,\ell}^i - 1)$$

and it originates a linear system of equations of the form

$$\begin{aligned} 2\cdot w_{j,1}^i &\quad \cdot x_j^i(1-x_j^i)R_{j,1}^i &&-\lambda = 0\\ \vdots &&&\vdots = \vdots\\ 2\cdot w_{j,N_j}^i &\cdot\ x_j^i(1-x_j^i)R_{j,N_j}^i &&-\lambda = 0\\ \sum_{\ell=1}^{N_j} w_{j,\ell}^i &&&-1 = 0\end{aligned}$$

whose solution provides us with the minimum values of the weights as follows $\forall \ell \in \{1, \ldots, N_j\}$:

$$w_{j,\ell}^i = \frac{\prod_{m=1,m\neq\ell}^{N_j} R_{j,m}^i}{\sum_{k=1}^{N_j}\prod_{m=1,m\neq k}^{N_j} R_{j,m}^i} = \frac{(R_{j,\ell}^i)^{-1}}{\sum_{k=1}^{N_j}(R_{j,k}^i)^{-1}}.$$

Appendix 3

Error estimation and correction. Estimation of the type-specific errors follows the supplementary material of (Orlando *et al.* 2013). Assume having one observed sequenced individuals affected by base-transition errors. This individual has an associated 4×4 error matrix $\boldsymbol{e}$, such that the entry $\boldsymbol{e}(a, b)$ is the probability of observing a base of type b when the true base is of type a. Consider the tree ((T,R),O), in which the leaves are sequenced genomes affected by type-specific errors (T), an individual without errors, used as reference for the error correction (R), and an outgroup individual (O).

Assume that loci are independent and that the errors between pairs of alleles are independent given a base o in the outgroup and the error matrix $\boldsymbol{e}$. Then the likelihood of the base t in the observed individual can be decomposed as a product through the loci:

$$\mathbb{P}(T = t|O = o, \boldsymbol{e}) = \prod_{i=1}^{M} \mathbb{P}(T_i = t_i|O_i = o_i, \boldsymbol{e}).$$

Marginalize any ith factor of the above equation over the true alleles before error $g_i \in \{A, C, G, T\}$ of the underlying true genotype:

$$\begin{aligned}\mathbb{P}(T_i = t_i|O_i = o_i, \boldsymbol{e}) &= \sum_{g_i \in \{A,C,G,T\}} \mathbb{P}(T_i = t_i, G_i = g_i|O_i = o_i, \boldsymbol{e}) \\ &= \sum_{g_i \in \{A,C,G,T\}} \mathbb{P}(T_i = t_i|G_i = g_i, O_i = o_i, \boldsymbol{e}) \mathbb{P}(G_i = g_i|O_i = o_i) \\ &= \sum_{g_i \in \{A,C,G,T\}} \boldsymbol{e}(g_i, t_i) \mathbb{P}(G_i = g_i|O_i = o_i),\end{aligned}$$

where the true genotype g_i is independent of the error rates for each $i = 1, \ldots, M$. One can approximate the probability of observing g_i conditionally to o_i with the relative frequency of the base g_i in the error-free individual R, for loci where the outgroup is o_i, that is

$$\mathbb{P}(G_i = g_i|O_i = o_i) = \mathbb{P}(R_i = g_i|O_i = o_i).$$

It is possible to perform a maximum likelihood estimation by numerical optimization to obtain an estimate of the error matrix. Note that every entry $\boldsymbol{e}(g_i, t_i)$ is the same over all loci.

The rationale behind the error correction is that the count of each base in the genomes T and R should be the same, otherwise an excess of counts in T is due to error.This approach to error estimation has been applied in (Orlando *et al.* 2013) to study type-specific errors in ancient horses' genomes.

Assume that the error matrix $\boldsymbol{e}_\ell$ has been estimated for every individual ℓ in each jth group. For a specific genome ℓ we have the following equation for each locus i

$$\begin{aligned}\mathbb{P}(T_i = t_i|\boldsymbol{e}_\ell) = \mathbb{P}(T_i = t_i|\boldsymbol{e}_\ell, G \to t_i) \boldsymbol{e}_\ell(t_i, t_i) \\ + \textstyle\sum_{\tilde{t}_i \neq t_i} \mathbb{P}(T_i = t_i|\boldsymbol{e}_\ell, G = \tilde{t}_i) \boldsymbol{e}_\ell(\tilde{t}_i, t_i).\end{aligned}$$

The same equation can be expressed in matrix form as follows:

$$\boldsymbol{p}_T^i = \boldsymbol{e}_\ell \boldsymbol{p}_G^i,$$

where $\boldsymbol{p}_T^i$ and $\boldsymbol{p}_G^i$ are the vectors of probabilities of observing alleles at locus i, respectively in the T and R genome. If the error matrix e_ℓ is invertible, we can find the error corrected allele frequencies as

$$\boldsymbol{p}_G^i = \boldsymbol{e}_\ell^{-1} \boldsymbol{p}_T^i. \tag{14}$$

The correction performed in equation (14) makes the estimated allele frequencies unbiased. The unbiasedness allows the numerator of the D-statistic to have mean zero, and makes the D-statistic calculated with error-corrected frequencies convergent to a standard normal distribution (see Appendix 1). In fact, consider for a certain locus the di-allelic scenario with alleles A and B. Let n be the number of observed bases. The number of alleles A in absence of errors is

$$m \sim Bin(n, x),$$

where x is the population frequency. Let $\epsilon_{A,B}$ and $\epsilon_{B,A}$ be the probabilities of having a transition from A to B and from B to A, respectively. Then the total number of observed A alleles is given by the sum of the two following variables:

$$\begin{aligned}m_0 &\sim Bin(m, 1 - \epsilon_{A,B}), \\ m_1 &\sim Bin(n - m, \epsilon_{B,A}).\end{aligned}$$

The expected population frequency is given by

$$\begin{aligned}\frac{1}{n}\mathbb{E}[m_0 + m_1] &= \frac{1}{n}\mathbb{E}\big[\mathbb{E}[m_0|m]\big] + \frac{1}{n}\mathbb{E}\big[\mathbb{E}[m_1|m]\big] \\ &= x(1 - \epsilon_{A,B}) + (1 - x)\epsilon_{B,A}.\end{aligned}$$

The error matrix and its inverse for the di-allelic case are expressed as follows:

$$\boldsymbol{e} = \begin{bmatrix} 1 - \epsilon_{A,B} & \epsilon_{B,A} \\ \epsilon_{A,B} & 1 - \epsilon_{B,A} \end{bmatrix}, \qquad \boldsymbol{e}^{-1} = \frac{1}{C}\begin{bmatrix} 1 - \epsilon_{B,A} & -\epsilon_{B,A} \\ -\epsilon_{A,B} & 1 - \epsilon_{A,B} \end{bmatrix},$$

where $C = (1 - \epsilon_{A,B})(1 - \epsilon_{B,A}) - \epsilon_{A,B}\epsilon_{B,A}$ is the constant arising from the inversion of a 2×2 matrix.

The formula in equation (14) is rewritten as

$$\begin{bmatrix} \hat{x} \\ 1 - \hat{x} \end{bmatrix} = \frac{1}{C}\begin{bmatrix} 1 - \epsilon_{B,A} & -\epsilon_{B,A} \\ -\epsilon_{A,B} & 1 - \epsilon_{A,B} \end{bmatrix}\begin{bmatrix} \hat{z} \\ 1 - \hat{z} \end{bmatrix}, \tag{15}$$

where $\hat{x}$ is the estimator of the error-corrected population frequency, while $\hat{z}$ is the estimated population frequency prior to error correction:

$$\hat{z} = \frac{m_0 + m_1}{n}.$$

From equation (15) it is possible to deduce the following equality:

$$\begin{aligned}\mathbb{E}[\hat{x}] &= \frac{1}{C}(1 - \epsilon_{B,A})\mathbb{E}[\hat{z}] - \frac{1}{C}(1 - \mathbb{E}[\hat{z}])\epsilon_{B,A} \\ &= \frac{1}{C}x(1 - \epsilon_{B,A} - \epsilon_{A,B}) = x.\end{aligned}$$

This proves that the error-corrected estimators of the allele frequencies are again unbiased, therefore calculating the D-statistic using error-corrected allele frequencies leaves the convergence results unchanged.

LITERATURE CITED

Alexander, D. H., J. Novembre, and K. Lange, 2009 Fast model-based estimation of ancestry in unrelated individuals. Genome Research .

Altshuler, D., R. Durbin, G. Abecasis, D. Bentley, A. Chakravarti, *et al.*, 2010 A map of human genome variation from population-scale sequencing. NATURE **467**: 1061–1073.

Black, J. S., M. Salto-Tellez, K. I. Mills, and M. A. Catherwood, 2015 The impact of next generation sequencing technologies on haematological research - a review. Pathogenesis **2**: 9–16.

Busing, F. M. T. A., E. Meijer, and R. V. D. Leeden, 1999 Delete-m jackknife for unequal m. Statistics and Computing **9**: 3–8.

Chatters, J. C., 2000 The recovery and first analysis of an early holocene human skeleton from kennewick, washington. American Antiquity **65**: 291–316.

Consortium, I. H., 2003 The international hapmap project. Nature **426**: 789–796.

Ewing, G. and J. Hermisson, 2010 Msms: a coalescent simulation program including recombination, demographic structure and selection at a single locus. Bioinformatics **26**: 2064–2065.

Green, R. E., J. Krause, A. W. Briggs, T. Maricic, U. Stenzel, *et al.*, 2010 A draft sequence of the neandertal genome. Science **328**: 710–722.

Johnson, O., 2004 *Information Theory And The Central Limit Theorem*. Imperial College Press.

Kent, W., C. Sugnet, T. Furey, K. Roskin, T. Pringle, *et al.*, 2002 The human genome browser at ucsc. Genome Res. **12**: 996–1006.

Lalueza-Fox, C. and M. T. P. Gilbert, 2011 Paleogenomics of archaic hominins. Current Biology **21**: R1002–R1009.

Lamperti, J. W., 1996 *Probability: A Survey of the Mathematical Theory, Second Edition*. John Wiley & Sons.

Li, Y., N. Vinckenbosch, G. Tian, E. Huerta-Sanchez, T. Jiang, *et al.*, 2010 Resequencing of 200 human exomes identifies an excess of low-frequency non-synonymous coding variants. Nature Genetics **42**: 969–972 IF:35.209.

Meyer, M., M. Kircher, M.-T. Gansauge, H. Li, F. Racimo, *et al.*, 2012 A high-coverage genome sequence from an archaic denisovan individual. Science **338**: 222–226.

Nielsen, R., J. Paul, A. Albrechtsen, and Y. Song, 2011 Genotype and snp calling from next-generation sequencing data. Nature Reviews. Genetics **12**: 443–451.

Orlando, L., A. Ginolhac, G. Zhang, D. Froese, A. Albrechtsen, *et al.*, 2013 Recalibrating Equus evolution using the genome sequence of an early Middle Pleistocene horse. Nature **499**: 74–78 IF:38.597.

Patterson, N. J., P. Moorjani, Y. Luo, S. Mallick, N. Rohland, *et al.*, 2012 Ancient admixture in human history. Genetics .

Pesaran, M. H., 2015 *Time Series and Panel Data Econometrics*. Oxford University Press.

Pickrell, J. K. and J. K. Pritchard, 2012 Inference of population splits and mixtures from genome-wide allele frequency data. PLoS Genet **8**: 1–17.

Pritchard, J., M. Stephens, and P. Donnelly, 2000 Inference of population structure using multilocus genotype data. Genetics **155**: 945–959.

Raghavan, M., M. DeGiorgio, A. Albrechtsen, I. Moltke, P. Skoglund, *et al.*, 2014 The genetic prehistory of the New World Arctic. Science **345**.

Raghavan, M., P. Skoglund, K. E. Graf, M. Metspalu, A. Albrechtsen, *et al.*, 2013 Upper Palaeolithic Siberian genome reveals dual ancestry of Native Americans. Nature **505**: 87–91.

Raghavan, M., M. Steinrücken, K. Harris, S. Schiffels, S. Rasmussen, *et al.*, 2015 Genomic evidence for the pleistocene and recent population history of native americans. Science .

Rasmussen, M., S. Anzick, M. Waters, P. Skoglund, M. DeGiorgio, *et al.*, 2014 The genome of a late pleistocene human from a clovis burial site in western montana. Nature **506**: 225–229.

Rasmussen, M., Y. Li, S. Lindgreen, J. S. Pedersen, A. Albrechtsen, *et al.*, 2010 Ancient human genome sequence of an extinct Palaeo-Eskimo. Nature **463**: 757–762.

Reich, D., R. E. Green, M. Kircher, J. Krause, N. Patterson, *et al.*, 2010 Genetic history of an archaic hominin group from Denisova cave in Siberia. Nature **468**: 1053–1060.

Reich, D., N. Patterson, M. Kircher, F. Delfin, M. Nandineni, *et al.*, 2011 Denisova Admixture and the First Modern Human Dispersals into Southeast Asia and Oceania. The American Journal of Human Genetics **89**: 516–528.

Reich, D., K. Thangaraj, N. Patterson, A. Price, and L. Singh, 2009 Reconstructing indian population history. Nature **461**: 489–494.

Skoglund, P., S. Mallick, M. C. Bortolini, N. Chennagiri, T. Hünemeier, *et al.*, 2015 Genetic evidence for two founding populations of the Americas. Nature **525**: 104.

Slutsky, E., 1925 Uber stochastische asymptoten und grenzwerte. Internationale statistische Zeitschrift **5**: 3–89.

Stoneking, M. and J. Krause, 2011 Learning about human population history from ancient and modern genomes. Nature Reviews **12**.

Wall, J. D., M. A. Yang, F. Jay, S. K. Kim, E. Y. Durand, *et al.*, 2013 Higher levels of neanderthal ancestry in east asians than in europeans. Genetics .

Supplemental Material.

The Supplemental Material contains two tables with numeric results related to a real data scenario, and five figures regarding the power of the method, the asymptotic behaviour of D_{ext}, the estimates of type-specific errors, the behaviour of the D-statistic and the correction for external introgression.

Table S1. European Introgression into Native American Individuals. The table contains the values of the different types of D-statistics used to create the plot of Figure 4C, reporting the D-statistic for the tree (((PEL,CHB)CEU)YRI). The first column denote if we are illustrating either the extended D-statistic, D_{ext}, or the D-statistic that uses a sampled base, D_{1base}. The column denoted by D is the D-statistic over all blocks of loci, used to estimate the standard deviation (third column) by bootstrapping. The Z-score represents the D-statistic normalized by its standard deviation. The last column represents the ratio between the estimated standard deviations of D_{1base} and D_{ext}.

D-statistic	D	stdev(D)	Z-score	$\frac{\sigma_{1base}}{\sigma_{ext}}$
D_{ext}	-0.032638	0.002449	-13.114101	-
D_{1base}	-0.038171	0.006164	-6.223641	2.51
D_{1base}	-0.032786	0.006244	-5.253267	2.54
D_{1base}	-0.030950	0.006708	-4.602315	2.74
D_{1base}	-0.038730	0.006480	-5.999972	2.64
D_{1base}	-0.033640	0.006244	-5.353646	2.55

Table S2. Estimated Error Rates. Estimated type-specific error rates for the ancient individuals Saqqaq and Canadian Dorset Mi'kmaq used in the tree of Figure 3B.

Individual	$A \to C$	$A \to G$	$A \to T$	$C \to A$	$C \to G$	$C \to T$
Saqqaq	1.90e-04	6.08e-04	3.27e-04	7.52e-04	1.22e-04	6.32e-04
Dorset	8.86e-05	1.15e-03	1.62e-04	2.04e-04	8.52e-05	5.22e-03
	$G \to A$	$G \to C$	$G \to T$	$T \to A$	$T \to C$	$T \to G$
Saqqaq	6.35e-04	1.26e-04	7.52e-04	3.28e-04	6.08e-04	1.91e-04
Dorset	5.21e-03	9.01e-05	2.06e-04	1.64e-04	1.15e-03	9.04e-05

Table S3. Extended D-Statistic in Real Data Scenario with Ancient Genomes. Table comparing the extended D-statistic with the application of error correction and/or transition removal for the tree of Figure 5B, where the ancient individuals Saqqaq and Canadian Dorset Mi'kmaq are affected by high type-specific error rates.

Correction	D_{ext}	$sd(D_{ext})$	$Z-score$	$p-value$
None	-5.26e-2	5.4e-3	-9.81	0
Trans.Rem.	1.01e-2	7.1e-3	1.41	1.57e-1
Error.Corr.	5.64e-3	6.1e-3	0.93	3.51e-1
Err.Corr & Tr.Rem	8.77e-4	7.3e-3	0.12	9.04e-1

Figure S1. Effect of the number of individuals per population in detecting admixture. Results from the simulation of the scenario of Figure 2A, subject to a migration from H_3 to H_1, using either 1, 2, 5, 10 or 20 individuals per population sequenced at depth 0.2X. (A) Power of the extended D-statistic for increasing values of the number of individuals per group. (B) The value of the standard deviation of D_{ext} for different number of individuals per population.

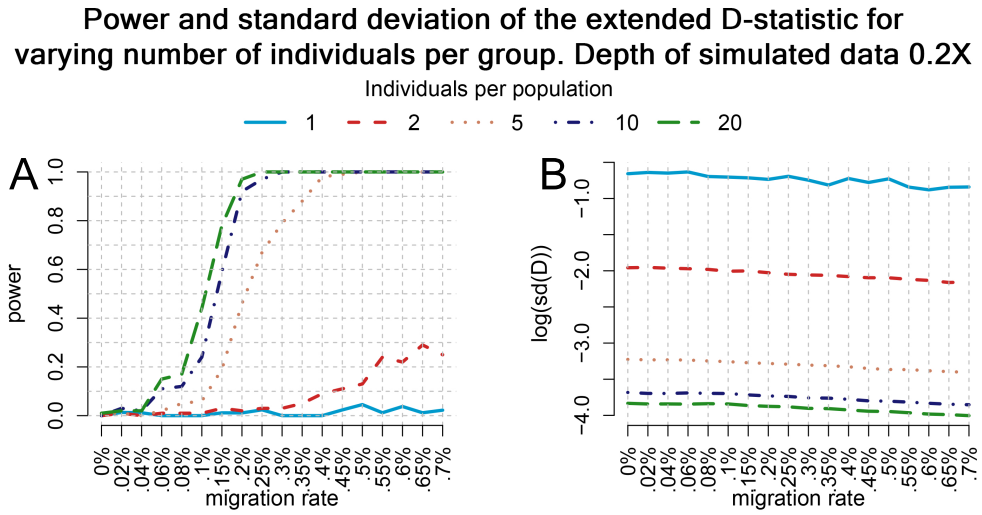


Figure S2. Asymptotic convergence of the extended D-statistic. QQ-plot of the observed log-pvalues from 5000 simulations of the null hypothesis of Figure 2B, where we have used 5 individuals per population and depth 2X. Each individual has 200 regions of length 5Mb. Despite that, the extended D-statistic D_{ext} shows already good properties of asymptotic convergence to the standard normal, with a slight problem due to few extreme pvalues.

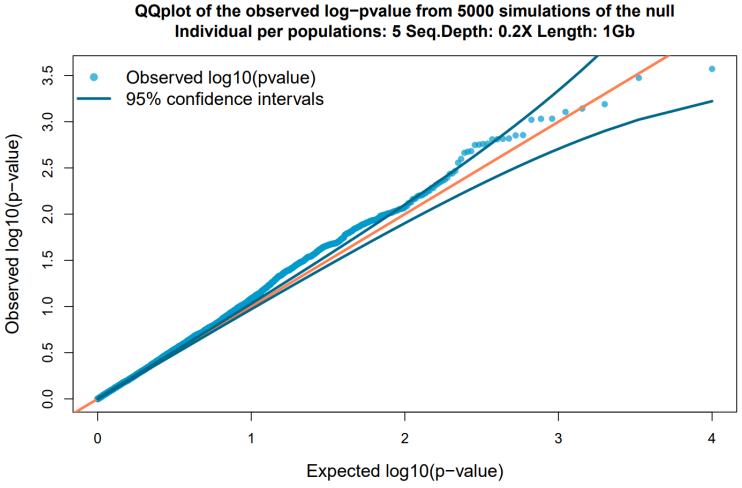


Figure S3. Subtrees of interest in a scenario subject to external introgression. (A) Case of a 4-population tree subject to introgression from an external population H_5. Consider H_2 being the population subject to introgression from H_5. (B) The subtree $T_{1:4}$ includes the 4-population tree excluding the admixing population. (C) The subtree T_{out} replaces the admixed population with the population source of introgression. (D) The subtree T_{un}, where H_2' represents H_2 when it has not yet undergone admixture, reflects the null hypothesis of correctness for the genetic relationship between four populations.

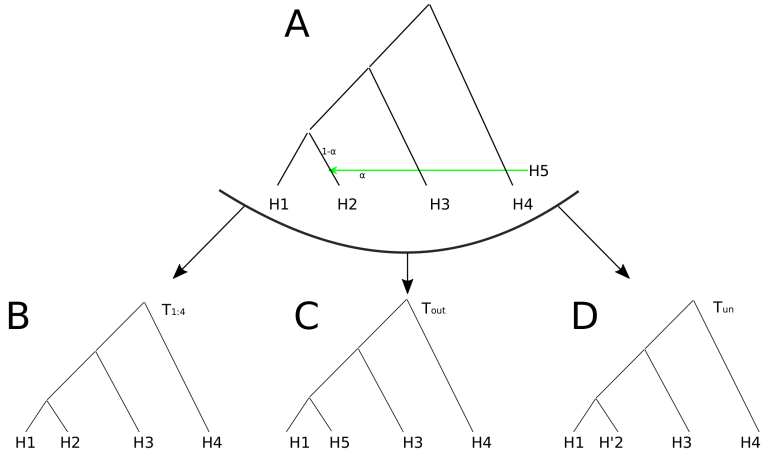


Figure S4. Estimates of Type-Specific Errors for Ancient Genomes. Estimated type-specific error rates for the Saqqaq, Mi'qmak and French genomes of the real data scenario illustrated in Fig 4B.

