bioRxiv preprint doi: https://doi.org/10.1101/127852; this version posted September 4, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

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

Samuele Soraggi[1*], Carsten Wiuf[1], Anders Albrechtsen[2]

1 Department of Mathematical Sciences, Faculty of Science, University of Copenhagen, Copenhagen, Denmark
2 Department of Mathematical Sciences, Faculty of Science, University of Copenhagen, Copenhagen, Denmark
3 Center for Bioinformatics, Faculty of Science, University of Copenhagen, Copenhagen, Denmark

* samuele@math.ku.dk

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.

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 [2–5, 8, 12–15, 22, 23, 33].

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

bioRxiv preprint doi: https://doi.org/10.1101/127852; this version posted September 4, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

high-throughput sequencing technologies [6], such as Next Generation Sequencing (NGS) technologies [7]. 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 [8], a Neandertal individual [2] and a Paleoamerican individual [9].

There are many different methods for inferring and analyzing admixture events using genome-scale data. Popular methods such as `STRUCTURE` [10] and `ADMIXTURE` [11] 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 Fig 1). A variant of the D-statistic (called the F_4-statistic) was first used in [12] 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 [4].

In the pivotal study [2] 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 [4].

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 [22, 33].

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 [8]. 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 [13].

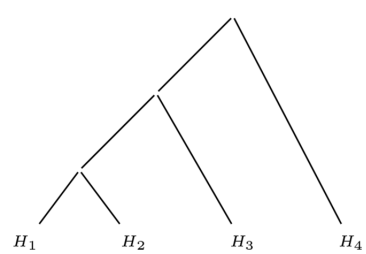


Fig 1. Tree topology for the D-statistic. Hypothesis of genetic relationship between four populations H_1, H_2, H_3, H_4.

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 [2]. This approach is still widely used, and amongst the available computational tools implementing this approach is the `doAbbababa` program of `ANGSD` [16] (supporting low depth NGS data) or the `fourpop` program of `TreeMix` [17] (supporting di-allelic genotype data and microsatellite data). The program `qpDstat` of `ADMIXTOOLS` [15] 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.

bioRxiv preprint doi: https://doi.org/10.1101/127852; this version posted September 4, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

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 Fig 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. 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

bioRxiv preprint doi: https://doi.org/10.1101/127852; this version posted September 4, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

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 Fig 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:

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

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

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

$$\mathcal{H}_0 : \mathbb{E}[(x_1^i - x_2^i)(x_3^i - x_4^i)] = 0 \ \text{ for } \ i = 1, \ldots, M, \tag{3}$$

where the expectation is the difference between eq 1 and eq 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

$$D_M := \frac{X_{(M)}}{Y_{(M)}} = \frac{\sum_{i=1}^{M} (\hat{x}_1^i - \hat{x}_2^i)(\hat{x}_3^i - \hat{x}_4^i)}{\sum_{i=1}^{M} (\hat{x}_1^i + \hat{x}_2^i - 2x_1^i x_2^i)(\hat{x}_3^i + \hat{x}_4^i - 2x_3^i x_4^i)}. \tag{4}$$

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 [21] 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 [2] 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 5 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 [2].

bioRxiv preprint doi: https://doi.org/10.1101/127852; this version posted September 4, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Multiple individuals per group

The D-statistic defined in eq 4 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^i_j, $j = 1, 2, 3, 4$, at locus i for the jth population, we model the observed data as independent binomial trials with parameters n^i_j and x^i_j, where n^i_j 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 $x^i_{j,\ell}$ be the frequency of allele A at locus i, with estimator $\hat{x}^i_{j,\ell}$ the frequency of allele A for $\ell = 1, \ldots, N_j$. Define $\hat{q}^i_j$ as the weighted sum

$$\hat{q}^i_j := \sum_{\ell=1}^{N_j} w^i_{j,\ell} \cdot \hat{x}^i_{j,\ell}, \tag{5}$$

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

$$w^i_{j,\ell} \propto \frac{2n^i_{j,\ell}}{n^i_{j,\ell} + 1}. \tag{6}$$

The estimator $\hat{q}^i_j$ in eq (5) is an estimator for the jth population frequency at locus i with minimal variance (see Appendix 2). Substituting the estimated population frequencies in eq (4) with the weighted estimators determined by eq (5), 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 [18]. 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 [19] 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:

$$\boldsymbol{p}^i_G = \boldsymbol{e}^{-1} \boldsymbol{p}^i_T. \tag{7}$$

where $\boldsymbol{p}^i_G$ and $\boldsymbol{p}^i_T$ are the true and observed vectors of allele frequencies locus i, respectively, and $(\boldsymbol{e}(a,b))_{a,b}$ is considered to be invertible. Note that estimating $(\boldsymbol{e}(a,b))_{a,b}$ and correcting the allele frequencies is a process best applied before the calculation of weighted allele frequencies for multiple individuals.

bioRxiv preprint doi: https://doi.org/10.1101/127852; this version posted September 4, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

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 eq (7) 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 eq (7) 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 eq (7) 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 eq (5), 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:

$$\boldsymbol{E}(a_{1:4}, b_{1:4}) = \boldsymbol{e}_1(a_1, b_1) \cdot \boldsymbol{e}_2(a_2, b_2) \cdot \boldsymbol{e}_3(a_3, b_3) \cdot \boldsymbol{e}_4(a_4, b_4). \tag{8}$$

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 eq (5).

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 eq 7 it holds that

$$\boldsymbol{P}_{corr} = \boldsymbol{E}^{-1}\boldsymbol{P}_{error}. \tag{9}$$

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 eq 6. 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 favour 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

bioRxiv preprint doi: https://doi.org/10.1101/127852; this version posted September 4, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

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 Supplementary Figure 7) 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 eq (5) 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

$$p_{1:4} = (1-\alpha)p_{un} + \alpha p_{out}. \tag{10}$$

By solving for p_{un} it follows that

$$p_{un} = \frac{1}{1-\alpha}(p_{1:4} - \alpha p_{out}). \tag{11}$$

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 Supplementary Figure 7A.

