Estimating admixture pedigrees of recent hybrids without a contiguous reference genome

Garcia-Erill, Genís; Hanghøj, Kristian; Heller, Rasmus; Wiuf, Carsten; Albrechtsen, Anders

Published in:
Molecular Ecology Resources

DOI:
10.1111/1755-0998.13830

Publication date:
2023

Document version
Publisher's PDF, also known as Version of record

Document license:
CC BY-NG

Citation for published version (APA):
Estimating admixture pedigrees of recent hybrids without a contiguous reference genome

Genís Garcia-Erill1 | Kristian Hanghøj1 | Rasmus Heller1 | Carsten Wiuf2 | Anders Albrechtsen1

1Department of Biology, University of Copenhagen, Copenhagen, Denmark
2Department of Mathematical Sciences, University of Copenhagen, Copenhagen, Denmark

Correspondence
Genís Garcia-Erill, Department of Biology, University of Copenhagen, Copenhagen, Denmark.
Email: genis.grcerll@gmail.com

Funding information
Danmarks Frie Forskningsfond, Grant/Award Number: 8021-00360B and 8049-00098B; H2020 European Research Council, Grant/Award Number: 853442

Handling Editor: Nick Hamilton Barton

Abstract
The genome of recently admixed individuals or hybrids has characteristic genetic patterns that can be used to learn about their recent admixture history. One of these are patterns of interancestry heterozygosity, which can be inferred from SNP data from either called genotypes or genotype likelihoods, without the need for information on genomic location. This makes them applicable to a wide range of data that are often used in evolutionary and conservation genomic studies, such as low-depth sequencing mapped to scaffolds and reduced representation sequencing. Here we implement maximum likelihood estimation of interancestry heterozygosity patterns using two complementary models. We furthermore develop apoh (Admixture Pedigrees of Hybrids), a software that uses estimates of paired ancestry proportions to detect recently admixed individuals or hybrids, and to suggest possible admixture pedigrees. It furthermore calculates several hybrid indices that make it easier to identify and rank possible admixture pedigrees that could give rise to the estimated patterns. We implemented apoh both as a command line tool and as a Graphical User Interface that allows the user to automatically and interactively explore, rank and visualize compatible recent admixture pedigrees, and calculate the different summary indices. We validate the performance of the method using admixed family trios from the 1000 Genomes Project. In addition, we show its applicability on identifying recent hybrids from RAD-seq data of Grant’s gazelle (Nanger granti and Nanger petersii) and whole genome low-depth data of waterbuck (Kobus ellipsiprymnus) which shows complex admixture of up to four populations.

KEYWORDS
Hybridization, non-model organisms, admixture, genomics, RADseq, lowdepth sequencing

1 | INTRODUCTION

Hybridization and admixture is a common outcome when different populations or species come into secondary contact, with an increasingly recognized role in evolution, for example during speciation (Feder et al., 2012) or adaptation (Edelman & Mallet, 2021). An important step towards understanding the role of hybridization in natural populations is the ability to reliably detect and characterize the presence of recent hybrids or admixed individuals from genetic data. Hybrid inference based on hybrid indexes or admixture proportions,
which model the global proportion of the individual's genome originating from each admixing population, has a long history in population genetics (Beugin et al., 2018; Buerkle, 2005; Pritchard et al., 2000; Vaha & Primmer, 2006). However, it has been shown that these methods can lead to spurious conclusions regarding admixture (Lawson et al., 2018), and furthermore they do not directly distinguish between very recent hybridization, happening within up to five generations, and admixed individuals where hybridization happened several generations ago. Other methods have harnessed the characteristic signals that admixture leaves in the genomes in the first few generations following the admixture event. One such characteristic pattern of recent hybrids is an excess of regions in the genome where the two alleles at each locus derive from different ancestries. This pattern is known as ‘interancestry heterozygosity’ and is used by multiple existing methods to detect and characterize recent hybrids (Anderson & Thompson, 2002; Boecklen & Howard, 1997; Crouch & Weale, 2012; Fitzpatrick, 2012; Pfaffelhuber et al., 2022; Shastry et al., 2021). Another useful signal to study recent hybridization is the length of continuous stretches of the genome with the same ancestry, known as ‘admixture tracts’. Admixture tract lengths decrease each generation after the admixture event through recombination, and this is used by several methods (Chakraborty & Rannala, 2023; Pei et al., 2020; Zou et al., 2015) to learn about recent admixture history. Finally, a combination of both interancestry heterozygosity and admixture tract lengths has been shown to be most powerful, allowing inference of admixture pedigree and time since admixture (Avadhanam & Williams, 2022; Frandsen et al., 2020).