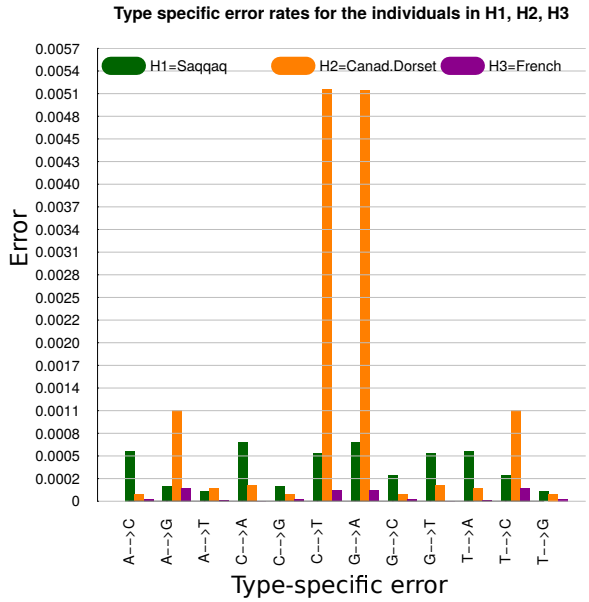
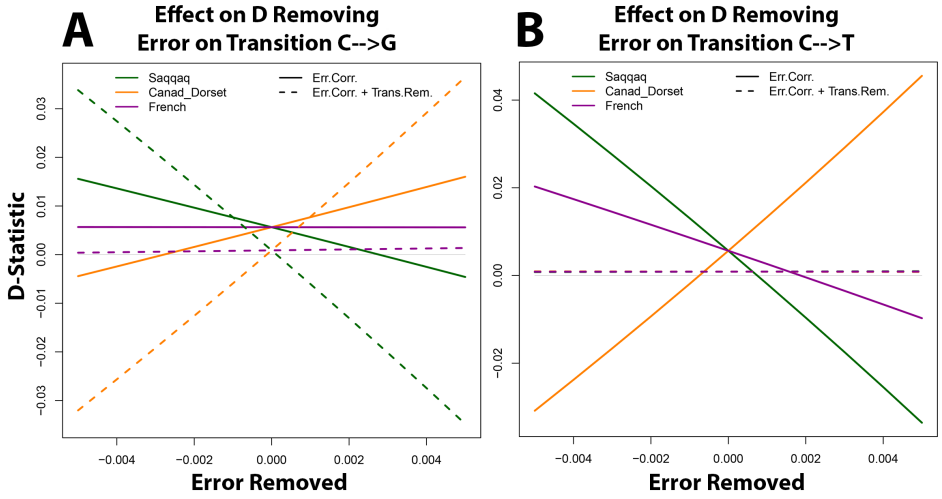


Figure S5. Behaviour of the D-Statistic in Function of the Type-Specific Error. Effect of increasing and decreasing the removal of error for the base transitions $C \to G$ and $C \to T$ for one of the Greenlandic Saqqaq, Canadian Dorset and French genomes. This corresponds to the addition of a value in the entry $\boldsymbol{e}(G, C)$ or $\boldsymbol{e}(T, C)$ of the estimated error matrix of one of the individuals, as if the estimated error rate was higher or lower. In solid lines are represented the values of D_{ext} for which the correction is performed. The dashed lines represent the analogous values where ancient transitions are not considered.



Manuscript 2

Background Theory for Admixture Graphs and F-statistics

Samuele Soraggi, Carsten Wiuf

Status: prepared for submission in Bulletin of Mathematical Biology

Contribution

In this manuscript the admixture graphs and the F-statistics (F_2, F_3 and F_4) are analyzed in a mathematical framework under the point of view of applications in population genetics. In fact, admixture graphs are at the basis of many computational tools for inferring or testing for gene flow [13, 20, 21], but their properties have not been formalized.

Here, formal definitions and proofs of properties for the admixture graphs and the F-statistics are provided. It is possible to relate some topological properties of the graphs to admixture rates and paths between nodes, and the renowned graphical method to calculate the F-statistics [13] is proven as a consequence of this theory.

Moreover, the relationship between this background theory and population genetics are highlighted in the formalization of the drifts and their role in defining the F-statistics. For the F_2-statistics, a canonical decomposition related to the graph topology and a theorem with minimal condition for their linear independence are proven.

Future perspectives

The results in this manuscript are related to the studies applying admixture graphs and F-statistics for inference/test of genetic relationships between populations [13, 20, 21]. However, assumptions and intuitions on the graphs and the F-statistics used in those studies are not always proven or fulfilled. The formalization for the necessary hypothesis that lead to such properties are given in this manuscript and could be implemented in the computational tools to perform a preliminary test on the topology.

Other interesting directions can be explored in relation to admixture graphs and F-statistic. For example, one could look into the possibility of implementing further F-statistic apart from the current ones (F_2, F_3 and F_4).

Moreover, the F_2 is a metric under some specific conditions. This fact connects to the topic of split theory [73, 74]. Here, a metric is decomposed as the sum of weighted metrics on subgraphs called splits and a residual term. Such a decomposition is not trivial because of the relation between F_2, the graph topology and the admixture rates, but it is an interesting development of the theory.

A fundamental result of this manuscript is the set of conditions for the linear independence of F_2-statistics. Here the result holds for a graph with two potential roots (ancient populations with unknown genetic relationship). it is still necessary to study the possibility of proving a similar theorem for an arbitrary number of roots.

Noname manuscript No.
(will be inserted by the editor)

Background theory for admixture graphs and F-statistics

Samuele Soraggi · **Carsten Wiuf**

Received: date / Accepted: date

Abstract The widespread availability of genome data for many organisms - including humans - has lead to a deeper understanding of the genetic relationships between populations. An important role in inferring and testing such relationships is played by model-based methods, where the evolutionary history of populations is modeled through graphs or networks, based on which a mathematical formulation of the problem can be expressed.

In particular, the admixture graph has become popular in methods to infer and test complex reticulates involving complex histories of populations. Most recent methods are based on moment statistics called F-statistics. However, a formal mathematical formulation of the admixture graphs and the F-statistics and their properties has been lacking.

The goal of this paper is to provide a background mathematical theory where the admixture graphs are defined, and their properties formally demonstrated. Applying the theory of chain graphs, the properties of the F-statistics are deduced in a stochastic framework. Assumptions and motivations for the population genetics framework are analyzed, and some examples from applications in population genetics are studied.

Keywords Admixture · Gene Flow · Admixture Graph · F-statistic

1 Introduction

The inference of demographic history from a genetic perspective, that is the study of gene flow and introgression between populations, the assessment of migrations and

Samuele Soraggi
Department of Mathematics,
Universitetsparken 5, 2200 Copenhagen, DK
E-mail: samuele@math.ku.dk

Carsten Wiuf
Universitetsparken 5, 2200 Copenhagen, DK
E-mail: wiuf@math.ku.dk

admixture or splitting of populations using genetic data, has been a topic of wide interest in population genetics [3,4,19,7,17,22,21] since the early availability of genetic data.

Inferring information on the past history of populations have been a challenge for population geneticists. Early traditional population genetics methods are based on comparing the expected value of genetic statistics under demographic and mutation scenarios, such as heterozygosity, to their value calculated from genetic data [15]. In this way, it is possible to infer information about the past history of a population. For example, the study of variations in population size can be indirectly informative on past migrations [25]. With the advent of Kingman's theory of the coalescent, and the possibility of genetic simulations [11,10], the focus on populations' history have increased and lead to new inference techniques based on the MCMC framework or likelihood-based approaches [16,8,25].

With the development of high-throughput techniques [14,20] such as NGS, scientists have been provide with large amounts of data, with the potential for providing much more informations. However, computational methods that are computationally performant and model complex populations' history are needed. A possible way to describe such complex genetic relationships between populations is through graphs or network, where each node represents a population. In term of data, those models associate a genetic characteristic, such as allele frequency, to each node. In this way, one is also able to bypass mutation-based models, that are not reliable on relatively short time periods, due to the low frequency of mutations.

Two first attempt to describe past histories of populations with a graph is through the phylogenetic tree [3,4] and the admixture graph [19,17,18]. A phylogenetic tree describes the evolutionary relationship between a set of populations admitting only splits giving rise to two descendants. Distinct nodes cannot be merged, therefore a phylogenetic tree does not describe gene flow or migrations. A more complex reticulate of relationships is described with the admixture graphs. Those admit gene flow between populations. In such a model more populations can merge and generate an admixed population [23,5,17,19]. Even though an admixture graph is still a simplified model of a more complex genetic history, it is able to describe more complex scenario compared to a phylogenetic tree.

Many computational model-based methods have been developed to infer demographic histories through admixture graphs. A first example is the tool `qpgraph` [2], where the authors use a heuristic method to exclude unlikely edges, by building specific subgraphs denoted as qp-graphs. The software `AdmixTools` [17] formulates the relationships of an user-defined graph in terms of quantities called F-statistics. The F-statistics are calculated from allele frequencies, and in the software `Admixtools` are used to define a system of equations from which admixture parameters can be inferred, and the graph can be tested for fitness to the data. The softwares `TreeMix` [1] and `MixMapper` [13] first build a graph without admixtures. Thereafter they apply different techniques to add best-fitting admixture branches according to the data, and solving equations in the same way as in `Admixtools`.

The methods `TreeMix`, `AdmixTools` and `MixMapper` use the F-statistics as main tool in their implementations. The F-statistics have been a particularly successful in population genetics [19,7,17,18], since they allow a greater computational efficiency

in graph-based methods, compared for example to earlier studies based on computationally intensive likelihood optimizations [5,23], that limit the applications to small sets of populations. The F-statistics are three parameters defined between two, three and four populations, respectively, in an admixture graph. The interpretation of the F-statistics is based on the analogy between common branches on paths between populations and shared amount of genetic drift that characterizes such populations [19, 7,17]. Other possible interpretations consider the F-statistics in term of expected coalescent times between populations, covariances between population frequencies and heterozygozities [17,18] under specific model topologies.

The mentioned softwares and interpretations lack a formal treatment and analysis of the properties of the admixture graphs and the F-statistics. Nonetheless they apply a wide range of assumptions by characterizing the F-statistics through admixture proportions and a specific type of paths between nodes of the admixture graph. The relationships between admixture graphs and F-statistics is essential in many of those assumptions and turn out to be as important in analyzing the F-statistics [18,17].

In this paper, the goal is to provide and analyze a formal mathematical background for the admixture graphs and their properties. In this way it is possible to motivate and extend the interpretations and definitions considered in the current literature. The definition of a stochastic structure through the Markov chain graphs [6, 12,24] allows to study in depth the F-statistics and to find fundamental results for current applications. For example, the graphical method to calculate the F_2-statistics [17,13,19], the additivity of the F_2-statistics and their linear independence [17,13] are are proven in the theoretical framework of this paper. Those are at the base of the methods used to infer or test admixture graphs, because they connect the topology of the graph and the genetic distances between populations through equations involving admixture proportions. Finally, connections with the interpretation of the theory in terms of population genetics is provided through examples related to applications of the three- and four-populations test.

2 Admixture Graphs

In this section, admixture graphs are defined. We consider labeled graphs with directed and undirected edges, and use the notations $i \leftrightarrow j$ (equivalently $j \leftrightarrow i$) and $i \rightarrow j$ (equivalently $j \leftarrow i$) for an undirected edge between i and j, and a directed edge from i to j, respectively. An edge $i \rightarrow j$ is said to be ingoing to j and outgoing of i. The undirected egde's notation is symmetric but we consider its two associated roots as an ordered pair according some criteria. In what follows we consider two roots i, j ordered as (i, j), where $i < j$, and the nodes of a directed edge $i \rightarrow j$ ordered as (i, j). This will also be the order used whenever the nodes are indices in the notation. For brevity we will also use the alternative notation e for an edge of type $i \rightarrow j$ or for an undirected edge $i \leftrightarrow j$. The set $par(i)$ denotes the parents of j, that is, $par(j) = \{i \mid i \rightarrow j\}$.

Definition 1 (Admixture graph) An admixture graph is an edge labeled graph $\mathscr{G} = (\mathscr{V}, \mathscr{E}, \mathscr{L})$ without directed cycles. The triplet consists respectively of the set of nodes, edges and labels. The set of nodes $\mathscr{V}$ is divided into:

- *roots* $\mathscr{R}$, nodes without ingoing edges. All pairs of roots are connected by an undirected edge and only these,
- *admixed nodes* $\mathscr{A}$, nodes that have ingoing directed edges,
- *leaves* $\mathscr{A}_0 \subseteq \mathscr{A}$, admixed nodes without outgoing directed edges.

An edge between two roots $r_1, r_2 \in \mathscr{R}$ has label $\alpha_{r_1 r_2} = 1$. For labels between $par(j)$ and $j \in \mathscr{A}$ we assume

$$\sum_{i \in par(j)} \alpha_{ij} = 1,$$

where $\alpha_{ij} \in (0,1]$ denotes the label of the edge $i \to j$. We will often denote $\alpha_e = \alpha_{ij}$ if e involves nodes i, j.

By definition, the graph is connected. In the following, we assume that an admixture graph is not trivial, meaning that it does not consist of only roots and undirected edges. To keep the notation uncluttered, we do not put any order in the two indices of a label, so $\alpha_{ji} = \alpha_{ij}$. See Figure 1 for examples.

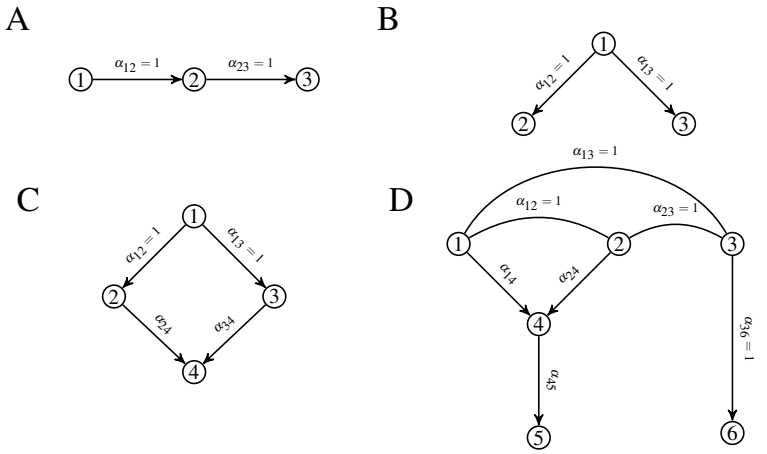


Fig. 1 Examples of admixture graphs. (A) An admixture graph where node 1 is the root, 2 an admixed node and 3 a leaf. All edges' labels are equal to one. (B) An admixture graph where node 2 and 3 are leaves. (C) An admixture graph where node 4 is a leaf with two parents. (D) An admixture graph with three roots and two leaves.

Definition 2 (Admixture path between two nodes) Given an admixture graph $\mathscr{G}$ and two nodes $i, j \in \mathscr{V}$, $i \neq j$, an admixture path (or just path) γ from i to j is an ordered sequence of nodes that starts i and ends in j such that

$$(i_k, i_{k-1}, \ldots, i_1, i_0, i'_0, i'_1, \ldots, i'_{k'-1}, i'_{k'}),$$

with no nodes being repeated and $i_k = i$, $i'_{k'} = j$, where $k, k' \geq 0$. Two adjacent nodes i_m, i_{m-1} are connected by an edge $i_m \leftarrow i_{m-1}$ for $m = 1, \ldots, k$, and by $i_{m-1} \to i_m$ for $m = 1, \ldots, k'$. The case $i_0 \neq i'_0$ is admitted only if i_0, i'_0 are roots. If $k = 0$ then $i'_0 = i$, and if $k' = 0$ then $i_0 = j$. A path from i to j is denoted by $i \Rightarrow j$ and the set of such

paths is denoted by Γ_{ij}. A subpath γ' of a path γ, or with an abuse of notation $\gamma' \subseteq \gamma$, is an ordered sequence of nodes found in γ with the same order.
An edge $e = i \to j$ or $e = i \leftrightarrow j$ is in a path γ if its nodes are adjacent in γ. With a slight abuse of notation this is denoted by $e \in \gamma$. The sign of an edge e in a path γ, $\mathrm{sgn}_\gamma(e)$, has value $+1$ if the nodes of e have the opposite order than in γ, otherwise $\mathrm{sgn}_\gamma(e) = -1$.

The label p_γ of a path $\gamma \in \Gamma_{ij}$ is the product of labels

$$p_\gamma := \prod_{e \in \gamma} \alpha_e.$$

A path can at most contain two roots, in particular, a path from the root r_1 to the root r_2 consists of the roots themselves. A path $\gamma \in \Gamma_{ij}$ is not symmetric, meaning that it is not considered the same as the path $\gamma' \in \Gamma_{ji}$ composed by the edges of γ in the opposite order. Therefore $\Gamma_{ij} \neq \Gamma_{ji}$. Note that the labels of γ and γ' are identical.

Remark. Note that an admixture path $\gamma \in \Gamma_{ij}$ is not a path according to the standard definition of graphs, where a path is defined by following the direction of the edges [9]. An admixture path is defined through a sequence of ordered nodes whose connecting edges follow specific constraints. For example $(i_k, i_{k-1}, \ldots, i_1, i_0)$, where $k > 0$, $i_k = i$ and $i_0 = j$, is an admixture path from i to j with edges of type $i_m \leftarrow i_{m-1}$ for $m = 1, \ldots, k$.

Example 1 In Figure 1B, there is only one possible path between the nodes 2 and 3, defined by $\gamma = (2,1,3)$, with label $p_\gamma = \alpha_{12}\alpha_{13} = 1$. In Figure 1C, the path $\gamma = (2,1,3)$ is the only path of Γ_{23}, because the sequence $(2,4,3)$ does not fulfill Definition 2. In Figure 1C, $\gamma_1 = (3,1)$ and $\gamma_2 = (3,2,1)$ are the only two paths of Γ_{31}. Their labels are α_{13} and $\alpha_{23}\alpha_{12}$, respectively.

Proposition 1 *Consider two nodes $i, j \in \mathscr{V}$ of an admixture graph $\mathscr{G}$. Then $\Gamma_{ij} \neq \emptyset$. Further, the sum of the labels is one, $\sum_{\gamma \in \Gamma_{ij}} p_\gamma = 1$.*

In what follows we characterize - in terms of paths and labels - when an admixture graph is a tree or a forest with connected roots. Here a forest is a set of trees with roots connected by undirected edges.

Theorem 1 *For an admixture graph $\mathscr{G}$, the following statements are equivalent:*

1. *for each pair of nodes in $\mathscr{V}$, there is only one path γ connecting them,*
2. *every path γ on $\mathscr{G}$ has probability* 1,
3. *the admixture graph consists of a forest of R trees, where R is the number of roots, pairwise connected by undirected edges linking the roots.*

Definition 3 (Root weights of a node) Let $\ell \in \mathscr{A}$ be an admixed node and $r \in \mathscr{R}$ a root of an admixture graph $\mathscr{G}$. Let $\Omega_{\ell r} \subseteq \Gamma_{\ell r}$ be the set of paths from ℓ to r that do not contain another root. The root weight of r with respect to ℓ is the probability

$$q_{\ell r} = \sum_{\gamma \in \Omega_{\ell r}} p_\gamma.$$

Proposition 2 *Given an admixture graph, the root weights of the roots with respect to an admixed node form a probability distribution.*

If the admixture graph is a tree or a forest (in the sense of Theorem 1), then for each node ℓ and root r, the probability $q_{\ell r}$ is equal to 1 if the node is in the tree with root r, and otherwise $q_{\ell r} = 0$.

For a subset of nodes of an admixture graph, we consider the subgraph given by all paths connecting any two nodes of the subset.

Definition 4 (Admixture graph spanned by a subset of nodes) Let $\mathscr{G} = (\mathscr{V}, \mathscr{E}, \mathscr{L})$ be an admixture graph and let $C \subseteq \mathscr{V}$. We define the admixture graph spanned by C as the graph $\mathscr{G}_C = \big(\mathscr{V}_C, \mathscr{E}_C, \mathscr{L}_C\big)$, where

$$\mathscr{V}_C = \big\{ i \mid i \text{ is in a path of } \Gamma_{jk} \text{ for some } j,k \in C \big\},$$
$$\mathscr{E}_C = \big\{ e \mid e \in \mathscr{E} \text{ connects two nodes of } \mathscr{V}_C \big\},$$

and $\mathscr{L}_C$ is the set of labels inherited from $\mathscr{G}$. In particular, $\mathscr{G}_{\mathscr{V}} = \mathscr{G}$.

It is immediate to verify that the graph $\mathscr{G}_C$ defined above is an admixture graph.

Definition 5 (Operations on paths)

Given two paths γ_1, γ_2 on the same admixture graph, their intersection, denoted by $\gamma_1 \cap \gamma_2$ with an abuse of notation, is the set of nodes that appear in both paths.

Proposition 3 *Let $\mathscr{G}$ be an admixture graph and $C \subseteq \mathscr{V}$ a subset of the nodes. Let $\mathscr{G}_C$ be the admixture graph spanned by C and let C_0 be the leaves of $\mathscr{G}_C$. One of the following two equivalent conditions holds*

1. *for each node $k \in C \backslash C_0$, there is a pair $i, j \in C_0$ such that $\gamma \cap \delta = \{k\}$ for some $\gamma \in \Gamma_{ik}$, $\delta \in \Gamma_{kj}$,*
2. *for each node $k \in C \backslash C_0$, there is an admixture path from i to j that includes node k, for some $i, j \in C_0$,*

if and only if $\mathscr{G}_C = \mathscr{G}_{C_0}$.
Moreover C_0 is the smallest set spanning $\mathscr{G}_C$, meaning that any other set that spans $\mathscr{G}_C$ contains C.