Simulations

Different scenarios have been generated using `msms` [20] to reproduce the trees of Fig 2A, Fig 2B and Fig 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 [20]. Migrations and admixtures, respectively, for the scenarios of Fig 2A and Fig 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 Fig 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.

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 Fig 2B.

bioRxiv preprint doi: https://doi.org/10.1101/127852; this version posted September 4, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

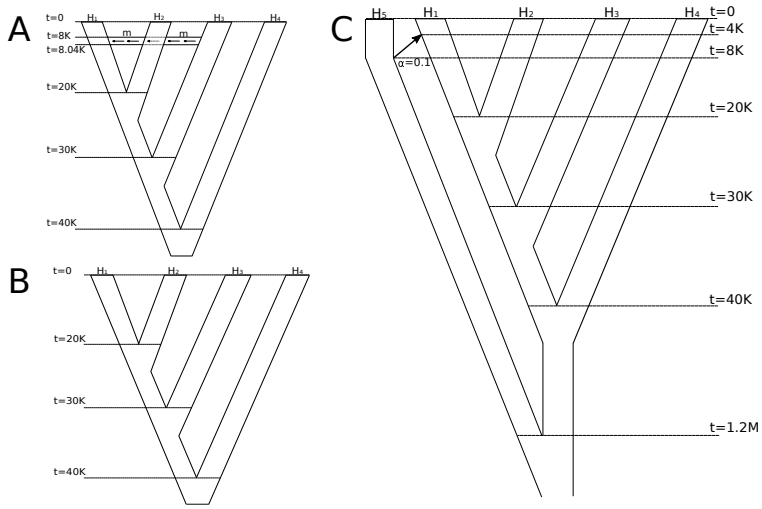


Fig 2. Simulated Scenarios. (A) Simulation of a tree in which migration occurs from population H_3 to H_1. The variable `m` is the (rescaled) migration rate varying between 0, 8, 16, 24, 32, 40 up to 280 with steps of size 20. Expressed in percentage, the migration rate varies between 0%, 0.02%, 0.04%, 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`. (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`.

bioRxiv preprint doi: https://doi.org/10.1101/127852; this version posted September 4, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Sequenced human populations

For the real data scenarios of Fig 3A, Fig 3B and Fig 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` [11]. 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 Supplementary Figure 7 for the barplot of the estimates of the type-specific error).

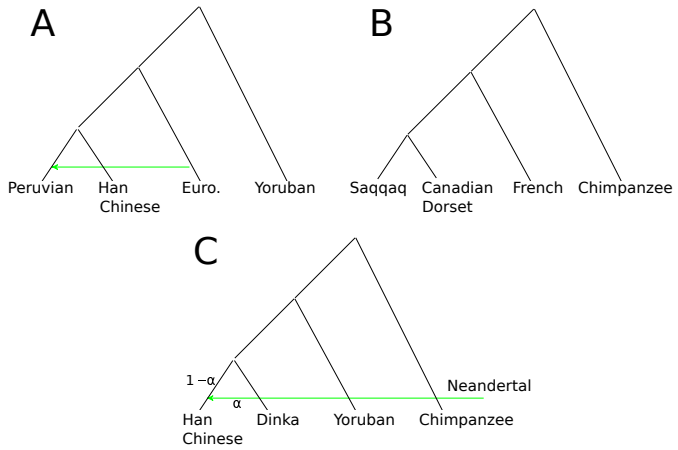


Fig 3. Real Data Scenarios. (A) Tree representing the southwestern European migration into the Americas during the European colonization. (B) Tree representing two independent migrations into northwestern Canada and Greenland. (C) Tree representing the presence of Neandertal genome into a modern non-african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

bioRxiv preprint doi: https://doi.org/10.1101/127852; this version posted September 4, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Genome	Population	Depth	Source
HG01923	Peruvian	6.3X	[28]
HG01974	Peruvian	11.9X	[28]
HG02150	Peruvian	7.3X	[28]
HG02259	Peruvian	6.5X	[28]
HG02266	Peruvian	3.8X	[28]
NA18526	Han Chinese	6.6X	[28]
NA18532	Han Chinese	7.3X	[28]
NA18537	Han Chinese	2.9X	[28]
NA18542	Han Chinese	7.3X	[28]
NA18545	Han Chinese	6.2X	[28]
NA06985	CEPH	12.8X	[28]
NA06994	CEPH	5.5X	[28]
NA07000	CEPH	9.4X	[28]
NA07056	CEPH	4.9X	[28]
NA07357	CEPH	5.7X	[28]
NA12778	CEPH	6.9X	[28]
NA18501	Yoruba	6.4X	[28]
NA18502	Yoruba	4.9X	[28]
NA18504	Yoruba	10.1X	[28]
NA18505	Yoruba	6.1X	[28]
NA18507	Yoruba	3X	[28]
HGDP00778	Han Chinese	23.4X	[29]
DNK02	Dinka	25.8X	[30]
HGDP00927	Yoruban	28X	[29]
AltaiNea	Neanderthal	44.9X	[2]
pantro2	Chimpanzee	-	[31]
saqqaq	Saqqaq	15.7X	[23]
MARC1492	ancient Canadian Dorset Mi'kmaq (New England)	1.1X	[35]
HGDP00521	French	23.8X	[29]

Table 1. List of the Genomes Used in Real Data Scenarios. The table contains the genome identification number, the major population division, the depth calculated using `ANGSD` and the study source of the data.

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, D_{1base} the D-statistic calculated by sampling 1 sequenced base per locus [2] and D_{geno} the D-statistic calculated with equation (4) 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 [21]. The threshold for rejection of the null hypothesis is set to a p-value 0.001, corresponding approximately to the two-tailed acceptance region $[-3, 3]$.

bioRxiv preprint doi: https://doi.org/10.1101/127852; this version posted September 4, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

The formula for calculating the D-statistic is given in eq (4) and finds amongst its current implementations, the ones in [15] and [16], 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 [20]. We simulated without sequencing error four populations with a varying amount of migration from H_3 to H_1 (see Fig 2A) and applied the D-statistic based on five individuals from each population for two different sequencing depths. Fig 4A and Fig 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 behaviour 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.