Studies in the field of evolutionary biology and ecological or conservation genetics of wild populations often focus on understudied organisms, for which there is rarely a chromosome-level reference genome and much less a known recombination map available. Even if a good reference genome is available, resequencing projects often need to use cost-effective sequencing approaches, like restriction site-associated DNA sequencing (RADseq) (Andrews et al., 2016), or low coverage whole genome sequencing (WGS) (Lou et al., 2021), to maximize sample sizes at the expense of genome-wide coverage. While allowing larger sample sizes at an affordable cost, these sequencing techniques reduce the amount of information that can be extracted from the sequencing data. This often prevents the application of state-of-the-art methods developed for human or other model organisms. For example, it is usually not possible to model the positional auto-correlation of ancestry along the genome when only a subset of the genome is covered by sequencing reads, as in reduced representation sequencing, or when genotypes cannot be reliably called, as with low-depth sequencing data. The same limitations apply in the absence of a highly contiguous reference genome where the sequencing data can be mapped to. In these situations, only the patterns of interancestry heterozygosity can be used to model the recent admixture history (Gompert et al., 2017).

Although there are several methods available to detect recent admixture or hybridization based on inferring genome-wide proportions of interancestry heterozygosities (Anderson & Thompson, 2002; Crouch & Weale, 2012; Fitzpatrick, 2012; Pfaffelhuber et al., 2022; Shastry et al., 2021; Wringe et al., 2017a, 2017b), all of them have at least one of two limitations. Most of these methods, with the exception of ENTROPY (Shastry et al., 2021), do not have an efficient implementation that can be applied to genome-scale data sets with large sample sizes. Moreover, only ENTROPY can be run from genotype likelihoods, allowing for analyses of low-depth sequencing data. Additionally, most methods do not explicitly infer the underlying recent admixture pedigree and provide it as an output to the user. Only the Bayesian method NEWHYBRIDS (Anderson & Thompson, 2002) and associated packages such as hybriddetection (Wringe et al., 2017b) classify the samples in different hybrid classes (F1, F2, backcrosses, etc.) and assign posterior probabilities for each class, hence quantifying the confidence in the classification. However, these methods require as input a pre-specified set of hybrid classes that can be detected, defined each by their expected values of interancestry heterozygosities. This will necessarily result in misclassification if the recent admixture history of a target sample is not included in the set of potential hybrid classes. Furthermore, it is limited to admixture between two parental populations \((K=2)\). Admixture in nature is often complex and can potentially involve more than two populations. Alternative approaches are based on directly interpreting the estimates of interancestry heterozygosity patterns relative to their expectation for different hybrid classes (Gompert et al., 2014; Pfaffelhuber et al., 2022; Shastry et al., 2021; Waples et al., 2021). These can potentially be applied to situations with multi-population admixture. However, none of these methods output straightforward evaluations of recent admixture histories compatible with the estimates, or compare between the relative fit of different admixture histories. Therefore, there is a considerable unexploited potential in methods that make more explicit use of the information contained in interancestry heterozygosity for inferring concrete recent admixture histories in a quantitative and easily interpretable framework.

To address this situation we implement two similar, but complementary models that infer patterns of interancestry heterozygosity, and use the estimates to explore and evaluate concrete recent admixture histories. The ‘paired ancestries model’ (Gompert et al., 2014; Nohr et al., 2021) models the global proportions of homozygous and heterozygous ancestries, without distinguishing between the two possible permutations of heterozygous ancestry (unordered paired ancestries). In contrast, the ‘parental admixture model’ (Crouch & Weale, 2012; Pfaffelhuber et al., 2022) allows us to estimate ordered paired ancestries that can distinguish between the two permutations, which increases the resolution to detect cases of recent admixture based on the paired ancestries (Figure S1). Both models have been previously described in studies addressing recent admixture and hybridization (Crouch & Weale, 2012; Gompert et al., 2014; Pfaffelhuber et al., 2022; Shastry et al., 2021), but its potential to directly infer recent admixture pedigrees has not yet been explored. Moreover, our implementation can be run from called genotypes or genotype likelihood data in reasonable times for large datasets, and is furthermore the first implementation of the parental admixture model for genotype likelihood data. We have implemented both methods as extensions to the software NGSremix (Nohr et al., 2021).
Our method can be applied on genomic data with reduced genome contiguity or low-depth sequencing (Figure 1a), and is therefore suitable for non-model organism analyses. Furthermore, we develop the independent software *apoh* (Admixture Pedigrees of Hybrids) to automate downstream interpretation of the estimated proportions in the context of recent admixture. We implement *apoh* both as a command-line tool and as a Graphical User Interface (GUI) that facilitates their use, and the interpretation of results, in applied cases.
2 | RESULTS