Corollary 1 *If an admixture graph $\mathscr{G}$ is such that $\mathscr{V}$ fulfils the hypothesis of Proposition 3, then $\mathscr{G}$ is spanned by its leaves $\mathscr{A}_0$, that is, $\mathscr{G}_{\mathscr{A}_0} = \mathscr{G}$.*

3 Stochastic Admixture Graphs

In this section we will add a stochastic structure to an admixture graph, such that the graph encodes conditional independencies. Specifically, we will assume the admixture graph with the stochastic structure is a chain graph, which is a special type of Markov graphical model [12,6,24]. Conditional independencies of this form are typically assumed in models in population genetics.

We first define how an admixture graph can be divided into blocks, based on the maximum number of edges necessary to reach a node starting from one of the roots.

Definition 6 (Blocks of an admixture graph) The blocks of an admixture graph consist of a ordered sequence $B_1,\dots,B_N$, that forms a partition of $\mathscr{V}$. If two nodes i,j are connected by a path such that $(i,i'_1,\dots,i'_{k'-1},j)$ for some $i'_1,\dots,i'_{k'-1}\in\mathscr{V}$, then $i\in B_{n_i}$, $j\in B_{n_j}$ and $n_i<n_j$. If $i,j\in\mathscr{R}$ are roots, then they are in the same block B_1.

Definition 7 (Stochastic admixture graph) Let $\mathscr{G}=(\mathscr{V},\mathscr{E},\mathscr{L})$ be an admixture graph. Construct a new graph $\mathscr{G}^*=(\mathscr{V}^*,\mathscr{E}^*)$ by augmenting the node set

$$\mathscr{V}^*=\mathscr{V}\cup\big\{(i,j)\,|\,i\to j\in\mathscr{E}\big\},$$

and splitting all directed edges into two, leaving the undirected edges as they are. That is, for $i\to j\in\mathscr{E}$, create $i\to(i,j)\in\mathscr{E}^*$ and $(i,j)\to j\in\mathscr{E}^*$, and erase $i\to j$. A stochastic variable with finite mean is associated with each node in $\mathscr{V}^*$, denoted by V_j if $j\in\mathscr{V}^*$ and C_{ij} if $(i,j)\in\mathscr{V}^*$. The variables C_{ij} are called contribution variables and the nodes (i,j) contribution nodes.

The admixture graph $\mathscr{G}$ is said to be a stochastic admixture graph if

(i) $\mathscr{G}^*$ is a chain graph (see Appendix A for the precise definition)
(ii) $V_j=\sum_{i\in par(j)}\alpha_{ij}C_{ij}$ for any admixed node $j\in\mathscr{A}$
(iii) $E(C_{ij}|V_i)=V_i$ for any admixed node $i\in\mathscr{A}$, where $E(X|Y)$ denotes the conditional expectation of a variable X given a variable Y.

An example of a stochastic admixture graph is shown in Figure 2. Here and elsewhere, an equality between two stochastic variables is equality almost surely with respect to the underlying probability measure.

The Markov structure of the chain graph implies in particular that for two contribution variables $C_{ij},C_{k\ell}$, where V_i,V_k are not necessarily distinct, it holds that

$$C_{ij}\perp C_{k\ell}\,|\,\{V_i,V_k\}. \tag{1}$$

Further, for a contribution variable C_{ij}, let B_{n_i} be the block in which node i is located. Then

$$C_{ij}\perp\bigcup_{n=1}^{n_i}\bigcup_{j\in B_n}V_j|V_i. \tag{2}$$

As a consequence of (2), if a node k is an element of $\bigcup_{n=1}^{n_j}B_n$, it follows that

$$C_{ij}\perp V_k|V_i. \tag{3}$$

The property in Definition 7(iii) does not hold between distinct root variables, unless these are identical variables, as shown below.

Theorem 2 *Let $\mathscr{G}$ be an admixture graph, and $R_1,\dots,R_k$ the variables associated with the roots, assuming $Var(R_i)<+\infty$, $i=1,\dots,k$. Then $E(R_i\,|\,R_j)=R_j$ holds for any pair of roots if and only if $R_1=R_2=\cdots=R_k$.*

Definition 8 (Drifts) Consider an admixture graph $\mathscr{G}$ and a pair of nodes $i,j\in\mathscr{V}$. The drift between i and j is defined as the difference between the associated variables,

$$D_{ij}:=V_j-V_i.$$

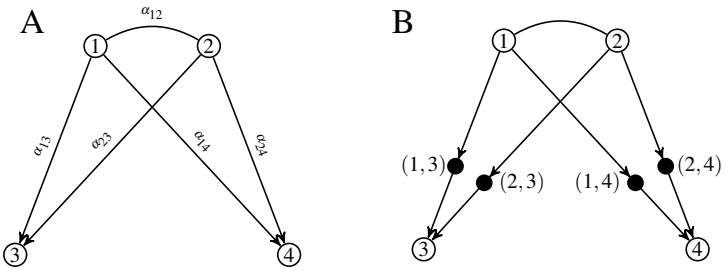


Fig. 2 Contribution nodes. (A) Admixture graph as in Definition 7. (B) The augmented graph derived from Figure A. The black dots represent the nodes associated to contribution variables.

Given an edge $e = k \to \ell$ or $e = k \leftrightarrow \ell$ with $k < \ell$, the partial drift of e is defined as the difference between the contribution variable from k to ℓ and the variable of the parent generating it:

$$d_e = d_{k\ell} := C_{k\ell} - V_k,$$

The partial drift of e on a path γ such that $e \in \gamma$ is defined as

$$d_e^\gamma = \mathrm{sgn}_\gamma(e) d_e.$$

Note that $D_{ji} = -D_{ij}$. In case e is undirected or when k is the only parent of ℓ, the partial drift coincides with the drift between k and ℓ.

Remark. The sign in a path γ of an undirected edge $i \leftrightarrow j \in \gamma$ is independent on the order chosen between the nodes.

We show that the drift between two nodes can be decomposed along the paths connecting the nodes as a linear function of the probabilities of such paths and of the partial drifts.

Theorem 3 (Canonical decomposition of a drift along paths) *Given $i, j \in \mathscr{V}$, the drift D_{ij} is the sum over Γ_{ji} of the probabilities of the paths multiplied by the sum of the partial drifts between subsequent nodes of each path, that is,*

$$D_{ij} = \sum_{\gamma \in \Gamma_{ji}} \left(p_\gamma \sum_{e \in \gamma} d_e^\gamma \right). \tag{4}$$

The Markov structure implies that the partial drifts are on average orthogonal to each other.

Proposition 4 *Consider two edges e_1, e_2, where at least one is directed. The product of their partial drifts is on average orthogonal in the sense that*

$$E(d_{e_1} d_{e_2}) = 0. \tag{5}$$

Note that the same statement holds for the partial drifts $d_{e_1}^{\gamma_1}, d_{e_2}^{\gamma_2}$ in any pair of paths γ_1, γ_2.
The same statement does not apply for two undirected edges. In fact, in this case it is not possible to use conditional independencies to make two partial drifts along distinct undirected edges orthogonal. In terms of chain graphs, this happens because the roots form a chain component.

4 F-statistics

This section defines the F-statistics F_2, F_3 and F_4, and gives various results for these. The F_2-statistic describes the distance between two nodes as the averaged squared difference of the drift. It is often assumed that the F_2-statistic is additive [17,19, 18]. We give conditions under which additivity holds. Further, we show that the F_2-statistics form a basis of a vector space [17]. Lastly some specific models that involve the F_3- and F_4-statistics to infer the presence of populations admixture in population genetics [19,7,17,18] are analyzed and commented.

Definition 9 **(F_2-statistic)** Let i, j in $\mathscr{V}$. The F_2-statistic between i and j is defined as

$$F_2(i,j) = E(D_{ij}^2). \tag{6}$$

Note that the F_2-statistic is guaranteed to be non-negative and symmetric by definition. According to Theorem 3, it is possible to write the drift D_{ij} as a sum of partial drifts over the paths $j \Rightarrow i$. In the following theorem, using the drift decomposition and the orthogonality of the partial drifts, we rewrite (6) in terms of squared partial drifts along the paths of Γ_{ji}.

We first define some quantities concerning partial drifts. Let Γ_{ij}^e denote the set of paths of Γ_{ij} containing edge e.

Definition 10 **(A- and B-coefficients on edges)** Let i, j be two nodes of an admixture graph. For a directed edge e consider the quantity A_e taking takes values in $[0,1]$ and defined by

$$A_e = \sum_{(\gamma_1,\gamma_2)\in(\Gamma_{ij}^e \times \Gamma_{ij}^e)} \mathrm{sgn}_{\gamma_1}(e)\mathrm{sgn}_{\gamma_2}(e) p_{\gamma_1} p_{\gamma_2}.$$

Let e_1 and e_2 be two undirected edges and define $B_{e_1e_2}$ as

$$B_{e_1e_2} = \sum_{(\gamma_1,\gamma_2)\in\Gamma_{ij}^{e_1} \times \Gamma_{ij}^{e_2}} \mathrm{sgn}_{\gamma_1}(e_1)\mathrm{sgn}_{\gamma_2}(e_2) p_{\gamma_1} p_{\gamma_2}.$$

The quantities A_e and $B_{e_1e_2}$ are denoted respectively as the A-coefficient of edge e and the B-coefficient of edges e_1, e_2. Each term of the A-coefficient is influenced by the sign of e in pairs of paths γ_1, γ_2 where e appears. The sign of the edges allows to take into account if an edge assumes opposite sign in the two paths. Similarly, the B-coefficient considers pairs of paths where two undirected edges (not necessarily coincident) appear. Observe that the A- and B-coefficient are simmetric within respect Γ_{ij} and Γ_{ji}.

The A- and B-coefficients can be interpreted as weights of a directed edge e and a pair of undirected edges e_1, e_2, respectively. Note that $A_e = 1$ if and only if $\Gamma_{ij}^e = \Gamma_{ij}$, and a similar consideration holds for B_{e_1,e_2}
Let $\mathscr{E}_{ij}$ be the set of edges involved in at least one path of Γ_{ij}.

Proposition 5 *Given $i, j \in \mathscr{V}$, a directed edge $e \in \mathscr{E}$ and a pair of undirected edges $e_1, e_2 \in \mathscr{E}$, then the following properties hold:*

1. *$A_e \geq 0$ and $A_e = 0$ if and only if $e \notin \mathscr{E}_{ij}$ (equivalently $\Gamma_{ij}^e = \emptyset$),*

2. $B_{e_1,e_1} + B_{e_2,e_2} + B_{e_1,e_2} \geq 0$ *and it takes value zero if and only if* $e_1, e_2 \notin \mathscr{E}_{ij}$ *(equivalently* $\Gamma_{ij}^e = \emptyset$*).*

For the following theorem we partition $\mathscr{E}_{ij}$ into two subsets, the set $\mathscr{E}_{ij}^u$ of undirected edges and the set $\mathscr{E}_{ij}^d$ of directed edges.

Theorem 4 (Canonical decomposition of the F_2-statistics along paths) *Given* $i, j \in \mathscr{V}$*, the statistic* $F_2(i,j)$ *can be decomposed in term of A- and B-coefficients and partial drifts as follows:*

$$F_2(i,j) = E\left(\sum_{e \in \mathscr{E}_{ji}^d} A_e d_e^2 + \sum_{e_1,e_2 \in \mathscr{E}_{ji}^u} B_{e_1 e_2} d_{e_1} d_{e_2} \right), \tag{7}$$

Note that the partial drifts appear without dependence on the paths of Γ_{ji} because such dependence is taken into account in the A- and B-coefficients.

Assuming the contribution has larger variance than the variable of the node generating it, then the squared partial drifts might be given in terms of variances [18],

$$E(d_{ij}^2) = \mathrm{Var}(C_{ij}) - \mathrm{Var}(V_i).$$

In the special case of an admixture graph with only one root, the second term in (7) vanishes and the canonical decomposition of the F_2-statistic becomes

$$F_2(i,j) = \sum_{k \to \ell \in \mathscr{E}_{ji}^d} A_{k\to\ell} \big(\mathrm{Var}(C_{k\ell}) - \mathrm{Var}(V_k) \big).$$

In [17,19], a visual method to decompose the F_2-statistic is introduced. We formally motivate it here. The steps to calculate the F_2-statistic between two nodes i, j based on the visual method are the following:

1. Consider all possible (ordered) pairs of paths $\gamma_1, \gamma_2 \in \Gamma_{ij}$, including coincident paths,
2. For each pair γ_1, γ_2, multiply by $p_{\gamma_1} p_{\gamma_2}$ the sum of squared partial drifts related to edges found in both paths,
3. For each pair γ_1, γ_2, multiply by $p_{\gamma_1} p_{\gamma_2}$ the sum of partial drifts related to undirected edges in the two paths,
4. Sum over the pairs of paths the quantities determined above and calculate the expectation.

In step 2. the partial drifts involved in the sum are related to the edges that overlap when the paths γ_1, γ_2 are traced on the admixture graph by connecting the ordered nodes. The paths do not necessarily overlap between roots in step 3. Some of the products of partial drifts can appear in more than one term of a sum, and can be therefore collected as common factors, with coefficient resulting in either the A- or the B-coefficients.

Example 2 Consider the statistic $F_2(5,6)$ in the admixture graph of Figure 1D. There are only two possible paths, namely

$$\gamma_1 = (5,4,1,3,6) \quad \text{and} \quad \gamma_2 = (5,4,2,3,6),$$

highlighted in Figure 3, where the four possible pairs of paths (γ_1,γ_1), (γ_2,γ_2), (γ_1,γ_2) and (γ_2,γ_1) are represented with two distinct colours. Note that there are two pairs containing two distinct paths. For each pair of paths apply the visual method. We obtain the following:

$$\begin{aligned} F_2(5,6) = E\Big(&p_{\gamma_1}^2(d_{45}^2+d_{14}^2+d_{13}^2+d_{36}^2) + p_{\gamma_2}^2(d_{45}^2+d_{24}^2+d_{23}^2+d_{36}^2) \\ &+ 2p_{\gamma_1}p_{\gamma_2}(d_{45}^2+d_{36}^2+d_{13}d_{23})\Big). \end{aligned}$$

By collecting terms with the same partial drift, we obtain

$$\begin{aligned} F_2(5,6) &= E\Big(d_{45}^2+d_{36}^2+p_{\gamma_1}^2(d_{14}^2+d_{13}^2)+p_{\gamma_2}^2(d_{24}^2+d_{23}^2)\Big) \\ &= E\Big(d_{45}^2+d_{36}^2+p_{\gamma_1}^2d_{14}^2+p_{\gamma_1}^2d_{13}^2+p_{\gamma_2}^2d_{24}^2+p_{\gamma_2}^2d_{23}^2+2p_{\gamma_1}p_{\gamma_2}d_{13}d_{23}\Big). \end{aligned}$$

Here we recognize the A- and B-coefficients,

$$\begin{aligned} &A_{45} = A_{36} = 1, \quad A_{14} = A_{24} = p_{\gamma_1}^2, \\ &B_{23,23} = p_{\gamma_2}^2, \quad B_{13,13} = p_{\gamma_1}^2, \quad B_{13,23} = B_{23,13} = p_{\gamma_1}p_{\gamma_2}. \end{aligned}$$

The F_2-statistic is often assumed to be additive, which means that given three nodes i,j,k, where $i \to k$, $k \to j$, then $F_2(i,j) = F_2(i,k) + F_2(k,j)$ [15,19,17]. This is true in some cases depending on the stochastic admixture graph. Here we give conditions that guarantee additivity of the F_2-statistics.

Proposition 6 (Additivity of the F_2-statistic) *Consider three distinct nodes i,j,k. If any path of Γ_{ij} passes through k, then the F_2-statistic between i,j can be split as the sum of the F_2-statistics between i,k and k,j,*

$$F_2(i,j) = F_2(i,k) + F_2(k,j).$$

The following definition illustrates the F_3- and F_4-statistics. These are often used as parameters to detect the presence of population admixture in specific admixture graphs [19,7,17,22,18]. We end by showing two applications of the F-statistics and their properties.

Definition 11 (F_3- and F_4-statistics) Let $i,j,k,l \in \mathscr{V}$ be four nodes of an admixture graph. The F_3-statistic between nodes i,j,k and the F_4-statistic between nodes i,j,k,ℓ are defined as

$$F_3(i;j,k) = E(D_{ij}D_{ik}) \quad \text{and} \quad F_4(i,j;k,l) = E(D_{ij}D_{k\ell}),$$

respectively.

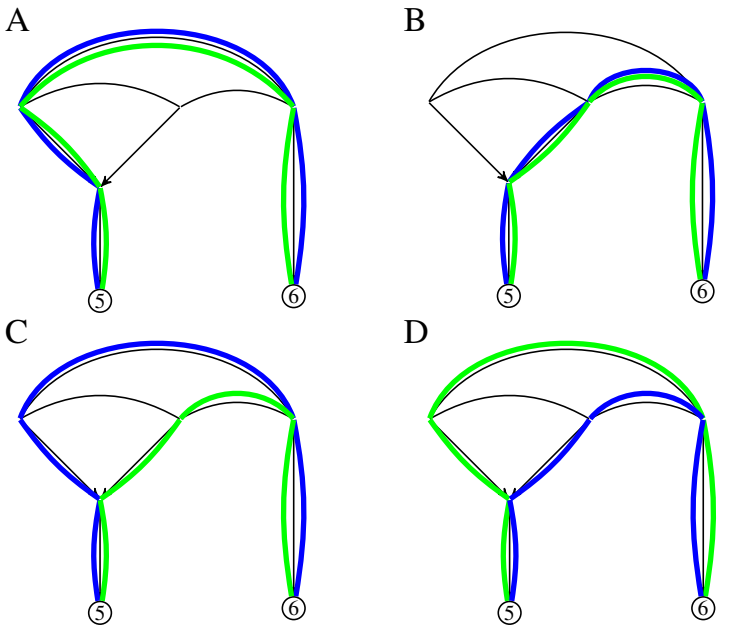


Fig. 3 Visual method for the F_2-statistic. Illustration of the visual method to calculate $F_2(5,6)$. Each of the four figures represent a possible pair of paths γ_1, γ_2. The overlapping directed edges and pairs of roots have the term contributing to the F_2-statistic written aside. (A,B) Each edge of γ_1 appears also in γ_2. Therefore their squared partial drifts contribute to the F_2-statistic between 5 and 6. (C,D) Edges $4 \to 5$ and $3 \to 6$ contribute with squared partial drifts to the $F_2(5,6)$. The pairs of undirected edges $(1 \leftrightarrow 3, 2 \leftrightarrow 3)$ and $(2 \leftrightarrow 3, 1 \leftrightarrow 3)$ contribute through the product of their drifts.

It is possible to apply the visual method to the F_3- and F_4-statistics and provide the following interpretations:

- the F_3-statistic is the weighted amount of overlapping edges of paths $i \Rightarrow j$ and $i \Rightarrow k$. If there is not any overlapping path, then the F_3-statistic assumes value zero,
- the F_4-statistic is the weighted amount of shared partial drifts along paths $j \Rightarrow i$ and $\ell \Rightarrow k$.

The F_3-statistic has an important role in defining the F_2-statistic as a distance between nodes. In fact the F_2-statistic does not necessarily comply with the definition of distance, depending on the configuration of admixed nodes in the admixture graph. Note that $F_2(i,j)$ can be rewritten as $F_2(i,k) + F_2(k,j) - 2F_3(k;i,j)$, therefore it fulfills the definition of metric only when $F_3(k;i,j) \geq 0$. We introduce a schematic representation of subgraphs to study in depth the properties of F_3 and the effect on F_2 as a metric.

Consider the schematic representation of a subgraph in Figure 4A (from now on denoted by cycle of type A). Here each directed dashed edge is a set of adjacent edges following the same direction. Some nodes are made explicit and represented by letters, and γ, δ represent two admixture paths.
If u, u' are roots, the thick edge between them represents an undirected edge, other-

wise a subpath of type $(u, i_{k-1}, \ldots, i_0, i'_0, \ldots, i'_{k'-1}, u')$. Along this part of the graph the two paths $\gamma \in \Gamma_{ki}$ and $\delta \in \Gamma_{kj}$ overlap with nodes in opposite order. In such a situation $F_3(k;i,j)$ has negative terms contributing to its value. There is no negative contribution if $u = u'$. Figure 4B illustrates a configuration (from now on denoted by cycle of type B) where there are at least two paths of type $u \Rightarrow k$. Figure 4C illustrates a sequence of cycles of type B followed by a cycle of type A.

Remark. The subgraph of Figure 4C is the only possible one where $F_3(k;i,j)$ has negative terms in its decomposition, because each path $\gamma \in \Gamma_{ki}$ and $\delta \in \Gamma_{kj}$ must respect the definition of admixture path. The cycles of type B and eventually some subpaths of the cycle of type A provide positive terms when γ, δ overlap, while the overlapping between u and u' provides negative terms if $u \neq u'$.

Consider two paths $\gamma \in \Gamma_{ki}$ and $\delta \in \Gamma_{kj}$ in one of the configurations of Figure 4. To ease the notation, we denote a subpath of γ and δ from ℓ to m by $\gamma_{\ell m}$ and $\delta_{\ell m}$, respectively, where ℓ, m are two nodes of the graph.