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 Fig 3A. This scenario represents the southwestern European gene flow into the ancestors of the Native Americans [13]. 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 Fig 4C). The extended D-statistic D_{ext} has much lower standard errors, that corresponds to a smaller

bioRxiv preprint doi: https://doi.org/10.1101/127852; this version posted September 4, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

p-value than in the case of D_{1base}, and therefore a more significant rejection. See Supplementary Table 2 for a better comparison of the values of the different D-statistics.

Error Impact and Correction

Sequencing or genotyping errors are known to have a large impact on the D-statistic [19]. 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 Fig 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 Fig 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 [23, 35], the two ancient genomes have high type-specific error rates as shown in Supplementary Table 3 and Supplementary Figure 7. 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 [22]. In Fig 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. Supplementary Table 4 shows the values related to the four D-statistics in this scenario. Supplementary Figure 9 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 $C \to T$.

bioRxiv preprint doi: https://doi.org/10.1101/127852; this version posted September 4, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

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 [2, 4] 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 Fig 2C, for which an external population H_5 is admixed with H_2 with rate $\alpha = 0.1$. The plot of Fig 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 eq (11) 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 Fig 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 [2, 4]. Green et al. [2] estimated α to be in the range $[0.01, 0.04]$, while Wall et al. [4] estimated it as being $\alpha = 0.0307$ with standard deviation 0.0049. The result is shown in Fig 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 [2]. The black dot coincides with the value of $\alpha = 0.0307$ calculated in [4]. 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 modelled 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.

bioRxiv preprint doi: https://doi.org/10.1101/127852; this version posted September 4, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

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 [2]. 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;

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.

bioRxiv preprint doi: https://doi.org/10.1101/127852; this version posted September 4, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

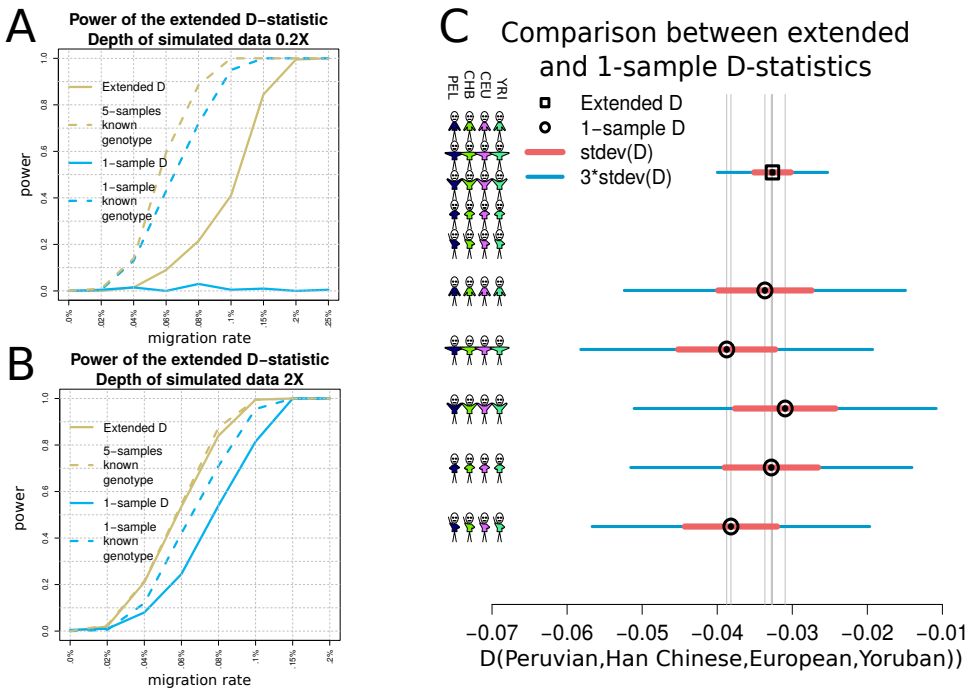


Fig 4. Detection of Admixture and Migration. (A,B) Rejection rate of the null hypothesis as a function of the migration rate in the tree $(((H_1, H_2)H_3)H_4)$, where a migration from H_3 to H_1 occurs. The yellow and blue solid lines represent respectively the power of the method related to D_{ext} 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 eq (4) 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.

bioRxiv preprint doi: https://doi.org/10.1101/127852; this version posted September 4, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

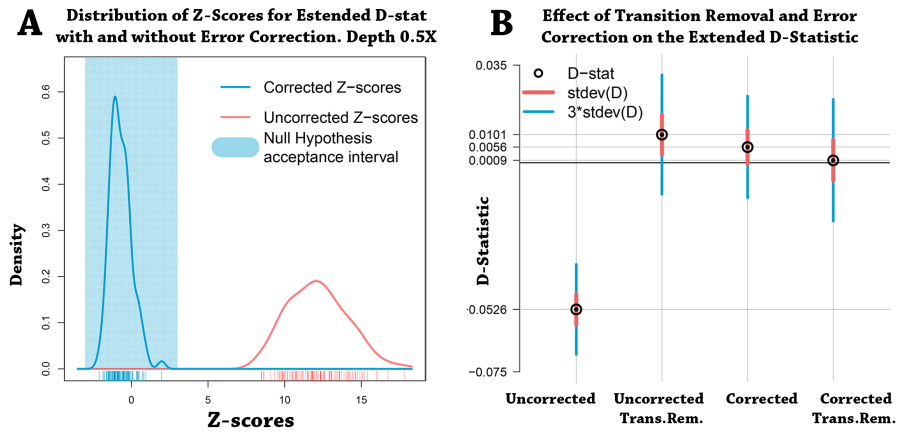


Fig 5. Effect of Error Estimation and Correction. (A) Estimated distributions of the Z-scores related to D_{ext} for the null hypothesis $(((H_1, H_2)H_3)H_4)$ in which H_1, H_3 and H_2 has probability 0.005 and 0.01 of transition from base A, respectively. The blue polygon represents the interval where 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.