We present a method to evaluate a series of probable admixture pedigrees consistent with the estimated patterns of interancestry heterozygosities for an admixed focal individual. The method works on a range of different data types, and notably both for low-depth and reduced representation sequencing data, including situations where such data are mapped to a draft genome assembly rather than a high-quality chromosome-level reference genome (Figure 1a). Note that here we define ‘recent admixture’ as the existence within the last five generations of two unadmixed ancestors of the focal individual with different ancestries, and we denote their interbreeding as the admixture event that we try to infer.

2.1 | Models of paired ancestries

The inference of recent admixture in the methods presented here is based on modelling the patterns of interancestry heterozygosities, using the genome-wide proportions of all possible ancestry configurations across a large number of loci. Assuming a diploid genome, an admixed individual carries at each locus two alleles that can have either the same or different population ancestries. The possible ancestry pair configurations are then equal to all possible pairwise combinations of $K$ ancestries. We refer to the vector $A = (a_1, a_2)$, which specifies the population of origin of a pair of alleles at a locus, as the ‘paired ancestry’. The paired ancestries can be either ordered or unordered. The ordered version distinguishes the ancestry of the maternal and paternal allele, while in the unordered version the parental origin is ignored (Figure 1b). We refer to the global genome-wide proportions of each paired ancestry as the ‘paired ancestry proportions’, which give the probability of observing a certain paired ancestry at a certain locus, $P(A = (a_i, a_j))$. The same or very similar quantities to the unordered paired ancestry proportions have been used before in the study of recent admixture, with different names such as ‘genomic proportions’ ( Fitzpatrick, 2012 ), ‘admixture class matrix’ ( Gompert et al., 2014 ), ‘admixture complement matrix’ ( Shastry et al., 2021 ) and ‘ternary ancestry fractions’ ( Waples et al., 2021 ).

We use three different ancestry models to infer the global paired ancestry proportions. We give here a brief description of each model, the information they contain and their assumptions, while a more formal description can be found in the Materials and Methods section.

**Standard admixture model** (used for the ‘independent pedigree’, see below) is the model of ancestry from the widely used software structure ( Pritchard et al., 2000 ) and ADMIXTURE ( Alexander et al., 2009 ). The ancestry of the pair of alleles at a locus is assumed to be independent given the individual’s admixture proportions $Q$, which is not the case in cases of recent hybridization. The estimated paired ancestries under this model, therefore, are consistent with the case where admixture and hybridization happened several generations ago such that $P(A = (a_1, a_2)) \approx P(a_1)P(a_2)$.

**Paired ancestries model** explicitly infers the probability of each paired ancestry combination. It does not distinguish between different permutations of ancestry combinations, so $P(A = (a_1, a_2)) = P(A = (a_2, a_1))$. It therefore infers what we call ‘unordered paired ancestries’ (Figure 1b). It can model both an excess of heterozygous ancestries and an excess of homozygous ancestries, relative to the expected under the standard admixture model. This means it can account for cases of recent admixture, that will lead to an excess of heterozygous ancestries, and also for inbreeding, that leads to an excess of homozygous ancestry.

**Parental admixture model** infers the admixture proportions of the two parents of the individual. The paired ancestry proportions are then given by the product of the parental admixture proportions. If the parents have different admixture proportions, the model will give different probability to each permutation of paired ancestry, $P(A = (a_1, a_2)) \neq P(A = (a_2, a_1))$, so it infers ‘ordered paired ancestries’ (Figure 1b). It can account for cases of recent admixture, where the parents will have different admixture proportions and therefore there will be an excess of heterozygous ancestries. However, it cannot account for inbreeding, since it does not allow for excess of homozygous ancestry.