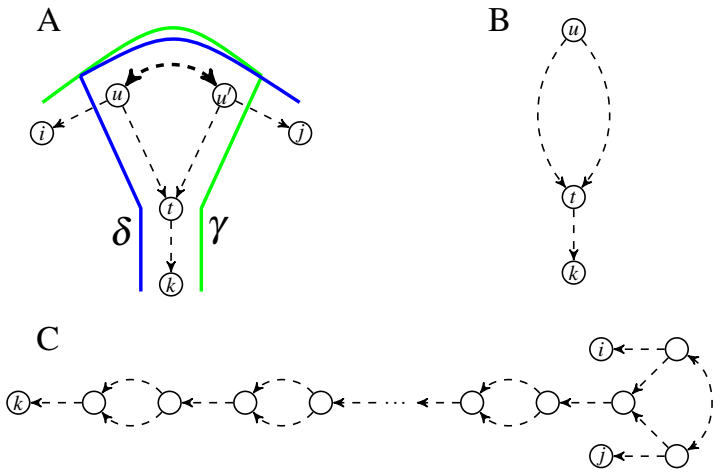


Fig. 4 Schematic representation of cycles in an admixture graph. Representation of two subgraphs where two admixture paths starting from k can form a cycle. The dashed arrows represent sequence of edges following the same direction. Note that each configuration allows peculiar cases by collapsing other nodes into one. For example the subgraph of type $t \to \cdots \to k$ is done with $t = u$ in configuration B. (A) Subgraph in which the two paths γ, δ starting in k overlap in the subpaths between nodes u, u' on edges with different sign. The products of the coincident partial drifts on this subpath can contribute to make $F_3(k;i,j) < 0$. This is the only configuration in which this can happen. (B) In this subgraph any two paths γ and δ starting in k overlap only on edges with the same sign. (C) In general any pair of paths contributing with negative terms to $F_3(k;i,j)$ starts in k, goes through a sequence of cycles of type B and ends in a cycle of type A.

Let a pair of paths $\gamma \in \Gamma_{ki}$, $\delta \in \Gamma_{kj}$ involve $c_{\gamma,\delta} \geq 0$ cycles in the configuration of Figure 4C, of which $c_{\gamma,\delta} - 1$ of type B and one of type A. The set of pairs of such paths is denoted by $\Gamma_{k;ij}$. Denote by $\mathscr{V}_{ij}^c$ the subset of nodes $k \in \mathscr{V}$ such that $\Gamma_{k;ij} \neq \emptyset$.

The F_3-statistics of type $F_3(k;i,j)$ that contain negative terms in their canonical decomposition, and therefore might assume negative value, are characterized in term of the nodes of $\mathscr{V}_{ij}^c$. The proof is a consequence of the remark discussed above.

Lemma 1 *The statistic $F_3(k;i,j)$ contains negative terms if and only if $k \in \mathscr{V}_{ij}^c$.*

The next theorem describes a necessary condition for having a non-negative F_3-statistic and therefore F_2 as a metric between two nodes. Given a pair $(\gamma,\delta) \in \Gamma_{k;ij}$, let $\mathscr{E}^+_{\gamma,\delta}$ and $\mathscr{E}^-_{\gamma,\delta}$ be the set of edges that have identical and opposite sign on the two paths, respectively.

Theorem 5 (Necessary and sufficient condition for F_2 being a metric) *Consider two nodes i,j of an admixture graph. Then $F_2(i,j)$ is a metric between i,j if the condition*

$$\sum_{(\gamma,\delta)\in\Gamma_{k;ij}} p_\gamma p_\delta \sum_{e\in\mathscr{E}^+_{\gamma,\delta}} E(d_e^2) \geq \sum_{(\gamma,\delta)\in\Gamma_{k;ij}} p_\gamma p_\delta \sum_{e\in\mathscr{E}^-_{\gamma,\delta}} E(d_e^2) \tag{8}$$

is fulfilled for each node $k \in \mathscr{V}_{ij}^c$.

Note that the right-hand side of (8) is zero on a tree, because a cycle of type A is not possible, since it involves an admixed node with more than one parent.

Corollary 2 *The F_2-statistic is always a measure if $\mathscr{G}$ is a tree.*

Example 3 Let $i=4$, $j=3$, $k=5$, in the tree of Figure 5A. The value of $F_3(5;3,4)$ is the length of the segment spanning from node 5 to the parent of nodes 3 and 4. In this setting $F_3(5;3,4) = F_2(2,5)$. The F_3-statistic is equal to zero if the length of such a segment is zero.

4.1 Application of the F_3- and F_4-statistics

In this subsection we analyze some specific models of admixture graphs (see Figure 5 and 6) on which the F_3- and F_4-statistics are applied in the population genetics' literature. Here the nodes of an admixture graph correspond to populations, and an admixture is the sum of contributions as in Definition 7(ii).

The F_3-statistic is often used to detect the recent admixture of two populations[19, 17], as represented in Figure 5B by node 4. The term recent refers to the assumption that branches $1\to 2$ and $1\to 6$ are significantly greater than $2\to 7, 6\to 7$ and $7\to 4$ in term of F_2-statistic [19] (in other words $1\to 2$ and $1\to 6$ have been undergoing a drift for longer time).
The recent admixture that generates population 4 is often detected through a negative value of $F_3(4;3,5)$. In fact

$$F_3(4;3,5) = F_2(7,4) + \alpha_{27}^2\mathbb{E}(d_{27}^2) + \alpha_{67}^2\mathbb{E}(d_{67}^2) - 2\alpha_{27}\alpha_{67}\big(F_2(1,2)+F_2(1,6)\big).$$

If $2\alpha_{27}\alpha_{67}\big(F_2(1,2)+F_2(1,6)\big)$ is larger than the sum of the other terms, then it holds that $F_3(4;3,5) < 0$.

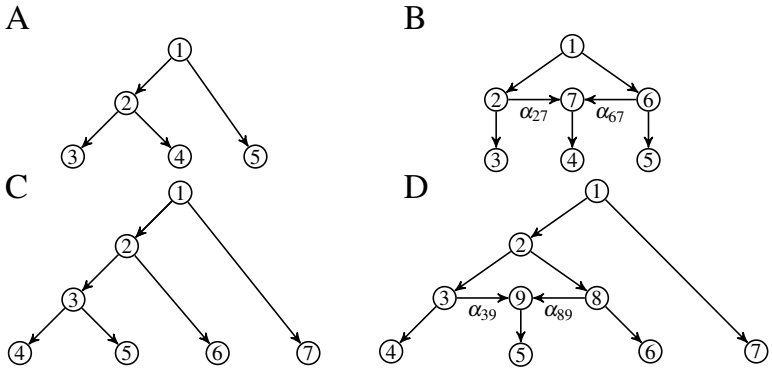


Fig. 5 Admixture graphs used to test for admixtures. (A) Admixture graph with three leaves, where no admixture is present. (B) Admixture graph involving 3 leaves and subject to an admixture. (C) Admixture graph with four leaves in which there are not admixed nodes. (D) Admixture graph with 4 leaves involving one node having two parents.

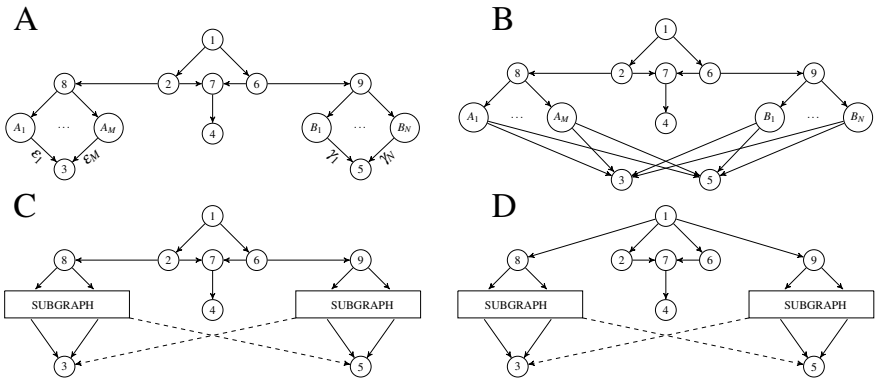


Fig. 6 Configurations for the analysis of the F_3-statistic. (A) Graph where $F_3(4;3,5)$ can assume negative values because of the presence of a cycle of type A. (B) A more general case of the graph in subfigure A, where the cycle of type A is still present. (C) In this general case, whichever structure is present in the subgraphs (that can have edges, represented with dashed arrows, connecting to nodes 3 and 5), edges $1 \to 6$ and $1 \to 2$ still contribute with negative terms to $F_3(4;3,5)$. (D) in this generic admixture graph it always holds $F_3(4;3,5) > 0$ because the cycle of type A is missing.

The negative F_3-statistic has been proven to be stable even if there is a more complex history for the parents of the admixed node 4 [19], where 3 and 5 have an arbitrary number of parents. However, this is true only in configurations of the type of Figure 6A-B (where eventually nodes $2,6$ can be connected by an edge to each node A_i, $i = 1,\ldots,M$ and B_j, $j = 1,\ldots,N$, respectively, bypassing nodes $8,9$). Those are of the type in Figure 4C. The nodes of Figure 4A can be matched in Figure 6A-B as it follows:

$$k = 4,\ t = 7,\ u = 2,\ u' = 6,\ i = 3,\ j = 5.$$

The admixture graph of Figure 6A involves an arbitrary number M and N of parents for node 3 and 5, respectively. For every $n = 1,\ldots,N$ and $m = 1,\ldots,M$, all

positive terms of the decomposition of $F_3(4;3,5)$ are of type

$$\alpha_{27}^2\varepsilon_m\gamma_n(F_2(7,4)+F_2(2,7))+\alpha_{67}^2\varepsilon_m\gamma_n(F_2(7,4)+F_2(6,7)). \tag{9}$$

The edges overlapping with opposite signs contribute to $F_3(4;3,5)$ with the following negative terms:

$$-2\alpha_{27}\alpha_{67}\varepsilon_m\gamma_n(F_2(1,2)+F_2(1,6)) \quad \text{for} \quad n=1,\dots,N,\ m=1,\dots,M. \tag{10}$$

Note that the sum of the quantities in (9) and (10) correspond to the left- and right-hand term of (8) fixing node $k=4$, respectively (apart from the negative sign due to the inequality). Again the F_3-statistic can be negative depending on the value of the partial drifts and labels. Analogous considerations hold for Figure 6B.

In general $F_3(4;3,5)$ will always contain negative terms in a structure of the type in Figure 6C. However the value of $F_2(4,7)+F_2(6,7)$ has to increase to compensate an increasing number of both labels and edges.
Any configuration of the type in Figure 6D provides a positive value of $F_3(4;3,5)$ because it misses a cycle of type B with $k=4, i=3, j=5$.

Another insight in the presence of admixture between two populations is given by the F_4-statistic. The F_4-statistic is applied to detect the presence of admixture in a graph with four leaves. Consider the tree of Figure 5C. Here $F_4(4,5;6,7)=0$ because Γ_{54} and Γ_{76} consist respectively of the path $\gamma=(5,3,4)$ and $\delta=(7,1,2,6)$, that do not overlap. In Figure 5D node 5 has two parents. The value of $F_4(4,5;6,7)$ is $\alpha_{89}^2F_2(2,8)$. Analogously, the F_4-statistic is negative if the node with two parents is 4. Therefore the F_4-statistic also discerns which nodes are involved in the admixture.

We provide a definition for sets of edges that appear in some decompositions of F_2-statistics with the same A-coefficient.

Definition 12 (Bottleneck edge) Let $C\subset\mathscr{V}$ and consider $S\subset\mathscr{E}$ a subset of edges with $|S|>1$. If each edge of S has the same A- or B-coefficient between different pairs of nodes of C, and S is maximal w.r.t. such a property, then S is called bottleneck. The bottleneck number of C is defined as follows:

$$n_{bot}^C=\sum\nolimits_{S\in\mathrm{bot}(C)}(|S|-1),$$

where $\mathrm{bot}(C)$ is the set of possible bottlenecks of C.

Denote by $\mathbb{F}_2^C$ the set of possible F_2-statistics between nodes of C. Then a bottleneck S can be seen as the set of edges having the same coefficients in the canonical decomposition of a subset of F_2^C related to the pairs of nodes of the bottleneck.

It is possible to prove[19,17] that $\mathbb{F}_2^C$ spans the linear space $\mathbb{F}_C$ of the F-statistics between nodes of C. Moreover, given a node $k\in C$, the set

$$\mathbb{F}_{2,3}^k=\big\{F_2(k,j),\,j\in C\backslash\{k\}\big\}\cup\big\{F_3(k;i,j),\,i,j\in C\backslash\{k\}\big\},$$

can be written in function of the elements of $\mathbb{F}_2^C$, and vice versa [19]. Further, note that the canonical decompositions of the elements of $\mathbb{F}_2^C$ involve only the nodes and edges of $\mathscr{G}_C$.

Theorem 6 *Let C be a subset of n nodes in an admixture graph $\mathscr{G}$. Let $\mathscr{G}_C$ have n_d edges and at most 2 roots. If for any triplet of nodes of C the hypothesis of Proposition 6 is not verified, and the following condition*

$$\binom{n}{2} < n_d - n_{bot}^C$$

holds, then

- *the set $\mathbb{F}_2^C$ is a basis for $\mathbb{F}_C$;*
- *a basis of $\mathbb{F}_C$ is also defined by $\mathbb{F}_{2,3}^k$, where $k \in C$;*
- $\dim \mathbb{F}_C = \binom{n}{2}$.

Proof The theorem is proved for the set $\mathbb{F}_2^C$, and it will hold also for the set $\mathbb{F}_{2,3}^k$. We keep the notation of the proof uncluttered by indexing with $F^1, F^2, \ldots, F^N$ the elements of $\mathbb{F}_2^C$ and with $d^1, \ldots, d^{n_d}$ the partial drifts on the edges of $\mathscr{E}_C$.
Consider the canonical decomposition of the F_2-statistic (7) and $F^t \in \mathbb{F}_2^C$. Then

$$F^t = \sum_e A_e^{(t)} E\big((d_e^{(t)})^2\big), \tag{11}$$

The index of the sum represents the edges involved in the canonical decomposition of F^t. Note that, in presence of 2 roots and only one directed edge e', the B-coefficient $B_{e',e'}^{(t)}$ is the same as $A_{e'}^{(t)}$ (where the sign of the undirected edge is used). Therefore (11) admits the presence of an undirected edge in case $\mathscr{G}_C$ has two roots.
Observe that each A-coefficient is dependent on the F_2-statistic F^t. Such a dependence is shown by the index (t). Every partial drift is independent of the F_2-statistic because is not influenced by the probabilities of the paths between nodes.

The objective is to prove that the elements of $\mathbb{F}_2^C$ are linearly independent, that is

$$\sum_{t=1}^N \omega_t F^t = 0 \iff \omega_1 = \cdots = \omega_N = 0 \tag{12}$$

over all the values that can be assumed by the partial drifts. The left-hand side of the double implication above can be rewritten as

$$\sum_{t=1}^N \omega_t F^t = \sum_{t=1}^N \omega_t \Big(\sum_e A_e^{(t)} E(d_e^2) \Big) = E\Big(\sum_e d_e^2 \Big(\sum_{t=1}^N \omega_t A_e^{(t)} \Big) \Big).$$

The coefficients $A_e^{(t)}$ are positive (see Proposition 5). Rewrite (4.1) as it follows:

$$\sum_{t=1}^N \omega_t F^t = \sum_{t=1}^N \omega_t \Big(\sum_e A_e^{(t)} E(d_e^2) \Big) = E\Big(\sum_e d_e^2 \Big(\sum_{t=1}^N \omega_t A_e^{(t)} \Big) \Big).$$

The terms d_e^2 and their expectations are positive. Thus the condition in (12) becomes a system in the edges e of the admixture graph:

$$\sum_{t=1}^N \omega_t A_e^{(t)} = 0 \tag{13}$$

The system of equations above has $\binom{n}{2}$ variables and $n_d - n_{bot}^C$ equations. In fact, for every bottleneck S there are n_S coincident equations involving $w_i \neq 0$ for each $F^i \in \mathbb{F}_2^C$ that contains the elements of the bottleneck. It is not possible to find other linear dependence relationships between equations, because by hypothesis the additivity of F_2-statistics is not verified. By hypothesis $\binom{n}{2} < n_d - n_{bot}^C$, therefore the only solution of equation (13) is $w_t = 0$ for every $t = 1, \ldots, N$.

Remark. The linear independence requires the presence of at most two roots in $\mathscr{G}_C$. The presence of more than two roots would change (4.1) into

$$\sum_{t=1}^{N}\omega_t F^t = \sum_{t=1}^{N}\omega_t\left(\sum_e A_e^{(t)}E(d_e^2)+\sum_{e_1,e_2}B_{e_1,e_2}^{(t)}E(d_{e_1}d_{e_2})\right) = E\left(\sum_e d_e^2\Big(\sum_{t=1}^{N}\omega_t A_e^{(t)}\Big)\right)$$
$$+E\left(\sum_{e_1,e_2}\sum_{t=1}^{N}\omega_t\big(B_{e_1,e_1}^{(t)}d_{e_1}^2+B_{e_2,e_2}^{(t)}d_{e_2}^2+2B_{e_1,e_2}^{(t)}d_{e_1}d_{e_2}\big)\right).$$

It is not possible to factorize the second term of the sum between the B-coefficients and the partial drifts, therefore one cannot obtain a system of equations as in (13) and a similar statement for the theorem with more than 2 roots.

Example 4 Consider the tree of Figure 5C, and let C be the subset of nodes $\{4,5,6,7\}$. The set $\mathbb{F}_2^C$ contains 6 elements and there are $n_d = 6$ directed edges in $\mathscr{E}_C$. The set $S = \{1\to 3,\ 1\to 7\}$ is a bottleneck for C, since the $\mathbb{F}_2$-statistics between $(4,7),(5,7)$ and $(6,7)$ have those edges with same coefficients in their decomposition. Therefore $n_{bot}^C = 1$. Thus the considered tree does not fulfill the hypothesis of Theorem 6, because $N = 6$ and $n_d - n_{bot}^C = 5$. It is immediate to see that decomposing the six F_2-statistics, there is a linear dependence relationship.
In an analogous way the same subset of nodes does not originate linearly independent F_2-statistics for the more complex admixture graph of Figure 5D, where $N = 6$ and $n_d = 9$. The set $S_1 = \{1\to 3,\ 1\to 7,\ 9\to 5\}$ is a bottleneck related to the pairs of nodes $(4,5),(5,6)$ and $(5,7)$, while $S_2 = \{1\to 3,\ 1\to 7\}$ is a bottleneck related to the pairs $(4,7),(5,7)$ and $(6,7)$. Thus $n_{bot}^C = 2+1 = 3$. In this case $N = n_d - n_{bot}^C$.
The system in (13) is expressed in this case as follows:

$$\begin{aligned}
\omega_1+\omega_2+\omega_3 &= 0 && F_2(4,3)\\
\omega_1+\omega_4+\omega_5 &= 0 && F_2(5,9),E(d_{29}^2),E(d_{89}^2) \qquad (14)\\
\alpha_{89}^2\omega_1+\omega_2+\omega_3+\alpha_{29}^2\omega_4+\alpha_{89}^2\omega_5 &= 0 && F_2(2,3)\\
\alpha_{89}^2\omega_1+\omega_2+\alpha_{29}^2\omega_4+\alpha_{89}^2\omega_5+\omega_6 &= 0 && F_2(3,8)\\
\omega_2+\omega_4+\omega_6 &= 0 && F_2(8,6)\\
\omega_3+\omega_5+\omega_6 &= 0 && F_2(3,1),F_2(1,7) \qquad (15)
\end{aligned}$$

On the right hand side is reported to which F_2-statistic or partial drift the equation is referred to. Equation (14) appears three times and (15) appears twice, as expected with $n_{bot}^C = 2+1$. The total number of distinct equations is therefore $n_d - n_{bot}^C = 6$, and the solution of the system is unique and in general different from $\omega_i = 0$ for $i = 1,\dots,6$.

5 Appendix 1

Any graph where the nodes are connected by either directed or undirected edges, and without directed cycles, is called a chain graph. The nodes of a chain graph can be partitioned in subsets, denoted as components, through an equivalence relation. Two

nodes i, j are part of the same component if there exist a path from i to j and vice versa. In this case a path is meant with the usual definition in the context of graphs, that is, following the direction of the edges. Following the notation for chain graphs, we denote by $i > j$ a path from i to j.

The future and the past of a node i are respectively defined as

$$\phi(V) := \{Q \,:\, V > Q\} \quad \text{and} \quad \pi(V) := \{Q \,:\, Q > V\}.$$

An analogous definition applies for the future and past of a component τ. A component τ is terminal if its future is empty. A set of nodes is an anterior set if it can be generated from the graph with a stepwise removal of terminal components.

Consider a subset of nodes $\mathscr{A}$ of a chain graph. Its border is defined as

$$bd(\mathscr{A}) := \left\{V \in \mathscr{V} \,:\, V \to A \text{ or } V \leftrightarrow A \text{ for some } A \in \mathscr{A}\right\},$$

where $V \to A$ and $V \leftrightarrow A$ denote the presence of a directed edge from V to A and an undirected edge between the two nodes, respectively. Given a chain graph $\mathscr{G} = (\mathscr{V}, \mathscr{E})$, its moral graph is $\mathscr{G}^m = (\mathscr{V}, \mathscr{E}^m)$, where E^m consists of the union between

- the set $\mathscr{E}^u$ where all elements of $\mathscr{E}$ are turned into undirected edges,
- the undirected edges connecting all pairs of nodes that are in the border of a component of $\mathscr{G}$.

The probabilistic conditional dependence that we assume in our treatment is the Global $\mathscr{G}$-markovian (GM) property. A probability measure defined on $(V_1, \ldots, V_N)$ is GM if

$$V \perp_{\mathbb{P}} Q \,|\, C$$

whenever C separates V and Q in $\mathscr{G}_{an(V \cup Q \cup C)^m}$, the moral graph of the smallest anterior set containing $V \cup Q \cup C$.