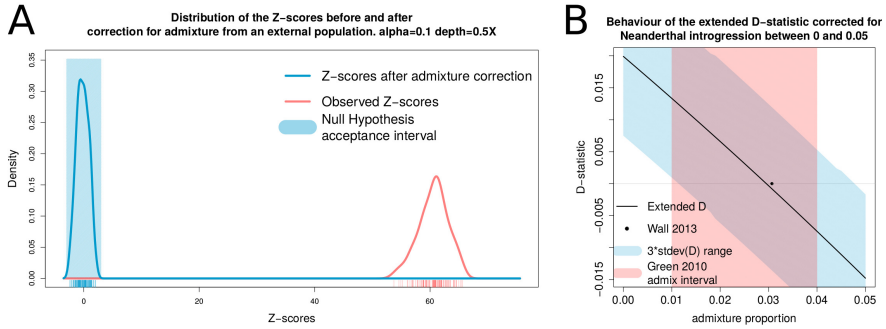


Fig 6. Effect of Correction from External Introgression. (A) Estimated distribution of the Z-scores related to D_{ext} from the 100 simulations of the null hypothesis $(((H_1, H_2)H_3)H_4)$ with introgression of rate $\alpha = 0.1$ from an external population H_5 into H_2. The Z-scores of the observed 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) Behaviour 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 [2] estimates α to fall in. The black dot coincides with the value of $\alpha = 0.0307$ calculated by [4] 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.

bioRxiv preprint doi: https://doi.org/10.1101/127852; this version posted September 4, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

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 Fig 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 $F_i(\boldsymbol{x})$, $i = 1, \ldots, M$ 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 Fig 1 is correct can be rewritten as follow:

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

where the expectation is done on the difference between the probabilities of ABBA and BABA events deduced in eq (1) and eq 2. Using the empirical frequencies as proxies for the expected values, 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 [12]. 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 Fig 1, it holds that* $\mathbb{E}[(x_1^1 - x_2^i)(x_3^1 - x_4^i)] = 0$ *for* $i = 1, \ldots, M$.

bioRxiv preprint doi: https://doi.org/10.1101/127852; this version posted September 4, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Proof. Let $x^i_{1:2}$, $x^i_{1:3}$ and $x^i_{1:4}$ be the frequencies of the ancestral populations of (x^i_1, x^i_2), (x^i_1, x^i_2, x^i_3) and the root of the tree, respectively, as illustrated in Fig 1. Let $\mathcal{X}$ be the set of those three frequencies. Using the martingale properties of the frequencies it follows that

$$\mathbb{E}[(x^i_1 - x^i_2)(x^i_3 - x^i_4)] = \mathbb{E}\Big[\mathbb{E}[(x^i_1 - x^i_2)(x^i_3 - x^i_4)|\mathcal{X}]\Big] \tag{12}$$

$$= \mathbb{E}\Big[\mathbb{E}[x^i_1 - x^i_2|\mathcal{X}]\mathbb{E}[x^i_3 - x^i_4|\mathcal{X}]\Big] \tag{13}$$

$$= \mathbb{E}\Big[\mathbb{E}[x^i_1 - x^i_2|x_{1:2}]\mathbb{E}[x^i_3 - x^i_4|\mathcal{X}]\Big] \tag{14}$$

$$= \mathbb{E}\Big[0 \cdot \mathbb{E}[x^i_3 - x^i_4|\mathcal{X}]\Big] = 0 \tag{15}$$

□

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

$$v_M \to \infty \quad \text{for} \quad M \to \infty. \tag{16}$$

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

$$\frac{1}{M}\sum_{i=1}^M \gamma_i \to \gamma \quad \text{for} \quad M \to \infty. \tag{17}$$

3. Denote with τ_i^2 the variance of Y_i. Then

$$\frac{1}{M^2}\sum_{i=1}^M \tau_i^2 \to 0 \quad \text{for } M \to \infty. \tag{18}$$

If the numerator and denominator are sums of iid variables, conditions (16), (17) and (18) are fullfilled. In fact, if every term X_i has variance σ^2, the sum of variances is $v_M = M\sigma^2$ and eq (16) holds. If every term Y_i has mean and variance γ and τ^2, respectively, eq (17) is still valid because the arithmetic mean is done on identical values. Moreover, eq (18) holds because

$$\frac{1}{M^2}\sum_{i=1}^M \tau^2 = \frac{1}{M}\tau^2, \tag{19}$$

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) [24], that will be applied to the numerator $X_{(M)}$. Subsequently we state the law of large number (LLN) [25] for not iid variables that is used for the denominator $Y_{(M)}$ of the D-statistic. Thereafter we enunciate one of the consequences of Slutsky's theorem [26, 27]. 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)*

bioRxiv preprint doi: https://doi.org/10.1101/127852; this version posted September 4, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

variables with zero mean and variances σ_i^2. Define v_M as $\sum_{i=1}^M \sigma_i^2$. Consider the following quantity

$$\Lambda_\epsilon(M) := \sum_{i=1}^M \mathbb{E}\left[\left(\frac{X_i}{\sqrt{v_M}}\right)^2 \mathbb{I}\left(\left|\frac{X_i}{\sqrt{v_M}}\right| \geq \epsilon\right)\right], \tag{20}$$

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 (17) and (18) are fulfilled, then for each $\epsilon > 0$*

$$\lim_{M\to\infty} \mathbb{P}\Big(|\bar{Y}_M - \frac{1}{M}\sum_{i=1}^M \gamma_i| \geq \epsilon\Big) = 0. \tag{21}$$

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 (16), (17) and (18), the D-statistic converges in distribution to a standard normal if rescaled by the constant:

$$c_M D_M \xrightarrow{d} \mathcal{N}(0,1) \quad \text{for } M \to \infty. \tag{22}$$

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

$$c_M := \gamma \frac{M}{\sqrt{v_M}}. \tag{23}$$

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 aritmetic 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

$$\Lambda_\epsilon(M) \leq \Big(\frac{1}{\sqrt{v_M}}\Big)^2 \sum_{i=1}^M \mathbb{E}\left[\mathbb{I}\left(\left|\frac{1}{\sqrt{v_M}}\right| \geq \epsilon\right)\right] \tag{24}$$

$$= \frac{1}{v_M}\mathbb{P}\left(|X_i| \geq \epsilon\sqrt{v_M}\right) \leq \frac{1}{v_M}\frac{\mathbb{E}[X_i]}{\epsilon\sqrt{v_M}} \leq \frac{1}{v_M}\frac{1}{\epsilon\sqrt{v_M}}, \tag{25}$$

bioRxiv preprint doi: https://doi.org/10.1101/127852; this version posted September 4, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