We implemented the ’paired ancestry model’ and the ’parental admixture model’ as part of the NGSremix software ( Nohr et al., 2021 ). They can be used to estimate paired ancestry proportions and provide information on the recent admixture process (Figure 1). The paired ancestry model ( Gompert et al., 2014 ; Nohr et al., 2021 ) models the global proportions of homozygous and heterozygous ancestries, without distinguishing between the two possible permutations of heterozygous ancestry (unordered paired ancestries). In contrast, the parental admixture model ( Crouch & Weale, 2012 ; Pfaffelhuber et al., 2022 ) allows to estimate ordered paired ancestries that can distinguish between the two permutations, which increases the resolution to detect cases of recent admixture based on the paired ancestries (Figure 1b). The increase in resolution allows for improved identification of admixture events more than four generations ago, and distinguishes between different pedigrees. For this reason, we base most of the downstream inference on the ordered paired ancestries estimated with the parental admixture model. We did not implement the standard admixture model, since its parameters can be calculated as a function of the estimates from either of these two models, or inferred with established software such as ADMIXTURE for genotype data ( Alexander et al., 2009 ) or NGSadmix for genotype likelihoods ( Skotte et al., 2013 ), among others.

2.2 | Inference of recent admixture pedigrees

We developed the software Apollo to automatically explore the most compatible admixture pedigrees and assess if there is evidence for the sample being very recently admixed. From the
ordered paired ancestry proportions estimated with the parental admixture model of an admixed individual, we explore the pedigrees that are most compatible with it assuming that either all ancestors where unadmixed four generations ago, or assuming an independent model where the ancestries of the two alleles are independent. Due to the inherent stochasticity in the parent-to-offspring transmission of haplotypes, we cannot always unambiguously identify a single admixture pedigree from the paired ancestry proportions, even if we had perfect estimates. Moreover, we want the method to work in situations where we have no information on genome location. In this situation, we cannot model the local correlation of ancestry such as local ancestry tracts, which makes us unable to account for the statistical uncertainty associated with the inference. To reflect this and other potential sources of uncertainty, we choose instead to exhaustively explore and compare multiple similar pedigrees that are most compatible with the estimated paired ancestry proportions.

Finally, as a baseline to assess the support for recent admixture, we use a pedigree where all ancestors four generations ago have the same admixture proportions. In this case, the paired ancestry proportions expected under the standard admixture model should constitute a good model fit. Because of the assumption of independence in this model, we refer to this pedigree as the ‘independent pedigree’. We note, however, that this type of pedigree includes situations where both parents are in fact recently admixed, but have equal admixture proportions (e.g. Figure S1, F2 admixture). It is therefore important to keep in mind that in our framework, an inferred lack of support for recent admixture does not necessarily rule out that recent admixture did in fact take place.

2.3 Summary indices

As a guide to help in interpreting the model estimates and their fit to inferred recent admixture pedigrees, we developed several summary indices. The main index is designed to assess the compatibility of a potential admixture pedigree to the estimated paired ancestry proportions, using the distance between the estimate and the paired ancestry proportions expected under that pedigree. We measure the distance as the Jensen–Shannon distance (JSD), which is a symmetrical measure of divergence between two probability distributions, to measure the distance between paired ancestry proportions. We use the ordered paired ancestries, since these have more information to distinguish between different admixture pedigrees (Figure S1). We then use the JSD to rank potential admixture pedigrees; this includes recent admixture pedigree but also the independent pedigree, which accounts for the case where there is not evidence of recent admixture. The distance is a useful but ad hoc metric to rank admixture pedigrees, and should not be taken as a formal statistical test for the support to a pedigree. We complement it with a block bootstrap approach to estimate uncertainty in the inference of recent admixture pedigrees. For each bootstrap replicate, we estimate paired ancestry proportions, and find the most compatible pedigree for each replicate. The bootstrap support for a given pedigree is then the proportions of bootstrap replicates in which that pedigree has the most support. However, we stress that in situations where we do not have information on genome location this will not account for all the uncertainty.