6 Appendix 2

This appendix contains the proofs of some statements presented in this paper. In this setup we consider admixture graphs, denoted by $\mathscr{G} = (\mathscr{V}, \mathscr{E}, \mathscr{L})$, that are not trivial, that is, they cannot contain only roots and undirected edges. The set $\mathscr{V}$ contains the nodes of the graph, $\mathscr{E}$ its edges and $\mathscr{L}$ its edges' labels.

A directed edge from i to j is written as $i \to j$, if it is undirected it is denoted by $i \leftrightarrow j$. Edges can be briefly denoted by e. An edge $i \to j$ is said to be ingoing to j and outgoing from i. Two nodes are connected by an undirected edge only if they are roots, from which only directed edges are only outgoing. Any other node is admixed if it has both ingoing and outgoing edges, or a leaf if it has only ingoing nodes. Each edge is ordered: directed edges $i \to j$ are considered ordered from i to j, while undirected edges can have an arbitrary order. The label of an edge $i \to j$ is denoted by α_{ij} or α_{ji}, and undirected edges have label 1. Directed edges ingoing to a node have positive labels that sum to 1.

Let i,j be two nodes. An admixture path γ from i to j consists of an ordered sequence $(i, i_{k-1},\dots,i_1,i_0,i'_0,i'_1,\dots,i'_{k'-1},j)$ of adjacent nodes beginning in i and ending in j. Any two adjacent nodes amongst the first $k \geq 0$ ordered nodes are connected by a directed edge with order opposite to the nodes' order. The node i_0 can be a root. Only a root can be adjacent to another root i'_0, otherwise $i_0 = i'_0$. Two adjacent nodes amongst the following $k' \geq 0$ are connected by a directed edge ordered in the same way as the nodes are. The set of all paths from i to j is denoted by Γ_{ij}. An edge has positive sign in γ, $\mathrm{sgn}_\gamma(e) = +1$, if the order of the edge is the same of its nodes in γ, otherwise $\mathrm{sgn}_\gamma(e) = -1$. The label of a path γ, p_γ, is the product of the labels of edges between the adjacent nodes of the paths.

Given an admixture graph $\mathscr{G} = (\mathscr{V},\mathscr{E},\mathscr{L})$, consider the augmented graph $\mathscr{G}^* = (\mathscr{V}^*,\mathscr{E}^*)$ with nodes $\mathscr{V}^* = \mathscr{V} \cup \{(i,j) \,|\, i \to j \in \mathscr{E}\}$, where each directed edge $i \to j \in \mathscr{E}$ is split into $i \to (i,j) \in \mathscr{E}^*$ and $(i,j) \to j \in \mathscr{E}^*$, and where undirected edges are unchanged. For each node of $\mathscr{G}^*$ associate a stochastic variable with finite mean denoted by V_j if $j \in \mathscr{V}^*$ and C_{ij} if $(i,j) \in \mathscr{V}^*$. The variables C_{ij} are called contribution variables and the nodes (i,j) contribution nodes. We denote $\mathscr{G}$ a stochastic admixture graph if:

(i) $\mathscr{G}^*$ is a chain graph, that is, it has no directed cycles.
(ii) $V_j = \sum_{i \in par(j)} \alpha_{ij} C_{ij}$ for any admixed node j.
(iii) $E(C_{ij}|V_i) = V_i$ for any admixed node i, with $E(X|Y)$ being the expectation of a variable X conditionally to a variable Y.

Define the drift between nodes i,j as $D_{ij} = V_j - V_i$. Given an edge $e = i \to j$ or $e = i \leftrightarrow j$, the partial drift between i,j is defined as $d_{ij} = C_{ij} - V_i$. The partial drift between i,j in a path γ is $d^\gamma_{ij} = \mathrm{sgn}_\gamma(e) d_{ij}$. Given four nodes i,j,k,ℓ, the F-statistics are defined as $F_2(i,j) = E[D_(ij)^2]$, $F_3(k;i,j) = E[D_{ik}D_{jk}]$ and $F_4(i,j;k,\ell) = E[D_{ij}D_{k\ell}]$. Those are respectively the F_2-statistic between i,j, the F_3-statistic between i,j,k and the F_4-statistic between i,j,k,ℓ.

We consider coefficients that assume value in $[0,1]$ for directed and undirected edges called A- and B-coefficients, respectively. Given two nodes $i,j \in \mathscr{V}$ and $e \in \mathscr{E}$, the A-coefficient of such edge is

$$A_e = \sum_{\gamma_1,\gamma_2 \in \Gamma^e_{ij}} \mathrm{sgn}_{\gamma_1}(e)\mathrm{sgn}_{\gamma_2}(e) p_{\gamma_1} p_{\gamma_2}.$$

The B-coefficient of two undirected edges e_1 and e_2 is

$$B_{e_1e_2} = \sum_{(\gamma_1,\gamma_2) \in \Gamma^{e_1}_{ij} \times \Gamma^{e_2}_{ij}} \mathrm{sgn}_{\gamma_1}(e_1)\mathrm{sgn}_{\gamma_2}(e_2) p_{\gamma_1} p_{\gamma_2}.$$

Proof of Proposition 1. Given $i,j \in \mathscr{A}$, there exists $r_1,r_2 \in \mathscr{R}$ such that they define two paths $(i, i_{k-1},\dots,i_1,r_1)$ and $(r_2,i'_1,\dots,i'_{k'-1},j)$ for some nodes $i_1,\dots,i_{k-1},i'_1,\dots,i'_{k'-1} \in \mathscr{A}$, $k,k' \geq 1$ (because the admixture graph is connected and there are no directed cycles). Either all these nodes are distinct (except perhaps for r_1,r_2) in which case they form a path from i to j, or there is $i_\ell = i'_{\ell'}$ for some ℓ,ℓ'. Choose ℓ,ℓ' such that $\ell+\ell' \leq k+k'$ is as large as possible. Then $(i, i_{k-1},\dots,i_{\ell+1},i_\ell,i'_{\ell'+1},\dots,i'_{k'-1},j)$

is a path from i to j by definition of a path. There cannot be any repeated nodes. If there was, then $\ell+\ell'$ would not be as large as possible. Hence $\Gamma_{ij} \neq \emptyset$.

If $i, j \in \mathscr{R}$, then they are trivially connected by a path. If $i \in \mathscr{R}$ and $j \in \mathscr{A}$, then there is a path $(j, i_{k-1}, \ldots, i_1, i)$ for some $i_1, \ldots, i_{k-1} \in \mathscr{A}$ (as before).

To prove the second part of the proposition, we proceed by induction in the length of the paths. For $i, j \in \mathscr{V}$, $i \neq j$, consider Γ_{ij} and let l_γ denote the number of edges in a path $\gamma \in \Gamma_{ij}$. Assume $l_\gamma \leq 1$ for all $\gamma \in \Gamma_{ij}$. Then either $i \to j$, $i \leftarrow j$ or $i \leftrightarrow j$ is the edge involved in a path (i, j). In the latter case the label is 1. In the former cases, there cannot be any node i' such that $i \leftarrow i'$ (similarly if $i \leftarrow j$) because then there would be a path from i to j via i' as $\Gamma_{ii'} \neq \emptyset$, implying the length is larger than one. Hence the label of $i \to j$ is by definition 1. Hence $\sum_{\gamma \in \Gamma_{ij}} p_\gamma = 1$.

Assume now the statement holds if $\sum_{\gamma \in \Gamma_{ij}} p_\gamma = 1$ and $l_\gamma \leq k$, $\gamma \in \Gamma_{ij}$, for some $k \geq 1$. Consider two nodes $i, j \in \mathscr{V}$ such that all paths between them fulfil $l_\gamma \leq k+1$. Then

$$\sum_{\gamma \in \Gamma_{ij}} p_\gamma = \sum_{\ell \in par(j)} \sum_{\gamma \in \Gamma_{i\ell}} p_\gamma \alpha_{\ell j} = \sum_{\ell \in par(j)} \alpha_{\ell j} \sum_{\gamma \in \Gamma_{i\ell}} p_\gamma = \sum_{\ell \in par(j)} \alpha_{\ell j} = 1$$

(potentially by reverting the roles of i, j such that all paths are ingoing to j), as all paths in $\Gamma_{i\ell}$ must have length at most k. □

Proof of Theorem 1. We first prove the double implication between 1 and 2. Assume 1. Let γ be the unique path $i \Rightarrow j$, where $i, j \in \mathscr{V}$. Since γ is unique, the admixed nodes of the path must have exactly one parent. In fact, if a node $i_\ell \in \gamma$ had a second parent, say $\tilde{i}_{\ell-1}$, then since $\mathscr{G}$ is connected it would be possible to create a new path involving the edge $\tilde{i}_{\ell-1} \to i_\ell$, not present in γ. But this contradicts uniqueness. It follows that the product of labels is one. Oppositely, if the label p_γ of a path is one, then $\Gamma_{ij} = \{\gamma\}$, according to Proposition 1.

Next we prove the double implication between 1 and 3. Assume 1. Let $\mathscr{A}_r$ be the nodes in $\mathscr{A}$ for which there is a (unique) path from $r \in \mathscr{R}$ to a node in $\mathscr{A}$, not involving any other root. Any $i \in \mathscr{A}$ is in at least one $\mathscr{A}_r$, and cannot be in two such sets $\mathscr{A}_{r_1}, \mathscr{A}_{r_2}$, because then there would be paths $(r_1, \ldots, i)$ and $(r_1, r_2, \ldots, i)$ from r_1 to i, contradicting uniqueness of paths. Assume 3. This implication is straightforward using the definition of a forest and taking into account the undirected edges between the roots.

□

Proof of Proposition 2. Consider $r_1 \in \mathscr{R}$. Let γ be a path from the leaf ℓ to another root r_2. The union of γ and the undirected edge between r_2 and r_1 is a path of $\Gamma_{\ell r_1}$ and its label is p_γ. Therefore $\cup_{r \in \mathscr{R}} \Omega_{\ell r} = \Gamma_{\ell r}$ and it follows that

$$\sum_{r \in \mathscr{R}} q_{\ell r} = \sum_{r \in \mathscr{R}} \sum_{\gamma \in \Omega_{\ell r}} p_\gamma = \sum_{\gamma \in \Gamma_{\ell r_1}} p_\gamma = 1.$$

□

We define some operations on paths. Note that the notation is a slight abuse of the one used for operations on sets.

Definition 13 (Operations on Paths) Let γ_1, γ_2 be two paths on the same admixture graph, such that the last m nodes of γ_1 and the first m nodes of γ_2 are the same, where $m \geq 1$. The union of the two paths, denoted by $\gamma_1 \cup \gamma_2$, is the ordered sequence of nodes consisting of the ordered nodes of γ_1 followed by the ordered nodes of γ_2. The nodes shared by γ_1 and γ_2 are not repeated.
Let γ be a path and C a set of nodes. Define $\gamma \backslash C$ as γ where the nodes of C are removed from the path.

Lemma 2 *Let γ_1, γ_2 and m be defined as in Definition 13. At least one between $\gamma_1 \backslash \{\gamma_1 \cap \gamma_2\}$ and $\gamma_2 \backslash \{\gamma_1 \cap \gamma_2\}$ is of type $(i_k, i_{k-1}, \ldots, i_{m+1}, i_m)$ or $(i'_0, i'_1, \ldots, i'_{k'-m'-1}, i'_{k'-m'})$, where $m = 0, \ldots, k$ and $m = 0, \ldots, k'$, if and only if $\gamma = \gamma_1 \cup \gamma_2$ is an admixture path.*

Proof Let $\gamma_1 \backslash \{\gamma_1 \cap \gamma_2\}$ be of type $(i_k, i_{k-1}, \ldots, i_{m+1}, i_m)$. Since γ_2 is a path, it is such that $(j_\ell, j_{\ell-1}, \ldots, j_1, j_0, j'_0, j'_1, \ldots, j'_{\ell'-1}, j'_{\ell'})$, where $\ell, \ell' \geq 0$ and eventually $j_0 = j'_0$. Then the union $\gamma_1 \backslash \{\gamma_1 \cap \gamma_2\} \cup \gamma_2 = \gamma_1 \cup \gamma_2$ still fulfills the definition of admixture path. Analogous considerations hold if $\gamma_1 \backslash \{\gamma_1 \cap \gamma_2\}$ is of type $(i'_0, i'_1, \ldots, i'_{k'-m-1}, i'_{k'-m})$, and if the roles of γ_1, γ_2 are inverted.
Let γ be the path $(i_k, i_{k-1}, \ldots, i_1, i_0, i'_0, i'_1, \ldots, i'_{k'-1}, i'_{k'})$ union of two subpaths γ_1, γ_2. If $\gamma_1 \backslash \{\gamma_1 \cap \gamma_2\}, \gamma_2 \backslash \{\gamma_1 \cap \gamma_2\}$ are neither of the first nor the second type in hypothesis, then $\gamma_1 \cup \gamma_2$ contains three adjacent nodes i_{n-1}, i_n, i_{n+1} such that $i_{n-1} \to i_n$ and $i_n \leftarrow i_{n+1}$ contradicting the definition of admixture path. □

Lemma 2 holds in the specific case of $m = 1$, where the two paths γ_1, γ_2 have only one node in common. The hypothesis can be further simplified without including the paths' intersection.

Corollary 3 *Let γ_1, γ_2 and $m = 1$ be defined as in Definition 5. At least one between γ_1 and γ_2 is of type $(i_k, i_{k-1}, \ldots, i_0)$ or $(i'_0, i'_1, \ldots, i'_{k'})$ if and only if $\gamma = \gamma_1 \cup \gamma_2$ is an admixture path.*

Proof of Proposition 3. The equivalence between conditions 1. and 2. holds from Corollary 3. From assumption 2 it follows that $\mathscr{V}_C = \mathscr{V}_{C_0}$, therefore we prove that $\mathscr{E}_C = \mathscr{E}_{C_0}$. Consider $e = k \to \ell \in \mathscr{E}_C \backslash \mathscr{E}_{C_0}$, then there is no path γ' between two leaves such that $e \in \gamma'$. Since $k \notin C_0$ and $\mathscr{G}_C$ is connected it is possible to find $i \in C_0$ and $\gamma \in \Gamma_{ik}$ of type $i \leftarrow \ldots \leftarrow k$, such that $e \in \gamma$. Since $k \in \mathscr{V}_{C_0}$ and $e \notin \mathscr{E}_{C_0}$, there exist an edge $e' = k \to \ell'$. There is another leaf j such that a path $\delta \in \Gamma_{jk}$, $e' \in \delta$; moreover $\delta \cap \gamma = \{k\}$, otherwise there would be $\delta' \in \Gamma_{jk}$ such that $e \in \delta'$. But from Corollary 3 $\gamma \cup \delta$ is an admixture path. Thus $e \in \mathscr{E}_{C_0}$ and $\mathscr{G}_C = \mathscr{G}_{C_0}$.
Let $\mathscr{G}_C = \mathscr{G}_{C_0}$. Then hypothesis 2 is fulfilled by definition of spanned graph.
Let $C'_0 \subset C_0$ such that $\mathscr{G}_{C'_0} = \mathscr{G}_{C_0}$. Consider $e = k \to i \in \mathscr{E}_{C'_0}$, $i \in C_0 \backslash C'_0$. All paths involving e must contain node i, therefore $e \notin \mathscr{E}_{C'_0}$ and $\mathscr{G}_{C'_0} \neq \mathscr{G}_{C_0}$, against the initial assumption. Thus C_0 is the smallest subset of nodes spanning $\mathscr{G}_C$. □

Proof of Theorem 2
Let R_i, R_j be two root variable, corresponding to distinct roots in $\mathscr{R}$. Assume

$$E(R_i \,|\, R_j) = R_j \quad \text{and} \quad E(R_j \,|\, R_i) = R_i. \tag{16}$$

Applying the law of total variance to R_i, it follows that

$$\mathrm{Var}\big(R_i\big) = E\Big(\mathrm{Var}\big(R_i|R_j\big)\Big) + \mathrm{Var}\Big(E\big(R_i|R_j\big)\Big).$$

Analogous equation holds for R_j. Applying (16) to the right-hand side of the equation above it holds that

$$\mathrm{Var}\big(R_i\big) = E\Big(\mathrm{Var}\big(R_i|R_j\big)\Big) + \mathrm{Var}\Big(R_j\Big), \quad \mathrm{Var}\big(R_j\big) = E\Big(\mathrm{Var}\big(R_j|R_i\big)\Big) + \mathrm{Var}\Big(R_i\Big).$$

Consider their sum:

$$\mathrm{Var}\big(R_i\big) + \mathrm{Var}\big(R_j\big) = \mathrm{Var}\big(R_i\big) + \mathrm{Var}\big(R_j\big) + E\Big(\mathrm{Var}\big(R_i|R_j\big)\Big) + E\Big(\mathrm{Var}\big(R_j|R_i\big)\Big).$$

Hence,

$$E\Big(\mathrm{Var}\big(R_i|R_j\big)\Big) + E\Big(\mathrm{Var}\big(R_j|R_i\big)\Big) = 0.$$

As the two variances are non-negative almost surely, then

$$\mathrm{Var}\big(R_i|R_j\big) = \mathrm{Var}\big(R_j|R_i\big) = 0. \tag{17}$$

Using (17) and the definition of $\mathrm{Var}\big(R_i|R_j\big) = E\Big(\big(R_i - E(R_i|R_j)\big)^2 \,|\, R_j\Big)$, we get $R_i = E(R_i|R_j)$, because $\big(R_i - E(R_i|R_j)\big)^2$ is non-negative. However by assumption $R_j = E(R_i|R_j)$, so $R_i = R_j$. It completes the proof. □

Remark. Note that, for any two nodes $i, j \in \mathscr{V}$, at least one between Γ_{ij} and Γ_{ji} contains only paths involving a parent of i and j, respectively.

Proof of Theorem 3 (Canonical decomposition of a drift along paths)
Assume that j is such that any path of Γ_{ji} involves a parent of j. We prove the statement by induction on the maximum number of edges n of the paths $\gamma \in \Gamma_{ji}$. If $n = 1$, there is only one path $\gamma = (j,i)$ from j to i, where $j \leftarrow i$. Therefore $p_\gamma = 1$ and $\sum_{e\in\gamma} d_e^\gamma = V_j - V_i$. Let (4) be true for $n = \ell$ and consider $n = \ell + 1$ for some $\ell \geq 1$. Since a path $\gamma \in \Gamma_{ji}$ always contains a node $j' \in par(j)$, then it holds that

$$\sum_{\gamma\in\Gamma_{ji}} p_\gamma = \sum_{j'\in par(j)} \alpha_{j'j} \sum_{\gamma'\in\Gamma_{j'i}} p_{\gamma'}. \tag{18}$$

Using (18) and the inductive hypothesis, the decomposition in (4) can be rewritten as:

$$\begin{aligned}
\sum_{\gamma\in\Gamma_{ji}} p_\gamma \sum_{e\in\gamma} d_e^\gamma &= \sum_{j'\in par(j)} \alpha_{j'j}\Big(d_{j'j}^{(j,j')} + \sum_{\gamma'\in\Gamma_{j'i}} p_{\gamma'} \sum_{e\in\gamma'} d_e^{\gamma'}\Big) \\
&= \sum_{j'\in par(j)} \alpha_{j'j}\Big(C_{j'j} - V_{j'} + V_{j'} - V_i\Big) \\
&= \sum_{j'\in par(j)} \alpha_{j'j} C_{j'j} - \sum_{j'\in par(j)} \alpha_{j'j} V_i = V_j - V_i.
\end{aligned}$$

Note that $D_{ji} = -D_{ij}$. Thus

$$D_{ji} = -\sum_{\gamma\in\Gamma_{ji}} p_\gamma \sum_{e\in\gamma} d_e^\gamma = \sum_{\gamma\in\Gamma_{ij}} p_\gamma \sum_{e\in\gamma} d_e^\gamma,$$

using the fact that the sign of an edge changes along a path, if the nodes of such a path are reordered from the last to the first. □

Proof of Proposition 4. Denote $e_1 = i \to j$ and $e_2 = k \to \ell$ (with one of the two eventually undirected). Let us assume that the blocks B_{n_i}, B_{n_k}, related to the nodes i,k are such that $n_i \geq n_k$. The expected value in (5) can be rewritten as

$$E(d_{ij}d_{k\ell}) = E(C_{ij}C_{k\ell}) - E(C_{ij}V_k) - E(C_{k\ell}V_i) + E(V_iV_k).$$

Consider the first expectation,

$$\begin{aligned} E(C_{ij}C_{k\ell}) &= E(E(C_{ij}C_{k\ell}|V_i,V_k)) = E(E(C_{ij}|V_i,V_k)E(C_{k\ell}|V_i,V_k)) && (19)\\ &= E(E(C_{ij}|V_i)E(C_{k\ell}|V_i,V_k)) = E(V_iE(C_{k\ell}|V_i,V_k)). && (20)\end{aligned}$$

Here we used (1) ,(3) in (19) and Definition 7 in (20). With similar considerations, it follows that

$$E(C_{k\ell}V_i) = E(E(C_{k\ell}V_i|V_i,V_k)) = E(V_iE(C_{k\ell}|V_i,V_k)).$$

Using the properties of conditional expectation, it holds that

$$E(V_iV_k) = E(E(V_iV_k|V_i)) = E(V_iE(V_k|V_i)).$$

The remaining term results in the following equation:

$$\begin{aligned} E(C_{ij}V_k) &= E(E(C_{ij}V_k|V_i)) && (21)\\ &= E(E(C_{ij}|V_i)E(V_k|V_i)) && (22)\\ &= E(V_iE(V_k|V_i)), \end{aligned}$$

where we have applied (3) and Definition 7(ii) in (21) and (22), respectively. Hence $E(d_{ij}d_{k\ell}) = 0$. □