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 (17) and (18) 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 eq 17. Thus the ratio

$$\frac{M}{\sqrt{v_M}}\frac{X_{(M)}}{Y_{(M)}} \tag{26}$$

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

$$c_M = \gamma\frac{\sqrt{v_M}}{M}. \tag{27}$$

□

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 [2] 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 eq 5 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 [18], 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 [2]. in fact, Theorem 1 still holds because in eq (12) 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

$$\hat{x}_j^i := \frac{n_j^{i,A}}{n_j^i}, \tag{28}$$

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.

bioRxiv preprint doi: https://doi.org/10.1101/127852; this version posted September 4, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

For locus i denote the allele frequency of individual ℓ in population j as $x^i_{j,\ell}$. We use as its unbiased estimator

$$\hat{x}^i_\ell := \frac{n^{i,A}_{j,\ell}}{n^i_{j,\ell}}, \tag{29}$$

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 $\hat{x}^i_\ell$ into a single value $\hat{q}^i_j$ that minimizes the variance of the sum of the estimated individuals' frequencies w.r.t. a set of normalized weights

$$\{w^i_{j,\ell}\}_{\ell=1}^{N_h}, \quad \sum_{\ell=1}^{N_h} w^i_{j,\ell} = 1$$

such that

$$\hat{q}^i_j := \sum_{\ell=1}^{N_h} w^i_{j,\ell} \cdot \hat{x}^i_{j,\ell}. \tag{30}$$

The estimated population frequency $\hat{q}^i_j$ is an unbiased estimator of the frequency of population j at the ith locus [18]. The aim of the weight estimate is to determine the set of weights that minimizes the variance of $\hat{q}^i_j$. 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 $\hat{x}^i_{j,\ell}$, defining

$$Y^i_{j,\ell} := \frac{\sum_{m=1}^{n^i_{j,\ell}} I_m}{n^i_{j,\ell}}, \tag{31}$$

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

$$\mathbb{V}[Y^i_{j,\ell}] = \frac{1}{(n^i_{j,\ell})^2}\left(\sum_{m=1}^{n^i_{j,\ell}} \mathbb{V}[I_m] + 2\sum_{r<t}^{n^i_{j,\ell}} Cov[I_r, I_t]\right). \tag{32}$$

The variance of the indicator function I_m

$$\mathbb{V}[I_m] = x^i_j(1 - x^i_j). \tag{33}$$

It remains to find the covariance

$$Cov[I_r, I_t] = \mathbb{E}[I_r I_t] - \mathbb{E}[I_r]\mathbb{E}[I_t] = \mathbb{E}[I_r I_t] - {x^i_j}^2, \tag{34}$$

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

$$\begin{aligned}
\mathbb{E}[I_r I_t] &= \sum_{g \in \{AA, AB, BB\}} \mathbb{P}(I_r I_t = 1, G = g) \\
&= \mathbb{P}(I_r I_t = 1 | G = AA)\mathbb{P}(G = AA) \\
&+ 2\mathbb{P}(I_r I_t = 1 | G = AB)\mathbb{P}(G = AB) \\
&+ \mathbb{P}(I_r I_t = 1 | G = BB)\mathbb{P}(G = BB) \\
&= 0 + \frac{1}{2} \cdot \frac{1}{2} \cdot 2x^i_j(1 - x^i_j) + 1 \cdot {x^i_j}^2 = \frac{1}{2}x^i_j(1 - x^i_j) + {x^i_j}^2.
\end{aligned} \tag{35}$$

Considering that the sum over $r < t$ in equation (32) is made over $\frac{1}{2}n^i_{j,\ell}(n^i_{j,\ell} - 1)$ equal

bioRxiv preprint doi: https://doi.org/10.1101/127852; this version posted September 4, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

expectations, we can write

$$\begin{aligned}\mathbb{V}[Y^i_{j,\ell}] &= \frac{1}{(n^i_{j,\ell})^2}[n^i_{j,\ell}x(1-x) + 2\frac{n^i_{j,\ell}(n^i_{j,\ell}-1)}{2}\frac{1}{2}x^i_j(1-x^i_j)] \\ &= \frac{1}{(n^i_{j,\ell})^2}[n^i_{j,\ell}x^i_j(1-x^i_j) + 2\frac{n^i_{j,\ell}(n^i_{j,\ell}-1)}{2}\frac{1}{2}x^i_j(1-x^i_j)] \\ &= \frac{n^i_{j,\ell}+1}{2n^i_{j,\ell}}x^i_j(1-x^i_j) = R^i_{j,\ell}x^i_j(1-x^i_j),\end{aligned} \tag{36}$$

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

$$\frac{n^i_{j,\ell}+1}{2n^i_{j,\ell}}. \tag{37}$$

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

$$\mathbb{V}[\hat{x}^i_j] = \sum_{\ell=1}^{N_j}(w^i_{j,\ell})^2\mathbb{V}[\hat{x}^i_{j,\ell}] \approx \sum_{\ell=1}^{N_j}(w^i_{j,\ell})^2\mathbb{V}[Y^i_{j,\ell}]. \tag{38}$$

Our objective is to perform a Lagrange-constrained optimization w.r.t. the weights, being sure to find a minimum since eq (38), as function of the weghts, is convex. This is easily done since the Lagrange-parametrized function is

$$\mathcal{L}(w^i_{j,1:N_j},\lambda) = \sum_{\ell=1}^{N_j}(w^i_{j,\ell})^2 x^i_j(1-x^i_j)R^i_{j,\ell} - \lambda(\sum_{\ell=1}^{N_j} w^i_{j,\ell} - 1) \tag{39}$$

and it originates a linear system of equations of the form

$$\begin{array}{lll} 2\cdot w^i_{j,1} & \cdot x^i_j(1-x^i_j)R^i_{j,1} & -\lambda = 0 \\ \vdots & & \vdots = \vdots \\ 2\cdot w^i_{j,N_j} & \cdot\ x^i_j(1-x^i_j)R^i_{j,N_j} & -\lambda = 0 \\ \sum_{\ell=1}^{N_j} w^i_{j,\ell} & & -1 = 0 \end{array} \tag{40}$$