We furthermore developed two additional summary indices that convey information contained in the estimates of paired ancestry proportions:

- The first one is the ‘inconsistency index’, which measures the distance between the unordered paired ancestries estimated with the paired ancestries model and the ones estimated with the parental admixture model. Since the two estimates are expected to be the same under a good model fit, a non-zero inconsistency index value suggests the estimates should be treated with caution. Non-zero distance can be due to inbreeding or to a bad model fit of the population allele frequencies used to model the sample ancestry.

- The second index is the ‘admixture index’, which, if the individual is in fact recently admixed, reflects the time since admixture. In the simplest case where the individual is the result of a single admixture event between two populations, the value of the index corresponds to the number of generations since admixture. In more complex cases, involving multiple admixture events and/or more than two ancestries, the index has a less straightforward interpretation, and is a lower bound on the sum of admixture times in the pedigree.

2.4 Implementation

We have implemented the parental admixture proportions model as an extension to the C++ software NGSremix (Nohr et al., 2021), where we previously implemented the paired ancestry model for estimation of relatedness. It can be run either from genotype data or from genotype likelihood data. We implement the downstream inference of recent admixture pedigrees, visualization and derived summary indexes as an the software apoh, and developed a GUI using the ‘shiny’ R package (Chang et al., 2021). The GUI allows users to upload estimates of parental admixture proportions and paired ancestry proportions from NGSremix, and perform the downstream inference. All methods are publicly available in github, NGSremix in https://github.com/KHanghoj/NGSremix and apoh in https://github.com/popgenDK/apoh.

2.5 Simulations

We first evaluated the performance of the parental admixture model under simple simulated scenarios of 1000–1000,000 unlinked markers. We used the true genotypes to emulate the potential usage of SNP chips and high-depth sequencing including RADSeq. We also simulated reads based on a Poisson distribution assuming 1X to
8X average depth of sequencing to emulate the use of low-depth sequencing data. We used the simple GATK genotype (McKenna et al., 2010) likelihoods model for the genotype likelihood simulations assuming a sequencing error rate of 0.1%.

For each data type, we simulated recently admixed individuals from two populations with varying number of generations since admixture and varying the amount of genetic differentiation between the two admixing populations. Ancestral allele frequencies were estimated from 20 unadmixed individuals from each ancestry together with 10 admixed individuals. For genotype call, this was done using ADMIXTURE and for low depth this was done using NGsadmix. A comparison of the true known minor parental ancestry with those estimated for the different admixture classes revealed that the inference is less accurate when the ancestral source differentiation is very low, and hence cannot distinguish between admixture classes when $F_{ST}$ is 0.01. However, at a still relatively minor $F_{ST}$ of 0.05, we are usually able to distinguish between different classes (Figure S2).

As expected, the accuracy decreases as we decrease the number of sites and/or the depth of sequencing. When having a large number of sites like 1 million, the inference still works reasonably well for the lowest tested depth of 1X.

### 2.6 Evaluation using admixed family trios

To assess our implementation of the parental admixture proportions inference, we used genotype data of family trios from the 1000 Genomes project (Byrska-Bishop et al., 2022). Because the parents of the offspring were included in the unrelated samples for which we ran ADMIXTURE, we could compare direct estimates of the parent’s admixture proportions estimated with ADMIXTURE, and the estimates of these same proportions using only their offspring’s genotypes inferred by applying the parental admixture model. This revealed a high accuracy of the parental admixture model estimates (Figure 2a). To illustrate that the method would also work on SNP chip data, RADseq data and low-depth sequencing data, we repeated the estimation with subsampled sites and simulated low-depth sequencing reads (Figure S3). In general, the results are similar but at 1X depth the estimates become visibly more noisy.

There are multiple cases where the two parents of an individual have different admixture proportions, which corresponds to a recent and ongoing admixture in these populations, as expected. In the figure, we show such an example with individuals HG01969 from the PEL population. The suggested pedigree from apoh suggests that one parent is admixed while the other parent has three European grandparents and five Native Americans. The estimated parental ancestry and the ancestry estimated directly from the parents are consistent with the suggested pedigree, in contrast to the independent pedigree which assumes the admixture is more ancient.