Proof of Proposition 5 (Properties of the A- and B-coefficients)
If both paths of the pair (γ_1,γ_2) contain edge e, then this is in both paths of the pairs (γ_1,γ_1) and (γ_2,γ_2). Therefore the A-coefficient can be rewritten as

$$A_e = \sum_{(\gamma_1,\gamma_2)\in\Gamma_{ji}^e\times\Gamma_{ji}^e} \left(p_{\gamma_1} + \mathrm{sgn}_{\gamma_1}(e)\mathrm{sgn}_{\gamma_2}(e)p_{\gamma_2}\right)^2, \tag{23}$$

where each term of the sum is non-negative. Similar considerations lead to express the sum of B-coefficients as

$$B_{e_1e_1} + B_{e_2e_2} + B_{e_1e_2} = \sum_{(\gamma_1,\gamma_2)\in\Gamma_{ji}^{e_1}\times\Gamma_{ji}^{e_2}} \left(p_{\gamma_1} + \mathrm{sgn}_{\gamma_1}(e)\mathrm{sgn}_{\gamma_2}(e)p_{\gamma_2}\right)^2. \tag{24}$$

Each term in the sums of (23) and (24) are non-negative and become zero if the sets $\Gamma_{ji}^e, \Gamma_{ji}^{e_1}, \Gamma_{ji}^{e_2}$ are empty. □

Proof of Theorem 4 (Canonical decomposition of the F_2-statistics along paths) Rewrite the definition of $F_2(i,j)$ using the canonical decomposition of drifts:

$$F_2(i,j) = E\left(\sum_{\gamma_1 \in \Gamma_{ji}} p_{\gamma_1} \sum_{e_1 \in \gamma_1} d_{e_1}^{\gamma_1} \sum_{\gamma_2 \in \Gamma_{ji}} p_{\gamma_2} \sum_{e_2 \in \gamma_2} d_{e_2}^{\gamma_2} \right). \tag{25}$$

Distributing the products and exploiting the linearity of the expectation, (25) is equivalent to

$$F_2(i,j) = \sum_{\gamma_1,\gamma_2 \in \Gamma_{ji}} \sum_{e_1 \in \gamma_1} \sum_{e_2 \in \gamma_2} p_{\gamma_1} p_{\gamma_2} E(d_{e_1}^{\gamma_1} d_{e_2}^{\gamma_2}). \tag{26}$$

Note that $\Gamma_{ji} = \cup_{e \in \mathscr{E}_{ji}} \Gamma_{ji}^e$. Therefore (26) can be rewritten as follows

$$F_2(i,j) = \sum_{e_1,e_2 \in \mathscr{E}_{ji}} \sum_{(\gamma_1,\gamma_2) \in \Gamma_{ji}^{e_1} \times \Gamma_{ji}^{e_2}} p_{\gamma_1} p_{\gamma_2} E(d_{e_1}^{\gamma_1} d_{e_2}^{\gamma_2}).$$

Observe that $\mathscr{E}_{ji}$ is partitioned by $\mathscr{E}_{ji}^d, \mathscr{E}_{ji}^u$. Moreover, the product of two distinct edges (where at least one is undirected) has expectation zero (see Proposition 4). Thus

$$\begin{aligned} F_2(j,i) &= \sum_{e \in \mathscr{E}_{ji}^d} \sum_{(\gamma_1,\gamma_2) \in \Gamma_{ji}^e \times \Gamma_{ji}^e} \mathrm{sgn}_{\gamma_1}(e)\mathrm{sgn}_{\gamma_2}(e) p_{\gamma_1} p_{\gamma_2} E(d_e^2) \\ &+ \sum_{e_1,e_2 \in \mathscr{E}_{ji}^u} \sum_{(\gamma_1,\gamma_2) \in \Gamma_{ji}^{e_1} \times \Gamma_{ji}^{e_2}} \mathrm{sgn}_{\gamma_1}(e_1)\mathrm{sgn}_{\gamma_2}(e_2) p_{\gamma_1} p_{\gamma_2} E(d_{e_1} d_{e_2}) \\ &= \sum_{e \in \mathscr{E}_{ji}^d} A_e \mathbb{E}(d_e^2) + \sum_{e_1,e_2 \in \mathscr{E}_{ji}^u} B_{e_1,e_2} E(d_{e_1} d_{e_2}). \end{aligned}$$

□

Proof of Proposition 6 (Additivity of the F_2-statistic)
Assume any path $\gamma \in \Gamma_{ji}$ is of type $(j, j_{\ell-1}, \ldots, k, \ldots, j_0, j_0', j_1', \ldots, j_{\ell'-1}', i)$, where $\ell, \ell' \geq 1$ and eventually $k = j_0$. By simmetry of the F_2-statistic the proof works also for k in the subpath $j_0' \Rightarrow i$. Any two paths $\gamma_1 \in \Gamma_{jk}$, $\gamma_2 \in \Gamma_{ki}$ are such that $\mathscr{E}_{jk}^d \cap \mathscr{E}_{ki}^d = \emptyset$. Therefore

$$\begin{aligned} F_2(i,j) &= \sum_{e \in \mathscr{E}_{ji}^d} A_e \mathbb{E}(d_e^2) + \sum_{e_1,e_2 \in \mathscr{E}_{ki}^u} B_{e_1 e_2} \mathbb{E}(d_{e_1} d_{e_2}) \\ &= \underbrace{\sum_{e \in \mathscr{E}_{jk}^d} A_e \mathbb{E}(d_e^2)}_{F_2(k,j)} + \underbrace{\sum_{e' \in \mathscr{E}_{ki}^d} A_{e'} \mathbb{E}(d_{e'}^2) + \sum_{e_1,e_2 \in \mathscr{E}_{ki}^u} B_{e_1 e_2} \mathbb{E}(d_{e_1} d_{e_2})}_{F_2(i,k)}. \end{aligned}$$

□

Proof of Theorem 5 (Necessary and sufficient condition for F_2 being a metric)
Consider $k \notin \mathscr{V}_{ij}$. Then $F_3(k;i,j) \geq 0$ because edges overlap only with the same sign in two paths of Γ_{ki} and Γ_{kj}, thus $F_2(i,j)$ is a measure between i,j.
Assume $k \in \mathscr{V}_{ij}$. Then $F_3(k;i,j)$ can be written as

$$\sum_{(\gamma,\delta) \in \Gamma_{k;ij}} p_\gamma p_\delta \sum_{e \in \mathscr{E}_{\gamma,\delta}} E(d_e^2),$$

where $\mathscr{E}_{\gamma,\delta}$ is the set of edges that overlap on γ and δ. Such a set can be partitioned in the edges overlapping with identical and opposite signs, that is, $\mathscr{E}_{\gamma,\delta} = \mathscr{E}^{+}_{\gamma,\delta} \cup \mathscr{E}^{-}_{\gamma,\delta}$. Therefore

$$F_3(k;i,j) = \sum_{(\gamma,\delta)\in\Gamma_{k;ij}} p_\gamma p_\delta \sum_{e\in\mathscr{E}^{+}_{\gamma,\delta}} E(d_e^2) - \sum_{(\gamma,\delta)\in\Gamma_{k;ij}} p_\gamma p_\delta \sum_{e\in\mathscr{E}^{-}_{\gamma,\delta}} E(d_e^2),$$

where the subtraction is due to the opposite sign of edges in $\cup\mathscr{E}^{-}_{\gamma,\delta}$. The theorem is proved by setting $F_3(k;i,j) \geq 0$. □

References

1. Inference of Population Splits and Mixtures from Genome-Wide Allele Frequency Data. PLoS Genetics **8**(11), 1–17 (2012)
2. Castelo, R., Roverato, A.: A Robust Procedure For Gaussian Graphical Model Search From Microarray Data With p Larger Than n. Journal of Machine Learning Research **7**, 2621–2650 (2006)
3. Cavalli-Sforza, L.L.: Population structure and human evolution. Proceedings of the Royal Society of London. Series B, Biological sciences **164**, 362–79 (1966)
4. Cavalli-Sforza, L.L., Edwards, A.W.: Phylogenetic analysis. Models and estimation procedures. American journal of human genetics **19**, 233–57 (1967)
5. Chikhi, L., Bruford, M.W., Beaumont, M.A.: Estimation of Admixture Proportions: A Likelihood-Based Approach Using Markov Chain Monte Carlo. Genetics **158**, 1347–1362 (2001)
6. Frydenberg, M.: The chain graph markov property. Scandinavian Journal of Statistics **17**, 333–353 (1990)
7. Green, R.E., Krause, J.e.a.: A draft sequence of the Neandertal genome. Science **328**, 710–22 (2010)
8. Griffiths, R.C., Tavare, S.: Ancestral Inference in Population Genetics. Statistical Science **9**(3), 307–319 (1994). DOI 10.1214/ss/1177010378
9. Harary, F.: Graph Theory. Addison Wesley (1969)
10. Hudson, R.R.: Gene genealogies and the coalescent process. Oxford Surveys in Evolutionary Biology **7**, 1–44 (1990)
11. Kingman, J.: The coalescent. Stochastic Processes and their Applications **13**(3), 235–248 (1982). DOI 10.1016/0304-4149(82)90011-4
12. Lauritzen, S., Richardson, T.: Chain graph models and their causal interpretations. J. R. Statist. Soc. **64**, 321–361 (2002)
13. Lipson, M., Loh, P.R., Levin, A., Reich, D., Patterson, N., Berger, B.: Efficient moment-based inference of admixture parameters and sources of gene flow. Molecular Biology and Evolution **30**, 1788–1802 (2013)
14. Metzker, M.L.: Sequencing technologies the next generation. Nature Reviews Genetics **11**(1), 31–46 (2010). DOI 10.1038/nrg2626
15. Nei, M.: Molecular evolutionary genetics. Columbia University Press (1987)
16. Nielsen, R.: A likelihood approach to populations samples of microsatellite alleles. Genetics **146**(2), 711–6 (1997)
17. Patterson, N.J., Moorjani, P., Luo, Y., Mallick, S., Rohland, N., Zhan, Y., Genschoreck, T., Webster, T., Reich, D.: Ancient Admixture in Human History. Genetics **192**, 1065–1093 (2012)
18. Peter, B.M.: Admixture, Population Structure and F -statistics. Genetics **202**, 1485–1501 (2016)
19. Reich, D., Thangaraj, K., Patterson, N., Price, A., Singh, L.: Reconstructing Indian population history. Nature **461**, 489–94 (2009)
20. Reuter, J., Spacek, D.V., Snyder, M.: High-Throughput Sequencing Technologies. Molecular Cell **58**(4), 586–597 (2015). DOI 10.1016/j.molcel.2015.05.004
21. Skoglund, P., Mallick, S., Bortolini, M.C., Chennagiri, N., Hünemeier, T., Petzl-Erler, M.L., Salzano, F.M., Patterson, N., Reich, D.: Genetic evidence for two founding populations of the Americas. Nature **525**, 104–108 (2015)
22. Wall, J.D., Yang, M.A., Jay, F., Kim, S.K., Durand, E.Y., Stevison, L.S., Gignoux, C., Woerner, A., Hammer, M.F., Slatkin, M.: Higher levels of Neanderthal ancestry in east Asians than in Europeans. Genetics **194**, 199–209 (2013)

23. Wang, J.: Maximum-Likelihood Estimation of Admixture Proportions From Genetic Data. Genetics **164**, 747–765 (2003)
24. Whittaker, J.: Graphical models in applied multivariate statistics. Wiley (1990)
25. Wilson, I.J., Balding, D.J., Griffiths, R.C., Donnelly, P.: Genealogical inference from microsatellite data. Genetics **150**(1), 499–510 (1998)

Manuscript 3

Inference of Chromosomal Ploidy from Short-Read Sequencing Data

Samuele Soraggi, Matteo Fumagalli

Status: both manuscript and results are preliminary. Note: This is just a preliminary template, NOT a submission in the journal GENETICS.

Contribution

This preliminary manuscript illustrates a discrete states Hidden Markov Model to infer ploidy numbers from NGS data. In this framework, the emissions of the hidden Markov chain consist of mean coverage and sequenced data. The aim is to use the sequencing depth to detect changes in ploidy, and to assign each variation to the right ploidy number through genotype likelihoods [35]. This overcomes the limitations of other computational techniques that are based on sequencing depth and/or allele frequencies to detect ploidy changes, and that are subject to wrong interpretations when ploidy numbers can be high (such as in plants) [93, 95, 104].

Future perspectives

The results are still preliminary, but show good performances on low-depth samples. However, more idea are still in the process of being implemented and tested. One idea is that this tool could be used to detect Copy Number Variations (CNV). The detection of CNVs could be achieved by detecting changes in ploidy, and see if the changes in depth are compatible with the ploidy suggested by genotype likelihoods, otherwise flag the change in ploidy as CNV.

Moreover, allele frequencies are calculated over all individuals, but the Hidden Markov Model is so far applied to the depth of a single individual. A possible development is to implement the EM algorithm for multiple observations. From another point of view, one could use the Hidden Markov Model on a subset of individuals (with same ploidy number and haploid depth) and develop a test for aneuploidy by comparing the likelihood of the model on different datasets.

All data is reduced into windows of loci to reduce the noise from overdispersed sequenced reads and have more informative values of the probability of sequenced data. However, the window size has to be arbitrarily chosen by the user. This could cause a change of ploidy to happen inside a window. It would be interesting to find a way to automatically choose each window, in order to have a size such that both depth and genotype likelihoods will not be used to infer a single ploidy number, when they actually contain information about two different ones.

As an application of this tool, it is planned to use the data of more than 200 Bd fungi to detect their ploidy numbers. This genus of fungi is acting as parasite on a host populations of frogs in UK, that suffering a heavy loss of biodiversity. Knowing the ploidy numbers could play an important role in determining mechanism of adaptation/speciation of the Bd fungi on the host organisms.

Inference of Chromosomal Ploidy from Short-Read Sequencing Data

Samuele Soraggi[*,1] **and Matteo Fumagalli**[†]
[*]Department of Mathematics, University of Copenhagen, Denmark, [†]Department of Life Sciences, Imperial College London, United Kingdom

ABSTRACT The inference of ploidy numbers from genetic data is an important yet challenging task for deciphering the evolutionary mechanisms underpinning genome evolution. High-throughput sequencing machines are now providing researchers with massive amount of genomic data. However, the data produced is typically affected by large sequencing errors and the assignment of individual genotypes is challenging when a low-depth strategy is employed.

Statistical methods that take genotype uncertainty into account have been introduced, allowing for an accurate estimation of nucleotide diversity even when little data is present. However, most of the available software and approaches are based on classic assumptions of random mating and diploidy. To solve this issue, we propose a novel statistical framework to estimate ploidy from sequencing data, taking into account base qualities and depth, through a Hidden Markov Model.

The method shows good performances in estimating trajectories of ploidy numbers even at low depth (2X) from simulated data. We also discuss how this method can be adopted to perform variant and genotype calling and estimation of summary statistics under an arbitrary number of ploidy directly from genotype likelihoods.

We finally demonstrate the utility of such method for estimating the chromosomal copy number variation in *Batrachochytrium dendrobatis* (Bd) from whole genome sequencing data. Bd is an amphibian fungus that is imposing a huge burden on its host. Genomes of Bd strains have been shown to be highly dynamic, with changes in ploidy observed even over short timescales. By analysing more than 200 samples from worldwide geographical locations, we aim to assess whether such rapid changes in chromosomal number copies are indeed associated to increased virulence. Unveiling how ploidy variation relates to fungal pathogenicity might hold the key for effective molecular monitoring of one of the most threatening epidemics for animal biodiversity.

KEYWORDS Ploidy; Genotipe Likelihoods; Poliploidy; Next Generation Sequencing; Genomics

Introduction

Ploidy number (or ploidy) is the number of sets of chromosomes in a cell. Humans are known to be diploid, but other species are often characterized by a different ploidy. When the ploidy of an organism is higher than two, it is usually referred to as poliploidy. The polyploidy state is often the consequence of hybridization or whole genome duplications, as often observed in plants. For instance, the genus of the perennial *Spartina* is characterized by triploid, hexaploid and dodecaploid species (Ainouche *et al.* 2003).

The changes in ploidy are considered to be playing an essential role in evolution of plants in natural populations (Adams and Wendel 2005) and is probably the most important factor concurring in speciation of plants (Otto and Whitton 2000). Moreover poliploidy can be an advantage for adapting to environmental factors when it causes alterations of the morphology and phenology of the organisms (Soltis and Soltis 2012). Those alterations can happen even as fast as one generation. For instance, poliploidy events have been detected in the ancestry of some types of crops and tomatoes (Schlueter *et al.* 2004), in lineages of the maize (Messing *et al.* 2004; Lai *et al.* 2004), in the common ancestry of cotton types (Rong *et al.* 2004; Blanc and Wolfe 2004) and

doi: 10.1534/genetics.XXX.XXXXXX
Manuscript compiled: Wednesday 31st January, 2018%
[1]Department of Mathematics, University of Copenhagen, Universitetsparken 5, 2100 Copenhagen, Denmark, samuele@math.ku.dk

soybeans (Schlueter *et al.* 2004), and in fungi (Todd *et al.* 2017; Wertheimer *et al.* 2016).

An experimental method to detect ploidy numbers in a genome is by using flow cytometry procedures (Kron *et al.* 2007). Flow cytometry is a high-throughput technique to obtain a quantification of optical properties, such as fluorescence, from particles floating in a special fluid. When flow cytometry is applied to a cell, it is possible to accurately determine the amount of genetic material in the nucleus, and estimate the ploidy number. Modern flow cytometry instruments are very sensible and reliable. However their cost is high (bennett and Leitch 2005; Greilhuber *et al.* 2007) and not justifiable when we are solely interested in the detection of ploidy numbers.

The advances in high-throughput sequencing techniques of the recent years, such as Next Generation Sequencing (NGS) (Goodwin *et al.* 2016; Reuter *et al.* 2015), have rapidly resulted in a vast amount of cost-effective high-throughput data available for a wide range of genetic studies. The available NGS protocols (Goodwin *et al.* 2016; Reuter *et al.* 2015; Metzker 2010) essentially result into an output that consists of short reads whose length is in the order of hundreds of bases, that are further aligned to a reference genome or *de novo* assembled in scaffolds. Many studies based on NGS data rely on low-depth sequencing ($< 10X$) because of cost-efficiency and/or degradation of the samples. Additionally NGS data is affected by a higher sequencing error than the one typical of Sanger sequencing (Ratan *et al.* 2013; Lam *et al.* 2012). These conditions may result in potentially unreliable estimates of allele frequencies in the data, and consequently a poor frequency-based estimation of genotypes.

Many of the current methods for the estimation of ploidy numbers in NGS data are based on analysis of sequencing depth and allele frequencies. For instance, `conPade` (Margarido and Heckerman 2015) detects the ploidy of a given contig/scaffold using allele frequencies. The tool `ploidyNGS` (Augusto Corrêa dos Santos *et al.* 2017) estimates allele frequencies and provides a visualization tool through which ploidy can be assigned. The visual approach is very commonly used to empirically estimate the ploidy (Yoshida *et al.* 2013). `AbsCN-seq` (Bao *et al.* 2014) combines the information on depth and allele counts to estimate, amongst other parameters related to tumor-specific applications, the ploidy from NGS data. Analogous data is applied to cancer cells' data with a different approach in the package `sequenza` (Favero *et al.* 2015).

We propose a method, called `hiddenMarkovPloidy`, dedicated to infer ploidy numbers from NGS data. In our method we build a Hidden Markov Model (HMM) (Cappe *et al.* 2005; Rabiner 1989) with a double set of observations, that consists of sequencing depths and observed reads. The formers are used to detect changes in ploidy, while the latters are based on the genotype likelihoods (Nielsen *et al.* 2011), and contribute in assigning each hidden state to its corresponding ploidy number. Notably this method is able to output the optimal number of ploidy numbers given an arbitrary initial interval of ploidies.

Simulations at haploid depth 2X show good performances in estimating ploidy numbers as high as five. We believe that this implementation can be also applied to the detection of Copy Number variants (CNV). Tools such as `CNVnator` Abyzov *et al.* (2011), `HadoopCNV` Yang *et al.* (2017) and `CNVfinder` Mccallum and Wang (2013) detect CNVs using sequencing depth and eventually allele frequencies. Here, we aim at using sequencing depth to detect changes in ploidy, and guess the levels based only on depths. Further, we can use genotype likelihoods to compare the guess on ploidy numbers to the ones estimated from genotype likelihoods, and flag the loci where those two estimates are different.

Emerging infectious diseases caused by fungi are a serious threat to global biodiversity and food security. The chytrid fungus *Batrachochytrium dendrobatidis* (Bd) is responsible for the dramatic decline of amphibians worldwide, causing one of the largest losses of biodiversity in recent times Fisher *et al.* (2012). Despite much interest, the genetic mechanisms that underpin Bd's virulence are not yet known but appear to be driven by a highly dynamic genomic landscape with frequent events of gain/loss of chromosomal copies. The geographic origins and the timing of Bd's spread are yet to be fully unravelled, making this one of the most controversial problems in disease ecology (Fisher 2017). Understanding the genetic mechanisms underlying Bd's virulence through an accurate mapping of ploidy numbers at different lineages is a fundamental goal to plan molecular monitoring.

Materials and Methods

This section describes the statistical framework in which the data is modelled and the Hidden Markov Model is built. In what follows data is assumed to be diallelic, without loss of generality.

Consider N sequenced individuals with M sequenced bases. Only the loci that are covered by at least one of the genomes are considered. For $i \in 1,\dots,M$, and $j \in 1,\dots,N$, let $Y_{j,i}$ be the ploidy number and $G_{j,i}$ be the genotype of individuals j at locus i. Denote with S_y the set of possible genotypes with ploidy $Y_{j,i} = y$, expressed as

$$S^y_{j,i} = \{0, 1, ..., y\},$$

where $\{0, 1, ..., y\}$ is the number of alternate (or derived) alleles per genotype.