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

$$w^i_{j,\ell} = \frac{\prod_{m=1,m\neq\ell}^{N_j} R^i_{j,m}}{\sum_{k=1}^{N_j}\prod_{m=1,m\neq k}^{N_j} R^i_{j,m}} = \frac{(R^i_{j,\ell})^{-1}}{\sum_{k=1}^{N_j}(R^i_{j,k})^{-1}}. \tag{41}$$

Appendix 3 Error estimation and correction. Estimation of the type-specific errors follows the supplementary material of [19]. Assume having one observed sequenced individuals affected by base-transition errors. This individual has an associated 4x4 error matrix $(\boldsymbol{e}(a,b))_{a,b}$, such that the entry $\boldsymbol{e}(a,b)$ is the probability of having sequenced allele b instead of allele 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 $(\boldsymbol{e}(a,b))_{a,b}$. Then the likelihood of the base t in the observed individual can be decomposed as a product through the loci:

$$\mathbb{P}(T=t|O=o,\boldsymbol{e}) = \prod_{i=1}^{M}\mathbb{P}(T_i=t_i|O_i=o_i,\boldsymbol{e}). \tag{42}$$

bioRxiv preprint doi: https://doi.org/10.1101/127852; this version posted September 4, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

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:

$$\begin{aligned}
\mathbb{P}(T_i = t_i | O_i = o_i, \boldsymbol{e}) &= \sum_{g_i \in \{A,C,G,T\}} \mathbb{P}(T_i = t_i, G_i = g_i | O_i = o_i, \boldsymbol{e}) && (43)\\
&= \sum_{g_i \in \{A,C,G,T\}} \mathbb{P}(T_i = t_i | G_i = g_i, O_i = o_i, \boldsymbol{e}) \mathbb{P}(G_i = g_i | O_i = o_i) && (44)\\
&= \sum_{g_i \in \{A,C,G,T\}} \boldsymbol{e}(g_i, t_i) \mathbb{P}(G_i = g_i | O_i = o_i), && (45)
\end{aligned}$$

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

$$\mathbb{P}(G_i = g_i | O_i = o_i) = \mathbb{P}(R_i = g_i | O_i = o_i). \tag{46}$$

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 [19] 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

$$\begin{aligned}
\mathbb{P}(T_i = t_i | \boldsymbol{e}_\ell) &= \mathbb{P}(T_i = t_i | \boldsymbol{e}_\ell, G \to t_i) \boldsymbol{e}_\ell(t_i, t_i) && (47)\\
&+ \textstyle\sum_{\tilde{t}_i \neq t_i} \mathbb{P}(T_i = t_i | \boldsymbol{e}_\ell, G = \tilde{t}_i) \boldsymbol{e}_\ell(\tilde{t}_i, t_i). && (48)
\end{aligned}$$

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

$$\boldsymbol{p}_T^i = \boldsymbol{e}_\ell \boldsymbol{p}_G^i, \tag{49}$$

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

$$\boldsymbol{p}_G^i = \boldsymbol{e}_\ell^{-1} \boldsymbol{p}_T^i. \tag{50}$$

The correction performed in eq (50) 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

$$m \sim Bin(n, x), \tag{51}$$

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:

$$\begin{aligned}
m_0 &\sim Bin(m, 1 - \epsilon_{A,B}),\\
m_1 &\sim Bin(n - m, \epsilon_{B,A}).
\end{aligned} \tag{52}$$

The expected population frequency is given by

$$\begin{aligned}
\frac{1}{n}\mathbb{E}[m_0 + m_1] &= \frac{1}{n}\mathbb{E}\big[\mathbb{E}[m_0 | m]\big] + \frac{1}{n}\mathbb{E}\big[\mathbb{E}[m_1 | m]\big]\\
&= x(1 - \epsilon_{A,B}) + (1 - x)\epsilon_{B,A}.
\end{aligned} \tag{53}$$

bioRxiv preprint doi: https://doi.org/10.1101/127852; this version posted September 4, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

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

$$\boldsymbol{e} = \begin{bmatrix} 1-\epsilon_{A,B} & \epsilon_{B,A} \\ \epsilon_{A,B} & 1-\epsilon_{B,A} \end{bmatrix}, \qquad \boldsymbol{e}^{-1} = \frac{1}{C}\begin{bmatrix} 1-\epsilon_{B,A} & -\epsilon_{B,A} \\ -\epsilon_{A,B} & 1-\epsilon_{A,B} \end{bmatrix}, \tag{54}$$

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 eq (50) is rewritten as

$$\begin{bmatrix} \hat{x} \\ 1-\hat{x} \end{bmatrix} = \frac{1}{C}\begin{bmatrix} 1-\epsilon_{B,A} & -\epsilon_{B,A} \\ -\epsilon_{A,B} & 1-\epsilon_{A,B} \end{bmatrix}\begin{bmatrix} \hat{z} \\ 1-\hat{z} \end{bmatrix}, \tag{55}$$

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

$$\hat{z} = \frac{m_0 + m_1}{n}. \tag{56}$$

From eq (55) it is possible to deduce the following equality:

$$\begin{aligned} \mathbb{E}[\hat{x}] &= \frac{1}{C}(1-\epsilon_{B,A})\mathbb{E}[\hat{z}] - \frac{1}{C}(1-\mathbb{E}[\hat{z}])\epsilon_{B,A} \\ &= \frac{1}{C}x(1-\epsilon_{B,A}-\epsilon_{A,B}) = x. \end{aligned} \tag{57}$$

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.

Supplemental Data.

The Supplemental Data contains two tables with numeric results related to a real data scenario, and three figures regarding the estimates of type-specific errors, the behaviour of the D-statistic and the correction for external introgression.

Table 2. European Introgression into Native American Individuals. The table contains the values of the different types of D-statistics used to create the plot of Fig 5C, 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 3. 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.