Furthermore, we also analysed the offspring in trios from a South Asian population, PJL, to explore the performance of the method in cases of spurious admixture due to the source population not being well represented in the estimation of the population allele frequencies. We estimated paired ancestries with the parental and paired ancestry models for the PJL samples, using the population frequencies previously estimated with ADMIXTURE without including the PJL population. The parental admixture proportions modelled all PJL individuals as being admixed between all three ancestral populations that represent African, Native American and European ancestries (Figure S4). Therefore, these allele frequencies are not appropriate for modelling the South Asian ancestry of the PJL samples. Individuals from the PJL population show a relatively elevated inconsistency index compared to all other populations, demonstrating that the inconsistency index can detect violations of the model assumptions. An individual with sample ID HG01279 from the CLM population, for which the population allele frequencies are appropriate, shows a relatively elevated inconsistency index (Figure S4).

However, this individual has unusually high inbreeding relative to the other samples (Figure S4), demonstrating the other situation in which the inconsistency index value reflects departures from the model assumptions. Furthermore, five individuals from PJL with outlying inconsistency indices also showed high inbreeding coefficients, revealing that the inconsistency index reflects the cumulative effects of poorly estimated ancestral allele frequencies and inbreeding. We visualized the estimated unordered paired ancestry proportions for the HG01279 individual and for HG03707 from PJL, which has zero inbreeding and for which the high inconsistency index is therefore exclusively driven by poorly modelled allele frequencies. The plot illustrates that both of the model violations lead to an increase in homozygous paired ancestry proportions relative to the independent pedigree and the parental admixture model estimates (Figure S4).

### 2.7 Real data applications

To explore the applicability of the methods in studies with non-model organisms presenting more challenging data, we used datasets from two African antelope species. Both of them included populations with recent gene flow between different species or sub-species. One dataset is low-depth WGS data (waterbuck), and the other is reduced representation sequencing (Grant’s gazelle). Therefore, these two datasets showcase the utility of our method for data typically used for non-model organisms where haplotype phasing or even genotype calling is not possible and where the position of SNPs on the chromosome is not known.

#### 2.7.1 Low-depth WGS: Waterbuck

We used low-depth (3X) WGS data for 38 waterbuck (Kobus ellipsiprymnus) samples (Wang et al., 2022) from four populations in East Africa. Two of the populations, Kidepo Valley National Park (KVNP, Uganda) and Maswa (Tanzania), belong to the defassa waterbuck subspecies (K. e. defassa). The other two populations, Samburu and...
Nairobi (Kenya), are in the range of the common waterbuck subspecies (*K. e. ellipsiprymnus*), but are known to have received recent gene flow from the defassa waterbuck (Figure 3a) (Lorenzen et al., 2006; Wang et al., 2022). Eight samples, six from Samburu and two from Nairobi, were modelled as having ancestry from more than two populations, all but one involving admixture between the two subspecies (Figure 3b). We ran our implementation of the paired ancestry and parental admixture models in these 8 samples, using the population allele frequencies estimated with NGSadmix at $K=4$, and estimated ordered and unordered paired ancestry proportions for all samples (Figure S5). All of the samples had parents with different admixture proportions, which is the main signal of recent admixture (Figure 3c). We then applied apoh to infer the most compatible pedigrees and calculate the recent admixture indexes (Table 1).

For all admixed samples from Samburu, the distance to the most compatible recent admixture pedigree showed a clear improvement with respect to the independent pedigree and most have high bootstrap support for recent admixture (see one example in Figure 3d and samples 398, 392, 408, 415, 393 and 414 in Table 1) and in...
Figure S5). This, together with a low inconsistency index, indicates there is strong evidence that the individuals are recently admixed and that the favoured admixture pedigree are a good approximation. In some cases, we cannot distinguish between alternative recent admixture pedigrees, which is identified with a low bootstrap support (Figure 3d, Table 1). In all cases, the most compatible recent admixture pedigrees show that admixture happened three to four generations ago, and all the samples are double or triple backcrosses of a defassa/common waterbuck hybrid (Figure S6). Sample 427 and 429, from Nairobi, show a more complicated admixture history, with ancestry from four and three different populations, respectively (Figure S6). However, these two samples also have a considerably more elevated inconsistency index, which means the paired ancestry estimates between the two models do not coincide (Table 1). This suggests the individuals are probably recently admixed from a defassa waterbuck population that is not well represented in the dataset, and therefore the allele frequencies are not accurate and prevent a reliable evaluation of its admixture pedigree.