Probability of Sequenced Data

Denote by O the sequenced data, and consider it independent between loci and individuals. Let $R_{j,i}$ the number of sequenced reads at locus i for individual j and $O_{j,i,r}$ be the r-th sequenced read for individual j at locus i, for $j = 1,\dots,N$, $i = 1,\dots,M$ and $r = 1,\dots,R_{j,i}$. Denote with $O_{j,i,*}$ all the sequenced reads of individual j at locus i. The probability of $O_{j,i,*}$ conditionally on the ploidy number $Y_{j,i} = y_{j,i}$, the alternate allele frequency x_i at locus i and the inbreeding coefficient I_j of individual j is expressed by

$$p(O_{j,i,*}|y_{j,i}, x_i, I_j) = \sum_{g_{j,i} \in S^y_{j,i}} p(O_{j,i,*}|g_{j,i}, y_{j,i}) p(g_{j,i}|y_{j,i}, x_i, I_j), \quad (1)$$

where the left-hand side of the equation has been marginalized over the genotypes, and the resulting probabilities have been rewritten as product of two terms using the tower property of the probability. The first factor of the product is the genotype likelihood (Nielsen *et al.* 2011); the second factor is the probability of the genotype given the frequency, the ploidy and the inbreeding coefficient. Throughout the analysis carried out in this paper, we assume absence of inbreeding and model such a probability with a binomial distribution.

Genotype Likelihood for Arbitrary ploidy number

The genotype likelihood is the probability of observing genotype $g_{j,i}$ for individual j at locus i, for $j = 1, \dots, N$, and $i = 1, \dots, M$, given the observed data. In its simplest formulation the genotype likelihood is determined considering the individual's base qualities as probabilities of incorrect sequenced bases, and assuming independence of the bases through the reads.
Let $R_{j,i}$ be the number of sequencing reads at a locus i for individual j, $O_{j,i,*}$ the individuals's observed data at that locus, $o_{j,i,r}$ and $q_{j,i,r}$ the observed nucleotide and the Phred base quality for the individual's read r at locus i, respectively. The i-th base of genotype g is denoted by g_i, $i \in 1, \dots, y$. The genotype likelihood of $g_{j,i}$ for ploidy number $y_{j,i}$ is expressed as

$$\ln p(O_{j,i,*}|g_{j,i}, y_{j,i}) = \sum_{r=1}^{R} \ln \Big(\sum_{i=1}^{y_{j,i}} \frac{1}{y_{j,i}} p(o_{j,i,r}|g_{j,i}, q_{j,i,r}, y_{j,i}) \Big)$$

where

$$p(o_{j,i,r}|g_{j,i}, q_{j,i,r}, y_{j,i}) = \begin{cases} 1 - \epsilon_{j,i,r}, & \text{if } o_{j,i,r} = g_{j,i} \\ \frac{\epsilon_{j,i,r}}{3} & \text{otherwise} \end{cases}$$

and $\epsilon_{j,i,r}$ is the Phred probability related to the score $q_{j,i,r}$. The probabilities of observing incorrect nucleotides are considered homogeneous through the possible nucleotides.

Consider $L_1, \dots, L_W$ a set of $W > 0$ non-overlapping windows of adjacent loci. We write $i \in L_w$, with abuse of notation, when locus i is in the w-th window, for $i = 1, \dots, M$ and $w = 1, \dots, W$. In each window only loci that are covered by at least one individual are considered. Under the hypothesis that loci are independent and the samples have the same ploidy number in each window, define

$$p_{j,L_w} = \prod_{i \in L_w} \prod_{j=1}^{N} p(O_{j,i,*}|y_{j,i}, x_i, I_j) \tag{2}$$

as the probability of the sequenced data in the w-th window for the j-th samples.

Estimation of population frequencies

If multiple samples are available, the population frequency x_i at each locus $i = 1, \dots, M$ is estimated assuming infinite ploidy. Consider, for each individual $j = 1, \dots, N$, the estimator $\hat{x}_{j,i}$ given by the relative frequency of the A allele. In each individual, the sequenced reads are a sample with replacement from the true genotype.

By assuming infinite ploidy, and therefore an infinitely long genotype for each individual, each sample can be considered as drawn from a different position of the genotype. Hence the reads are considered independent, and the amount of information contained in the estimator $\hat{x}_{j,i}$ is proportional to the number of reads at locus i for individual j. We define the population frequency estimator for x_i, say $\hat{x}_i$, as the weighted sum

$$\hat{x}_i = \sum_{j=1}^{N} \frac{R_{j,i}}{R_{*,i}} \hat{x}_{j,i},$$

where $R_{j,i}$ is the number of reads at locus i for individual j, and $R_{*,i} = \sum_{j=1}^{N} R_{j,i}$.

In case the sample size is limited, or even one single sample is analysed, $\hat{x}_i$ is not a valuable estimator of the population size and therefore (1) might be biased. In fact, in the case of a single sample the derived allele frequency provides the genotype, and therefore does not contain additional information. In this case it is thus assumed that the frequency is the same at each locus, in order to approximate the expected population allele frequency over all loci. Under this scenario, we further assume that one of the two alleles can be assigned to an ancestral (e.g. wild-type) state, while the other to a derived (e.g. mutant) state.

Under the standard coalescent model with infinite sites mutations (Tavaré 2004; Ewens 2004), the probability mass function of the population derived allele frequencies x in a sample of N individuals is (Kingman 1982):

$$f_X(x) = \frac{1/x^k}{\sum_{j=1}^{-1} \frac{1}{j^k}}, \tag{3}$$

with X the random variable describing the allele frequency and $k \in (0, \infty)$ being a positive real number, that determines whether the population is deviating from a model of constant population size. For instance, $k = 1$ is equal to the distribution of x under constant population size, while $k > 1$ models a population shrinkage and $k < 1$ population growth. Given the probability distribution (3), the expected derived allele frequency in a population of size n is:

$$E(X) = \sum_{j=1}^{n-1} \frac{x^{-k}}{\sum_{j=1}^{n-1} \frac{1}{w^k}}. \tag{4}$$

Using the expected value of the frequency it is then possible to calculate quantities that involve the allele frequencies when only few samples are available.

Unknown or Uncertain Ancestral Allelic State. One of our main assumptions for the single-sample case is that we know which allele can be assigned to an ancestral state, and which one to a derived state. However, in many practical cases, such assignment is either not possible or associated with a certain level of uncertainty due to, for instance, ancestral polymorphisms or genome from a closely related species not being available. Under these circumstances, we extend our formulation by adding a parameter underlying the probability that the assigned ancestral state is incorrectly identified.

Let us define v as the ancestral state. This can take value in V, the set of the two most likely alleles from $\{A, C, G, T\}$. Assume that the true ancestral state is contained in $\mathcal{V}$.

The log-probability of the data for a single sample is

$$\ln p(O|y) = \sum_{i=1}^{N} \ln \Big(\sum_{v \in V} \sum_{g \in S_y} p(O|g_i, y_i) p(g_i|y_i, v) p(v) \Big) \tag{5}$$

where $p(v)$ denotes the probability that allele v is the ancestral state and is invariant across sites. Note that $\sum_{v \in V} p(v) = 1$. If $p(v) = 0.5$ for each $v \in V$, then (5) refers to the scenario of folded allele frequencies, where each allele is equally probable to be the ancestral state.

Hidden Markov Model for Ploidy Inference

Under the assumption that in each window of loci the ploidy is constant, we infer the ploidy numbers using a hidden markov model (HMM) with double emissions. Let an HMM for ploidy inference be defined by a discrete process

$$\{Y_{j,L_w}, C_{j,L_w}, O_{j,L_w,*}\}_{w=1}^{W},$$

where W is the number of windows of adjacent loci considered. The unobservable chain Y_{j,L_w} represents the unknown ploidy numbers, C_{j,L_w} the observed depth and $O_{j,L_w,*}$ the observed sequenced data for the j-th individual in the w-th window. The transition probabilities of the unknown markov chain are denoted by $\boldsymbol{A} = \{a_{ij}\}_{i,j=1}^{T}$, and the stationary probability of the chain by the vector $\boldsymbol{\pi}$ of length T, where T is the number of ploidy numbers considered in the model.

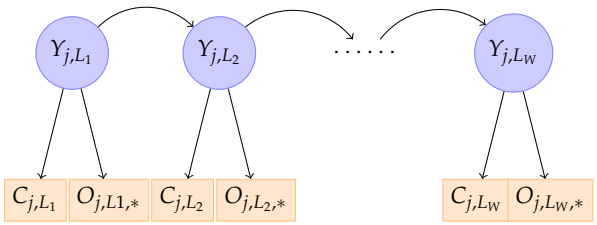


Figure 1 Hidden markov model for the detection of the unknown ploidy numbers Y_{j,L_w} of an individual j in adjacent windows of loci L_w, for $j = 1,\dots,N$ and $w = 1,\dots,W$. The ploidy-dependent emissions consist of the average coverage C_{j,L_w} and the sequenced data $O_{j,L_w,*}$.

Using the HMM defined above implies that some probabilistic relationships are assumed, amongst which:

- conditionally on the sequence of ploidy numbers, the average depth and the data in a window both depend on the ploidy at that window,
- the average depth and the data in a window are conditionally independent, given the ploidy number.

At each window, the average coverage given the ploidy number is modelled by a negative binomial distribution to capture the behaviour of overdispersed values. The observed data given the ploidy number at a certain window is described by the probability in equation (2).

The estimation of the parameters $\boldsymbol{A},\boldsymbol{\pi},\boldsymbol{\theta}$, where $\boldsymbol{\theta}$ characterizes the ploidy-dependent distributions of the depth, is performed through the EM-algorithm (Cappe *et al.* 2005; Rabiner 1989). The EM-algorithm is modified using the AIC criterion (Bishop 2006) to find the optimal number of ploidy numbers by following an approach similar to the one (Li and Biswas 1999). Here the EM algorithm is started with the maximum number of hidden states T of the Markov chain. When the convergence criteria of the EM procedure is met, one of the states is removed and the EM algorithm restarted. If the AIC criteria suggests that removing a state is not necessary, then the optimal number of states is found.

The genotype likelihoods solve the problem of the identifiability of the states (given T hidden states of the chain, there are $T!$ relabeled HMMs that provides the same result with the EM algorithm) (Rabiner 1989; Bishop 2006). The optimal sequence of ploidy numbers is inferred using the Viterbi algorithm, that detects the most probable sequence of ploidies once the parameters of the model have been optimized (Rabiner 1989; Bishop 2006; Viterbi and A. 1967; Forney 1973).

Simulations

To assess the accuracy of estimating ploidy from sequencing data, mapped reads in *mpileup* format are simulated for different scenarios of haploid depth and changes in ploidy numbers. Each site i, for $i = 1,\dots,M$, is treated as an independent observation, without modelling the effect of linkage disequilibrium. The number of reads is distributed as a Poisson(cy_i), where c is the haploid depth and y_i the ploidy at locus i.

At each locus, individual genotypes are randomly drawn according to a probability distribution defined by set of population parameters (e.g. shape of the site frequency spectrum). Once genotypes are assigned, sequencing reads (i.e. nucleotides' bases) are sampled with replacement with a certain probability. Such a probability is given by the quality scores.

All simulated configurations involve 20 individuals, known ancestral allele and absence of inbreeding. In the simulated scenario, 10^4 loci are simulated in two situations, with haploid depth $0.5X$ and $2X$. Here the ploidy changes every 1000 loci increasing from 1 to 5, and decreasing from 5 to 1.

Real Data

We applied our method to detect ploidy numbers to whole-genome sequencing data of Bd strains (Farrer *et al.* 2013). The assembled genome is 20Mbp long comprising more than 20 supercontigs. We first investigated changes in ploidy for a sample previously discovered to be highly variables in chromosomal copies. We will then aim at analysing more than 200 samples of Bd for different geographical locations, comprising the suggestive source of the panzootic (South Africa, North America, South America, Japan and East Asia).

Results and Discussion

In both simulations and real data scenarios, non overlapping windows with size of ten loci are used. In those, only the loci where the allele frequencies estimated with ANGSD fall in the interval $[0.1, 0.9]$ are selected.

In the simulated scenario of Figure 2, the Hidden Markov Model is able to recognize the simulated ploidy numbers from 1 to 4 with few errors at depth $0.5X$. However it does not identify ploidy number 5. This is likely due to two causes. The first is a poor estimation of allele frequencies from low-depth samples, causing the probability of observed data to be maximum for a lower ploidy number. Indeed, the higher is the ploidy number, the easier is that the bias on allele frequencies confound the selection of the correct ploidy value. The second cause is the lack of reads, and therefore the difficulty in inferring some of the genotypes. In fact, in some loci the number of reads available from the 20 individuals might be lower than five. However, the case of depth $0.5X$ is extreme and real data is in general at higher depth. Only minor issues are observed in case of haploid depth $2X$, where ploidy is estimated correctly except in few windows for levels 4 and 5.

Figure 3 shows the performances on 15 contigs of one strain of Bd. Each window of loci has a size of 50Kb. In the graph representing the depth, red lines represent the mean of the depth distribution. The bottom plot shows the minor allele frequencies estimated with `ANGSD` as an additional sanity check. Here the inferred ploidies are compatible with the ones that can be deduced by visual analyse at the sequencing depth variation. Minor errors observable are caused by oscillations in the sequencing depth.

Note that the frequency estimation needs a high number available samples, especially at low depth. In case of few samples, or even only one, are available, the use of the expect frequency

over all loci calculated in (4) is an alternative to estimate the frequencies used in the Hidden Markov Model framework.

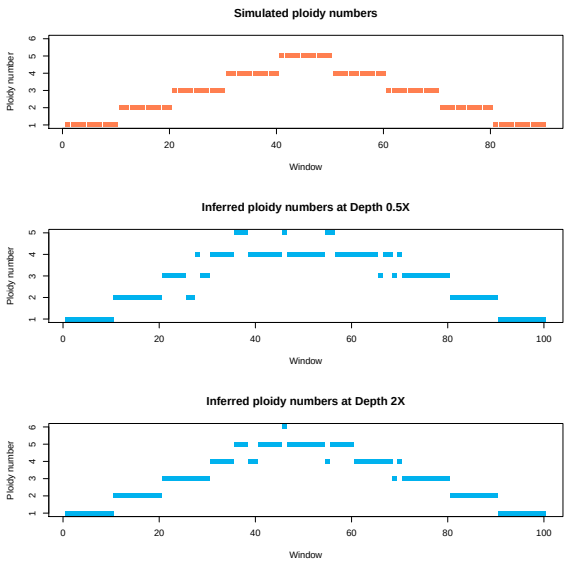


Figure 2 Ploidy inference from simulated data. Inference of simulated ploidy numbers (red), where the ploidy changes from 1 to 5 and is constant in each window of loci. In all plots the window size is 10 loci. The results are shown in blue dots for depth 0.5X and 2X.

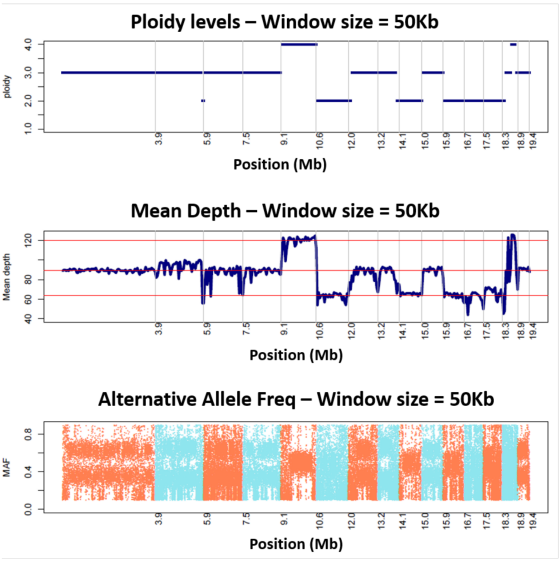


Figure 3 Ploidy inference from a strain of the Bd fungi. Inference of ploidy numbers from a strain of the Bd fungi. For each window of loci of size 50Kb, the plot shows the inferred ploidies, the average sequencing depth and the estimated minor allele frequencies.

Literature Cited

Abyzov, A., A. E. Urban, M. Snyder, and M. Gerstein, 2011 CNVnator: An approach to discover, genotype, and characterize typical and atypical CNVs from family and population genome sequencing. Genome Research **21**: 974–984.

Adams, K. L. and J. F. Wendel, 2005 Polyploidy and genome evolution in plants. Current Opinion in Plant Biology **8**: 135–141.

Ainouche, M. L., A. Baumel, A. Salmon, and G. Yannic, 2003 Hybridization, polyploidy and speciation in Spartina (Poaceae). New Phytologist **161**: 165–172.

Augusto Corrêa dos Santos, R., G. H. Goldman, and D. M. Riaño-Pachón, 2017 ploidyNGS: visually exploring ploidy with Next Generation Sequencing data. Bioinformatics **33**: 2575–2576.

Bao, L., M. Pu, and K. Messer, 2014 AbsCN-seq: a statistical method to estimate tumor purity, ploidy and absolute copy numbers from next-generation sequencing data. Bioinformatics **30**: 1056–1063.

bennett, M. D. and I. J. Leitch, 2005 Nuclear DNA Amounts in Angiosperms: Progress, Problems and Prospects. Annals of Botany **95**: 45–90.

Bishop, C. M., 2006 *Pattern recognition and machine learning*. Springer.

Blanc, G. and K. H. Wolfe, 2004 Widespread paleopolyploidy in model plant species inferred from age distributions of duplicate genes. The Plant cell **16**: 1667–78.

Cappe, O., E. Moulines, and T. Ryden, 2005 *Inference in Hidden Markov Models*. Springer Science+Business Media, Inc.

Ewens, W. J., 2004 *Mathematical population genetics : 1. Theoretical introduction*. Springer.

Farrer, R. A., D. A. Henk, T. W. J. Garner, F. Balloux, D. C. Woodhams, and M. C. Fisher, 2013 Chromosomal Copy Number Variation, Selection and Uneven Rates of Recombination Reveal Cryptic Genome Diversity Linked to Pathogenicity. PLoS Genetics **9**: e1003703.

Favero, F., T. Joshi, A. M. Marquard, N. J. Birkbak, M. Krzystanek, Q. Li, Z. Szallasi, and A. C. Eklund, 2015 Sequenza: allele-specific copy number and mutation profiles from tumor sequencing data. Annals of Oncology **26**: 64–70.

Fisher, M. C., 2017 Ecology: In peril from a perfect pathogen. Nature **544**: 300–301.

Fisher, M. C., D. A. Henk, C. J. Briggs, J. S. Brownstein, L. C. Madoff, S. L. McCraw, and S. J. Gurr, 2012 Emerging fungal threats to animal, plant and ecosystem health. Nature **484**: 186–194.

Forney, G., 1973 The viterbi algorithm. Proceedings of the IEEE **61**: 268–278.

Goodwin, S., J. D. McPherson, and W. Richard McCombie, 2016 Coming of age: ten years of next-generation sequencing technologies .

Greilhuber, J., E. M. Temsch, and J. C. M. Loureiro, 2007 Nuclear DNA Content Measurement. In *Flow Cytometry with Plant Cells*, pp. 67–101, Wiley-VCH Verlag GmbH & Co. KGaA, Weinheim, Germany.

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

Kron, P., J. Suda, and B. C. Husband, 2007 Applications of Flow Cytometry to Evolutionary and Population Biology. Annu. Rev. Ecol. Evol. Syst **38**: 847–76.

Lai, J., J. Ma, Z. Swigonová, W. Ramakrishna, E. Linton, V. Llaca, B. Tanyolac, Y.-J. Park, O.-Y. Jeong, J. L. Bennetzen, and J. Messing, 2004 Gene Loss and Movement in the Maize Genome.

Genome Research **14**: 1924–1931.
Lam, H. Y. K., M. J. Clark, R. Chen, R. Chen, G. Natsoulis, M. O'Huallachain, F. E. Dewey, L. Habegger, E. A. Ashley, M. B. Gerstein, A. J. Butte, H. P. Ji, and M. Snyder, 2012 Performance comparison of whole-genome sequencing platforms. Nature biotechnology **30**: 78.
Li, C. and G. Biswas, 1999 Temporal Pattern Generation Using Hidden Markov Model Based Unsupervised Classification. In *IDA 1999: Advances in Intelligent Data Analysis*, pp. 245–256, Springer, Berlin, Heidelberg.
Margarido, G. R. A. and D. Heckerman, 2015 ConPADE: Genome Assembly Ploidy Estimation from Next-Generation Sequencing Data. PLOS Computational Biology **11**: e1004229.
Mccallum, K. J. and J.-P. Wang, 2013 Quantifying copy number variations using a hidden Markov model with inhomogeneous emission distributions. Biostatistics **14**: 600–611.
Messing, J., A. K. Bharti, W. M. Karlowski, H. Gundlach, H. R. Kim, Y. Yu, F. Wei, G. Fuks, C. A. Soderlund, K. F. X. Mayer, and R. A. Wing, 2004 Sequence composition and genome organization of maize. Proceedings of the National Academy of Sciences of the United States of America **101**: 14349–54.
Metzker, M. L., 2010 Sequencing technologies — the next generation. Nature Reviews Genetics **11**: 31–46.
Nielsen, R., J. Paul, A. Albrechtsen, and Y. Song, 2011 Genotype and snp calling from next-generation sequencing data. Nature Reviews. Genetics **12**: 443–451.
Otto, S. P. and J. Whitton, 2000 Polyploid Incidence and Evolution. Annual Review of Genetics **34**: 401–437.
Rabiner, L., 1989 A tutorial on hidden Markov models and selected applications in speech recognition. Proceedings of the IEEE **77**: 257–286.
Ratan, A., W. Miller, J. Guillory, J. Stinson, S. Seshagiri, and S. C. Schuster, 2013 Comparison of sequencing platforms for single nucleotide variant calls in a human sample. PloS one **8**: e55089.
Reuter, J., D. V. Spacek, and M. Snyder, 2015 High-Throughput Sequencing Technologies. Molecular Cell **58**: 586–597.
Rong, J., C. Abbey, J. E. Bowers, C. L. Brubaker, C. Chang, P. W. Chee, T. A. Delmonte, X. Ding, J. J. Garza, B. S. Marler, C.-h. Park, G. J. Pierce, K. M. Rainey, V. K. Rastogi, S. R. Schulze, N. L. Trolinder, J. F. Wendel, T. A. Wilkins, T. D. Williams-Coplin, R. A. Wing, R. J. Wright, X. Zhao, L. Zhu, and A. H. Paterson, 2004 A 3347-locus genetic recombination map of sequence-tagged sites reveals features of genome organization, transmission and evolution of cotton (Gossypium). Genetics **166**: 389–417.
Schlueter, J. A., P. Dixon, C. Granger, D. Grant, L. Clark, J. J. Doyle, and R. C. Shoemaker, 2004 Mining EST databases to resolve evolutionary events in major crop species. Genome **47**: 868–876.
Soltis, P. S. and D. E. Soltis, 2012 *Polyploidy and genome evolution*. Springer.
Tavaré, S., 2004 *Ancestral Inference in Population Genetics*. Springer, Berlin, Heidelberg.
Todd, R. T., A. Forche, and A. Selmecki, 2017 Ploidy Variation in Fungi: Polyploidy, Aneuploidy, and Genome Evolution. Microbiology spectrum **5**.
Viterbi, A. and A., 1967 Error bounds for convolutional codes and an asymptotically optimum decoding algorithm. IEEE Transactions on Information Theory **13**: 260–269.
Wertheimer, N. B., N. Stone, and J. Berman, 2016 Ploidy dynamics and evolvability in fungi. Philosophical transactions of the Royal Society of London. Series B, Biological sciences **371**.
Yang, H., G. Chen, L. Lima, H. Fang, L. Jimenez, M. Li, G. J. Lyon, M. He, and K. Wang, 2017 HadoopCNV: A Dynamic Programming Imputation Algorithm To Detect Copy Number Variants From Sequencing Data. bioRxiv p. 124339.
Yoshida, K., V. J. Schuenemann, L. M. Cano, M. Pais, B. Mishra, R. Sharma, C. Lanz, F. N. Martin, S. Kamoun, J. Krause, M. Thines, D. Weigel, and H. A. Burbano, 2013 The rise and fall of the Phytophthora infestans lineage that triggered the Irish potato famine. eLife **2**: e00731.