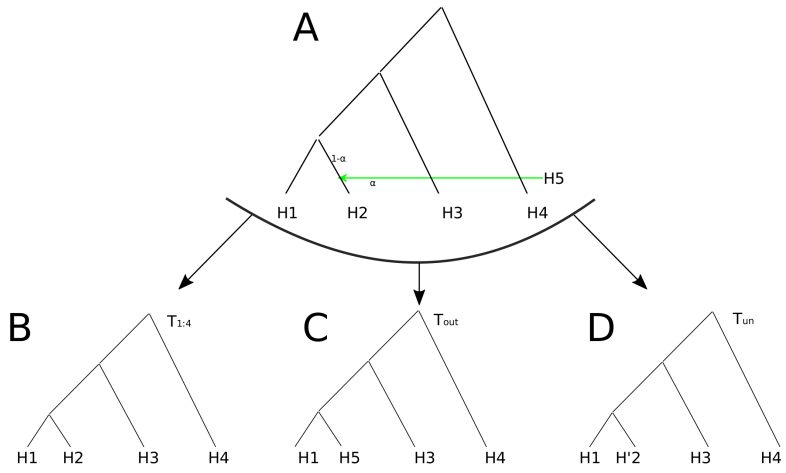
bioRxiv preprint doi: https://doi.org/10.1101/127852; this version posted September 4, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

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 4. 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 3B, 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 7. 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.



bioRxiv preprint doi: https://doi.org/10.1101/127852; this version posted September 4, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Figure 8. 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.

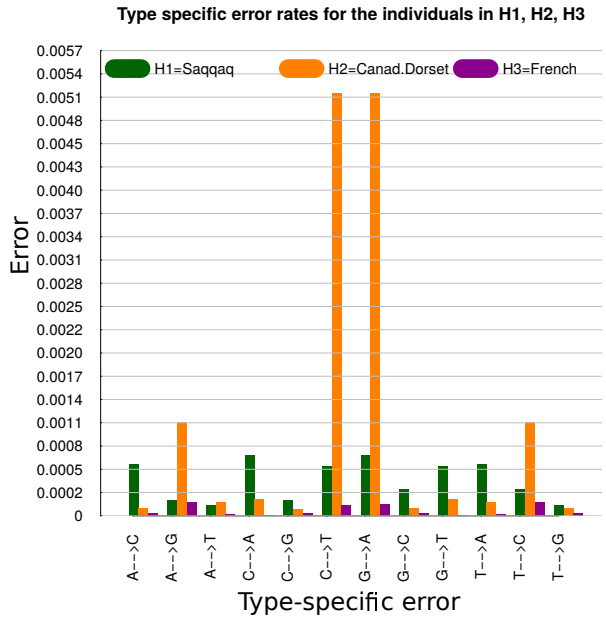
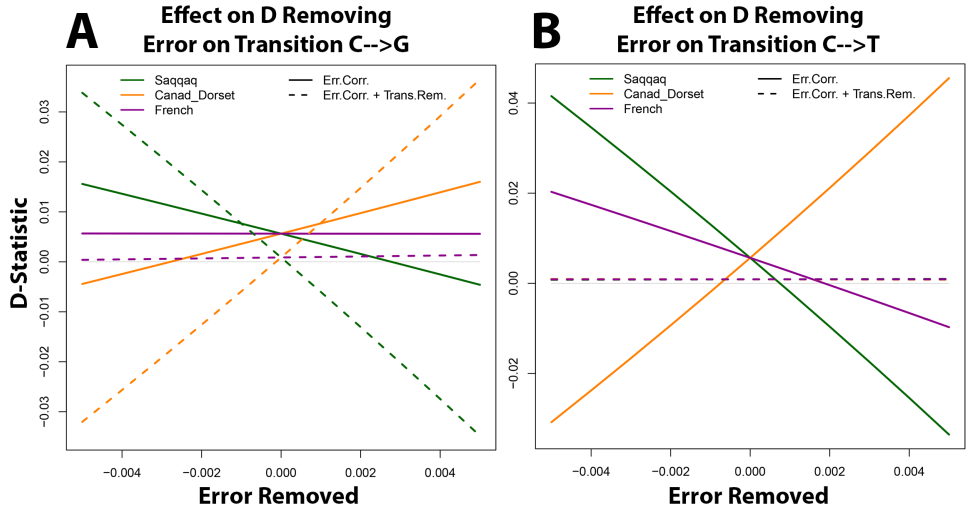


Figure 9. Behaviour of the D-Statistic in Function of the Type-Specific Error. Effect of increasing and decreasing the removal of error for the base transitions $C \to G$ and $C \to T$ for one of the Greenlandic Saqqaq, Canadian Dorset and French genomes. This corresponds to the addition of a value in 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.



bioRxiv preprint doi: https://doi.org/10.1101/127852; this version posted September 4, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

References

1. Haber M, Mezzavilla M, Xue Y, Tyler-Smith C. Ancient DNA and the rewriting of human history: be sparing with Occam's razor. Genome biology. 2016;17(1):1. doi:10.1186/s13059-015-0866-z.

2. Green RE, Krause J, Briggs AW, Maricic T, Stenzel U, Kircher M, et al. A Draft Sequence of the Neandertal Genome. Science. 2010;328(5979):710–722. doi:10.1126/science.1188021.

3. Lalueza-Fox C, Gilbert MTP. Paleogenomics of archaic hominins. Current Biology. 2011;21(24):R1002–R1009. doi:10.1016/j.cub.2011.11.021.

4. Wall JD, Yang MA, Jay F, Kim SK, Durand EY, Stevison LS, et al. Higher Levels of Neanderthal Ancestry in East Asians Than in Europeans. Genetics. 2013;doi:10.1534/genetics.112.148213.

5. Reich D, Patterson N, Kircher M, Delfin F, Nandineni M, Pugach I, et al. Denisova Admixture and the First Modern Human Dispersals into Southeast Asia and Oceania. The American Journal of Human Genetics. 2011;89(4):516–528. doi:http://dx.doi.org/10.1016/j.ajhg.2011.09.005.

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

7. Black JS, Salto-Tellez M, Mills KI, Catherwood MA. The impact of next generation sequencing technologies on haematological research - A review. Pathogenesis. 2015;2:9–16. doi:http://dx.doi.org/10.1016/j.pathog.2015.05.004.

8. Rasmussen M, Anzick S, Waters M, Skoglund P, DeGiorgio M, Stafford T, et al. The genome of a Late Pleistocene human from a Clovis burial site in western Montana. Nature. 2014;506(7487):225–229. doi:10.1038/nature13025.

9. Chatters JC. The Recovery and First Analysis of an Early Holocene Human Skeleton from Kennewick, Washington. American Antiquity. 2000;65(2):291–316.

10. Pritchard JK, Stephens M, Donnelly P. Inference of population structure using multilocus genotype data. Genetics. 2000;155:945–959.

11. Alexander DH, Novembre J, Lange K. Fast model-based estimation of ancestry in unrelated individuals. Genome Research. 2009;doi:10.1101/gr.094052.109.

12. Reich D, Thangaraj K, Patterson N, Price A, Singh L. Reconstructing Indian Population History. Nature. 2009;461(7263):489–494.

13. Raghavan M, Skoglund P, Graf KE, Metspalu M, Albrechtsen A, Moltke I, et al. Upper Palaeolithic Siberian genome reveals dual ancestry of Native Americans. Nature. 2013;505(7481):87–91. doi:10.1038/nature12736.

14. Reich D, Green RE, Kircher M, Krause J, Patterson N, Durand EY, et al. Genetic history of an archaic hominin group from Denisova Cave in Siberia. Nature. 2010;468(7327):1053–1060.

15. Patterson NJ, Moorjani P, Luo Y, Mallick S, Rohland N, Zhan Y, et al. Ancient Admixture in Human History. Genetics. 2012;doi:10.1534/genetics.112.145037.

16. Nielsen R, Paul J, Albrechtsen A, Song Y. Genotype and SNP calling from next-generation sequencing data. Nature Reviews Genetics. 2011;12(6):443–451. doi:10.1038/nrg2986.

bioRxiv preprint doi: https://doi.org/10.1101/127852; this version posted September 4, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

17. Pickrell JK, Pritchard JK. Inference of Population Splits and Mixtures from Genome-Wide Allele Frequency Data. PLoS Genet. 2012;8(11):1–17. doi:10.1371/journal.pgen.1002967.

18. Li Y, Vinckenbosch N, Tian G, Huerta-Sanchez E, Jiang T, Jiang H, et al. Resequencing of 200 human exomes identifies an excess of low-frequency non-synonymous coding variants. Nature Genetics. 2010;42(11):969–972 IF:35.209.

19. Orlando L, Ginolhac A, Zhang G, Froese D, Albrechtsen A, Stiller M, et al. Recalibrating Equus evolution using the genome sequence of an early Middle Pleistocene horse. Nature. 2013;499(7456):74–78 IF:38.597.

20. Ewing G, Hermisson J. MSMS: a coalescent simulation program including recombination, demographic structure and selection at a single locus. Bioinformatics. 2010;26(16):2064–2065. doi:10.1093/bioinformatics/btq322.

21. Busing FMTA, Meijer E, Leeden RVD. Delete-m Jackknife for Unequal M. Statistics and Computing. 1999;9(1):3–8. doi:10.1023/A:1008800423698.

22. Raghavan M, Steinrücken M, Harris K, Schiffels S, Rasmussen S, DeGiorgio M, et al. Genomic evidence for the Pleistocene and recent population history of Native Americans. Science. 2015;doi:10.1126/science.aab3884.

23. Rasmussen M, Li Y, Lindgreen S, Pedersen JS, Albrechtsen A, Moltke I, et al. Ancient human genome sequence of an extinct Palaeo-Eskimo. Nature. 2010;463(7282):757–762. doi:10.1038/nature08835.

24. Johnson O. Information Theory And The Central Limit Theorem. Imperial College Press; 2004. Available from: http://www.worldscientific.com/worldscibooks/10.1142/p341.

25. Lamperti JW. Probability: A Survey of the Mathematical Theory, Second Edition. John Wiley & Sons; 1996. Available from: http://eu.wiley.com/WileyCDA/WileyTitle/productCd-0471154075.html.

26. Slutsky E. Uber stochastische Asymptoten und Grenzwerte. Internationale statistische Zeitschrift. 1925;5(3):3–89.

27. Pesaran MH. Time Series and Panel Data Econometrics. Oxford University Press; 2015.

28. Altshuler D, Durbin R, Abecasis G, Bentley D, Chakravarti A, Clark A, et al. A map of human genome variation from population-scale sequencing. Nature. 2010;467:1061–1073. doi:10.1038/nature09534.

29. Consortium IH. The International HapMap Project. Nature. 2003;426(6968):789–796.

30. Meyer M, Kircher M, Gansauge MT, Li H, Racimo F, Mallick S, et al. A High-Coverage Genome Sequence from an Archaic Denisovan Individual. Science. 2012;338(6104):222–226. doi:10.1126/science.1224344.

31. Kent WJ, Sugnet CW, Furey TS, Roskin KM, Pringle TH, Zahler AM, et al. The human genome browser at UCSC. Genome Res. 2002;12(6):996–1006.

32. Lalueza-Fox C, Gilbert MT. Paleogenomics of Archaic Hominins. Current Biology. 2011;21(24):R1002 – R1009. doi:http://dx.doi.org/10.1016/j.cub.2011.11.021.

bioRxiv preprint doi: https://doi.org/10.1101/127852; this version posted September 4, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

33. Skoglund P, Mallick S, Bortolini MC, et al. Genetic evidence for two founding populations of the Americas Nature. 2015;525:104–108. doi:10.1038/nature14895

34. Durand E, Patterson N, Reich D, Slatkin M Testing for ancient admixture between closely related populations Molecular Biology and Evolution. 2011;28:2239–2252. doi:10.1093/molbev/msr048

35. Raghavan M, DeGiorgio M, Albrechtsen A, Moltke I, Skoglund P, SK Thorfinn, et al. The Genetic Prehistory of the New World Arctic Science. 2014;345. doi:10.1126/science.1255832