References

1. Mallick, S. *et al.* The Simons Genome Diversity Project: 300 genomes from 142 diverse populations. *Nature* **538,** 201–206. ISSN: 0028-0836 (2016).
2. Campbell, C. D. *et al.* Estimating the human mutation rate using autozygosity in a founder population. *Nature Genetics* **44,** 1277–1281. ISSN: 1061-4036 (2012).
3. Nachman, M. W. & Crowell, S. L. Estimate of the mutation rate per nucleotide in humans. *Genetics* **156,** 297–304. ISSN: 0016-6731 (2000).
4. Fisher, R. A. *Genetical Theory of Natural Selection* (The Clarendon Press, 1930).
5. Wright, S. Evolution in Mendelian Populations. *Genetics* **16,** 97–159. ISSN: 0016-6731 (1931).
6. Tavaré, S. *Ancestral Inference in Population Genetics* 1–188. doi:10.1007/978-3-540-39874-5_1 (Springer, Berlin, Heidelberg, 2004).
7. Ewens, W. J. *Mathematical population genetics : 1. Theoretical introduction* 417. ISBN: 9781441918987 (Springer, 2004).
8. Kimura, M. The Number of Heterozygous Nucleotide Sites Mainained in a Finite Population due to Steady Flux of Mutations. *Genetics* **61** (1969).
9. Tajima, F. Infinite-allele model and infinite-site model in population genetics. *Journal of Genetics* **75,** 27–31. ISSN: 0022-1333 (1996).
10. Nielsen, R. & Slatkin, M. *An introduction to population genetics : theory and applications* ISBN: 1605351539 (Sinauer Associates, 2013).
11. Gillespie, J. H. *Population genetics : a concise guide* 214. ISBN: 9780801880094 (Johns Hopkins University Press, 2004).
12. Reich, D., Thangaraj, K., Patterson, N., Price, A. & Singh, L. Reconstructing Indian Population History. *Nature* **461,** 489–494 (2009).
13. Patterson, N. J. *et al.* Ancient Admixture in Human History. *Genetics* **192,** 1065–1093 (2012).
14. Reich, D., Thangaraj, K., Patterson, N., Price, A. & Singh, L. Reconstructing Indian population history. *Nature* **461,** 489–94 (2009).
15. Green, R. E. *et al.* A Draft Sequence of the Neandertal Genome. *Science* **328,** 710–722. ISSN: 0036-8075 (2010).
16. Skoglund, P. *et al.* Genetic evidence for two founding populations of the Americas. *Nature* **525,** 104. ISSN: 0028-0836 (2015).
17. Moreno-Mayar, J. V. *et al.* Terminal Pleistocene Alaskan genome reveals first founding population of Native Americans. *Nature* **553,** 203–207. ISSN: 0028-0836 (2018).
18. Wall, J. D. *et al.* Higher levels of Neanderthal ancestry in east Asians than in Europeans. *Genetics* **194,** 199–209 (2013).
19. Soraggi, S., Wiuf, C. & Albrechtsen, A. Powerful Inference with the D-Statistic on Low-Coverage Whole-Genome Data. *G3 (Bethesda, Md.)* g3.300192.2017. ISSN: 2160-1836 (2017).
20. Pickrell, J. K. & Pritchard, J. K. Inference of Population Splits and Mixtures from Genome-Wide Allele Frequency Data. *PLoS Genet* **8,** 1–17 (Nov. 2012).
21. Lipson, M. *et al.* Efficient Moment-Based Inference of Admixture Parameters and Sources of Gene Flow. *Molecular Biology and Evolution* **30,** 1788–1802 (2013).
22. Black, J. S., Salto-Tellez, M., Mills, K. I. & Catherwood, M. A. The impact of next generation sequencing technologies on haematological research - A review. *Pathogenesis* **2,** 9–16. ISSN: 2214-6636 (2015).

23. Goodwin, S., McPherson, J. D. & Richard McCombie, W. Coming of age: ten years of next-generation sequencing technologies. doi:10.1038/nrg.2016.49 (2016).

24. Metzker, M. L. Sequencing technologies — the next generation. *Nature Reviews Genetics* **11,** 31–46. ISSN: 1471-0056 (2010).

25. Bentley, D. R. *et al.* Accurate whole human genome sequencing using reversible terminator chemistry. *Nature* **456,** 53–59. ISSN: 0028-0836 (2008).

26. Cock, P. J. A., Fields, C. J., Goto, N., Heuer, M. L. & Rice, P. M. The Sanger FASTQ file format for sequences with quality scores, and the Solexa/Illumina FASTQ variants. *Nucleic acids research* **38,** 1767–71. ISSN: 1362-4962 (2010).

27. Li, H. *et al.* The Sequence Alignment/Map format and SAMtools. *Bioinformatics (Oxford, England)* **25,** 2078–9. ISSN: 1367-4811 (2009).

28. Anders, S. & Huber, W. Differential expression analysis for sequence count data. *Genome Biology* **11,** R106. ISSN: 1465-6906 (2010).

29. Reuter, J., Spacek, D. V. & Snyder, M. High-Throughput Sequencing Technologies. *Molecular Cell* **58,** 586–597. ISSN: 10972765 (2015).

30. Rasmussen, M *et al.* Ancient human genome sequence of an extinct Palaeo-Eskimo. *Nature* **463,** 757–762. ISSN: 1476-4687 (Electronic) 0028-0836 (Linking) (2010).

31. Raghavan, M. *et al.* The genetic prehistory of the New World Arctic. *Science* **345.** ISSN: 0036-8075. doi:10.1126/science.1255832 (2014).

32. Ratan, A. *et al.* Comparison of sequencing platforms for single nucleotide variant calls in a human sample. *PloS one* **8,** e55089. ISSN: 1932-6203 (2013).

33. Lam, H. Y. K. *et al.* Performance comparison of whole-genome sequencing platforms. *Nature biotechnology* **30,** 78. ISSN: 1546-1696 (2012).

34. Orlando, L. *et al.* Recalibrating Equus evolution using the genome sequence of an early Middle Pleistocene horse. *Nature* **499,** 74–78 IF:38.597 (2013).

35. Nielsen, R., Paul, J., Albrechtsen, A. & Song, Y. Genotype and SNP calling from next-generation sequencing data. *Nature Reviews. Genetics* **12,** 443–451. ISSN: 1471-0056 (2011).

36. Patterson, N. J. *et al.* Ancient Admixture in Human History. *Genetics.* ISSN: 0016-6731. doi:10.1534/genetics.112.145037 (2012).

37. Raghavan, M. *et al. Nature* **505,** 87–91. ISSN: 0028-0836 (2013).

38. Wall, J. D. *et al.* Higher Levels of Neanderthal Ancestry in East Asians Than in Europeans. *Genetics.* ISSN: 0016-6731. doi:10.1534/genetics.112.148213 (2013).

39. Raghavan, M. *et al.* Genomic evidence for the Pleistocene and recent population history of Native Americans. *Science.* ISSN: 0036-8075. doi:10.1126/science.aab3884 (2015).

40. Rasmussen, M. *et al.* The genome of a Late Pleistocene human from a Clovis burial site in western Montana. *Nature* **506,** 225–229. ISSN: 0028-0836 (2014).

41. Reich, D. *et al.* Genetic history of an archaic hominin group from Denisova Cave in Siberia. *Nature* **468,** 1053–1060. ISSN: 00280836 (Dec. 2010).

42. Reich, D. *et al.* Denisova Admixture and the First Modern Human Dispersals into Southeast Asia and Oceania. *The American Journal of Human Genetics* **89,** 516–528. ISSN: 0002-9297 (2011).

43. Lalueza-Fox, C. & Gilbert, M. T. P. Paleogenomics of archaic hominins. *Current Biology* **21,** R1002–R1009. ISSN: 09609822 (2011).

44. Chatters, J. C. The Recovery and First Analysis of an Early Holocene Human Skeleton from Kennewick, Washington. *American Antiquity* **65,** 291–316. ISSN: 00027316 (2000).

45. Johnson, P. L. F. & Slatkin, M. Accounting for Bias from Sequencing Error in Population Genetic Estimates. *Molecular Biology and Evolution* **25,** 199–206. ISSN: 0737-4038 (2007).

46. Durand, E. Y., Patterson, N., Reich, D. & Slatkin, M. Testing for ancient admixture between closely related populations. *Molecular Biology and Evolution.* doi:10.1093/molbev/msr048 (2011).

47. Pritchard, J., Stephens, M. & Donnelly, P. Inference of population structure using multilocus genotype data. *Genetics* **155,** 945–959 (June 2000).

48. Alexander, D. H., Novembre, J. & Lange, K. Fast model-based estimation of ancestry in unrelated individuals. *Genome Research.* doi:10.1101/gr.094052.109 (2009).

49. Corander, J., Waldmann, P. & Sillanpää, M. J. Bayesian Analysis of Genetic Differentiation Between Populations. *Genetics* **163** (2003).

50. Bansal, V. & Libiger, O. Fast individual ancestry inference from DNA sequence data leveraging allele frequencies for multiple populations. *BMC Bioinformatics* **16,** 4. ISSN: 1471-2105 (2015).

51. Jørsboe, E., Hanghøj, K. & Albrechtsen, A. fastNGSadmix: admixture proportions and principal component analysis of a single NGS sample. *Bioinformatics* **33,** 3148–3150. ISSN: 1367-4803 (2017).

52. Skotte, L., Korneliussen, T. S. & Albrechtsen, A. Estimating Individual Admixture Proportions from Next Generation Sequencing Data. *Genetics* **195,** 693–702. ISSN: 0016-6731 (2013).

53. Dempster, A. P., Laird, N. M. & Rubin, D. B. *Maximum Likelihood from Incomplete Data via the EM Algorithm* 1977. doi:10.2307/2984875.

54. Wu, C. F. J. On the Convergence Properties of the EM Algorithm. *The Annals of Statistics* **11,** 95–103. ISSN: 0090-5364 (1983).

55. Menozzi, P, Piazza, A & Cavalli-Sforza, L. Synthetic maps of human gene frequencies in Europeans. *Science (New York, N.Y.)* **201,** 786–92. ISSN: 0036-8075 (1978).

56. Patterson, N., Price, A. L. & Reich, D. Population structure and eigenanalysis. *PLoS genetics* **2,** e190. ISSN: 1553-7404 (2006).

57. Liao, P., Satten, G. A. & Hu, Y.-J. Robust Inference of Population Structure from Next-Generation Sequencing Data with Systematic Differences in Sequencing. *Bioinformatics.* ISSN: 1367-4803. doi:10.1093/bioinformatics/btx708 (2017).

58. Li, Y. *et al.* Resequencing of 200 human exomes identifies an excess of low-frequency non-synonymous coding variants. *Nature Genetics* **42,** 969–972 IF:35.209. ISSN: 1061-4036 (2010).

59. Nei, M. *Molecular evolutionary genetics* 512 (Columbia University Press, 1987).

60. Kingman, J. The coalescent. *Stochastic Processes and their Applications* **13,** 235–248. ISSN: 0304-4149 (1982).

61. Hudson, R. R. Gene genealogies and the coalescent process. *Oxford Surveys in Evolutionary Biology* **7,** 1–44 (1990).

62. Wilson, I. J., Balding, D. J., Griffiths, R. C. & Donnelly, P. Genealogical inference from microsatellite data. *Genetics* **150,** 499–510. ISSN: 0016-6731 (1998).

63. Nielsen, R. A likelihood approach to populations samples of microsatellite alleles. *Genetics* **146,** 711–6. ISSN: 0016-6731 (1997).

64. Griffiths, R. C. & Tavare, S. Ancestral Inference in Population Genetics. *Statistical Science* **9,** 307–319. ISSN: 0883-4237 (1994).

65. Wang, J. Maximum-Likelihood Estimation of Admixture Proportions From Genetic Data. *Genetics* **164,** 747–765 (2003).

66. Chikhi, L., Bruford, M. W. & Beaumont, M. A. Estimation of Admixture Proportions: A Likelihood-Based Approach Using Markov Chain Monte Carlo. *Genetics* **158,** 1347–1362 (2001).

67. Cavalli-Sforza, L. L. Population structure and human evolution. *Proceedings of the Royal Society of London. Series B, Biological sciences* **164,** 362–79 (1966).

68. Cavalli-Sforza, L. L. & Edwards, A. W. Phylogenetic analysis. Models and estimation procedures. *American journal of human genetics* **19,** 233–57 (1967).

69. Saitou, N & Nei, M. The neighbor-joining method: a new method for reconstructing phylogenetic trees. *Molecular biology and evolution* **4,** 406–25. ISSN: 0737-4038 (1987).

70. Castelo, R. & Roverato, A. A Robust Procedure For Gaussian Graphical Model Search From Microarray Data With p Larger Than n. *Journal of Machine Learning Research* **7,** 2621–2650 (2006).

71. Peter, B. M. Admixture, Population Structure and F -statistics. *Genetics* **202,** 1485–1501 (2016).

72. Jones, B. & West, M. *Covariance Decomposition in Undirected Gaussian Graphical Models* doi:10.2307/20441235.

73. Bandelt, H.-J. & Dresst, A. W. M. Split Decomposition: A New and Useful Approach to Phylogenetic Analysis of Distance Data. *PHYLOGENETICS AND EVOLUTION* **1,** 242–252 (1992).

74. Bandelt, H.-J. & Dress, A. W. M. A Canonical Decomposition Theory for Metrics on a Finite Set. *ADVANCES IN MATHEMATICS* **92,** 47–105 (1992).

75. Schlueter, J. A. *et al.* Mining EST databases to resolve evolutionary events in major crop species. *Genome* **47,** 868–876. ISSN: 0831-2796 (2004).

76. Adams, K. L. & Wendel, J. F. Polyploidy and genome evolution in plants. *Current Opinion in Plant Biology* **8,** 135–141. ISSN: 13695266 (2005).

77. Rieseberg, L. H. HYBRID ORIGINS OF PLANT SPECIES. *Annual Review of Ecology and Systematics* **28,** 359–389. ISSN: 0066-4162 (1997).

78. Ainouche, M. L., Baumel, A., Salmon, A. & Yannic, G. Hybridization, polyploidy and speciation in Spartina (Poaceae). *New Phytologist* **161,** 165–172. ISSN: 0028646X (2003).

79. Marchant, C. J. Evolution in Spartina (Gramineae). **60** (1968).

80. Otto, S. P. & Whitton, J. Polyploid Incidence and Evolution. *Annual Review of Genetics* **34,** 401–437. ISSN: 0066-4197 (2000).

81. Soltis, P. S. & Soltis, D. E. *Polyploidy and genome evolution* ISBN: 3642314414 (Springer, 2012).

82. Levin, D. A. *The Role of Chromosomal Change in Plant Evolution* **292,** 230. ISBN: 0-19-513859-7 (Oxford University Press, 2002).

83. Messing, J. *et al.* Sequence composition and genome organization of maize. *Proceedings of the National Academy of Sciences of the United States of America* **101,** 14349–54. ISSN: 0027-8424 (2004).

84. Lai, J. *et al.* Gene Loss and Movement in the Maize Genome. *Genome Research* **14,** 1924–1931. ISSN: 1088-9051 (2004).

85. Rong, J. *et al.* A 3347-locus genetic recombination map of sequence-tagged sites reveals features of genome organization, transmission and evolution of cotton (Gossypium). *Genetics* **166,** 389–417. ISSN: 0016-6731 (2004).

86. Blanc, G. & Wolfe, K. H. Widespread paleopolyploidy in model plant species inferred from age distributions of duplicate genes. *The Plant cell* **16,** 1667–78. ISSN: 1040-4651 (2004).

87. Todd, R. T., Forche, A. & Selmecki, A. Ploidy Variation in Fungi: Polyploidy, Aneuploidy, and Genome Evolution. *Microbiology spectrum* **5.** ISSN: 2165-0497. doi:10.1128/microbiolspec.FUNK-0051-2016 (2017).

88. Wertheimer, N. B., Stone, N. & Berman, J. Ploidy dynamics and evolvability in fungi. *Philosophical transactions of the Royal Society of London. Series B, Biological sciences* **371.** ISSN: 1471-2970. doi:10.1098/rstb.2015.0461 (2016).

89. Kron, P., Suda, J. & Husband, B. C. Applications of Flow Cytometry to Evolutionary and Population Biology. *Annu. Rev. Ecol. Evol. Syst* **38,** 847–76 (2007).

90. bennett, M. D. & Leitch, I. J. Nuclear DNA Amounts in Angiosperms: Progress, Problems and Prospects. *Annals of Botany* **95,** 45–90. ISSN: 0305-7364 (2005).

91. Greilhuber, J., Temsch, E. M. & Loureiro, J. C. M. in *Flow Cytometry with Plant Cells* 67–101 (Wiley-VCH Verlag GmbH & Co. KGaA, Weinheim, Germany, 2007). ISBN: 9783527610921. doi:10.1002/9783527610921.ch4.

92. Margarido, G. R. A. & Heckerman, D. ConPADE: Genome Assembly Ploidy Estimation from Next-Generation Sequencing Data. *PLOS Computational Biology* **11** (ed Ioshikhes, I.) e1004229. ISSN: 1553-7358 (2015).

93. Augusto Corrêa dos Santos, R., Goldman, G. H. & Riaño-Pachón, D. M. ploidyNGS: visually exploring ploidy with Next Generation Sequencing data. *Bioinformatics* **33,** 2575–2576. ISSN: 1367-4803 (2017).

94. Yoshida, K. *et al.* The rise and fall of the Phytophthora infestans lineage that triggered the Irish potato famine. *eLife* **2,** e00731. ISSN: 2050-084X (2013).

95. Bao, L., Pu, M. & Messer, K. AbsCN-seq: a statistical method to estimate tumor purity, ploidy and absolute copy numbers from next-generation sequencing data. *Bioinformatics* **30,** 1056–1063. ISSN: 1460-2059 (2014).

96. Favero, F. *et al.* Sequenza: allele-specific copy number and mutation profiles from tumor sequencing data. *Annals of Oncology* **26,** 64–70. ISSN: 0923-7534 (2015).

97. Iafrate, A. J. *et al.* Detection of large-scale variation in the human genome. *Nature Genetics* **36,** 949–951. ISSN: 1061-4036 (2004).

98. Feuk, L., Carson, A. R. & Scherer, S. W. Structural variation in the human genome. *Nature Reviews Genetics* **7,** 85–97. ISSN: 1471-0056 (2006).

99. Aitman, T. J. *et al.* Copy number polymorphism in Fcgr3 predisposes to glomerulonephritis in rats and humans. *Nature* **439,** 851–855. ISSN: 0028-0836 (2006).

100. Hollox, E. J. *et al.* Psoriasis is associated with increased β-defensin genomic copy number. *Nature Genetics* **40,** 23–25. ISSN: 1061-4036 (2008).

101. Cappe, O., Moulines, E. & Ryden, T. *Inference in Hidden Markov Models* (Springer Science+Business Media, Inc, 2005).

102. Rabiner, L. A tutorial on hidden Markov models and selected applications in speech recognition. *Proceedings of the IEEE* **77,** 257–286. ISSN: 00189219 (1989).

103. Martin, S. H., Davey, J. W. & Jiggins, C. D. Evaluating the use of ABBA-BABA statistics to locate introgressed loci. *Molecular biology and evolution* **32,** 244–57. ISSN: 1537-1719 (2015).

104. Zhao, M., Wang, Q., Wang, Q., Jia, P. & Zhao, Z. Computational tools for copy number variation (CNV) detection using next-generation sequencing data: features and perspectives. *BMC Bioinformatics* **14,** S1. ISSN: 1471-2105 (2013).