2.7.2 | RADseq: Grant's gazelle

We also assessed the performance of the methods on reduced representation sequencing data by applying them to a RADseq data set (Garcia-Erill et al., 2021). The Grant’s gazelle species complex (Nanger sp.) comprises three species with a non-overlapping distribution restricted to East Africa. Previous analyses of the dataset revealed two contact zones between divergent lineages, one in the Maasai Mara (Kenya) between two N. granti subspecies, N. g. granti and N. g. robertsii, and another in Mkomazi (Tanzania), between the
N. granti and N. petersii species. Both populations show signs of ancient and recent admixture, suggesting lineages in this species complex have diverged with possibly recurrent periods of isolation and migration, and are currently in a time of lineage connectivity (Garcia-Erill et al., 2021). In addition to the admixed populations, we included samples from Maswa (Tanzania), which belongs to the N. g. robertsii subspecies of N. granti, Monduli (Tanzania) that belongs to N. g. granti subspecies, and Tsavo (Kenya), which belongs to the N. petersii species (Figure 4a).

We identified six samples with ancestry from multiple populations: two from Maasai Mara and four from Mkomazi (Figure 4b). We used the inferred population allele frequencies to estimate ordered and unordered paired ancestries with the two paired ancestry models for all six admixed samples (Figure S5). The results identified two samples, one from each population, with signs of being recently admixed (Figure 4c). Sample 283 from Maasai Mara shows recent ancestry from both N. granti subspecies. Two potential recent admixture pedigrees have nearly identical distance to the estimated ordered paired ancestry proportions, and show relatively similar bootstrap supports. The first pedigree includes an ancestor from each of the two minor ancestries four generations ago, while the second only includes an ancestor from N.g.robertsii and disregards the lowest N.g.granti ancestry component (Figure 4d,e). This suggests that, in this case, there is not enough information in the estimates to distinguish between two fundamentally different admixture histories. A potential explanation is that the recently admixed population source is not well represented in the data, and the minor ancestry from N.g.granti could therefore be an artefact. The inconsistency index does not detect it (Table 1, Figure S7), but it could be due to very low ancestry proportion from the potentially spurious source, N.g.granti. The other recently admixed sample is 3720 from Mkomazi, a historically admixed population in the contact zone between the N.petersii and N.granti species. This individual is modelled as a recent backcross admixed from an N.g.granti population. The JSD to the estimates show strong improvement of the most compatible pedigree with respect to both the independent pedigree and the second recent admixture pedigree, which shows the individual is likely a triple or quadruple backcross of a N.g.granti population (Figure S8), with its admixture index ~5 (Table 1) favouring a quadruple backcross.

The remaining samples did not show evidence of recent admixture. Sample 286 from Maasai Mara and sample 3719 from Mkomazi had an elevated inconsistency index (Table 1). This is likely due to inbreeding, since other individuals with similar ancestry profiles from the same populations show no increased inconsistency index (Figure S7), suggesting the ancestry sources are good approximations. In all cases, the bootstrap support for the most compatible pedigree is low relative to what we found with the waterbuck data set (Table 1). This is consistent with the inference being based on a much lower number of sites.

**Table 1** Summary indices for the two application to data sets of low-depth sequencing of waterbuck and RADseq of Grant’s gazelle.

<table>
<thead>
<tr>
<th>Species</th>
<th>SampleID</th>
<th>Inconsistency index</th>
<th>Distance to independent pedigree (bootstrap support)</th>
<th>Distance to pedigree 1 (bootstrap support)</th>
<th>Distance to pedigree 2 (bootstrap support)</th>
<th>Admixture index</th>
</tr>
</thead>
<tbody>
<tr>
<td>Waterbuck</td>
<td>398</td>
<td>0.01</td>
<td>0.30 (0%)</td>
<td>0.14 (60%)</td>
<td>0.14 (40%)</td>
<td>3.08</td>
</tr>
<tr>
<td>Waterbuck</td>
<td>392</td>
<td>0.01</td>
<td>0.29 (2%)</td>
<td>0.03 (94%)</td>
<td>0.12 (0%)</td>
<td>4.16</td>
</tr>
<tr>
<td>Waterbuck</td>
<td>408</td>
<td>0.01</td>
<td>0.26 (0%)</td>
<td>0.14 (46%)</td>
<td>0.15 (28%)</td>
<td>3.48</td>
</tr>
<tr>
<td>Waterbuck</td>
<td>415</td>
<td>0.02</td>
<td>0.24 (0%)</td>
<td>0.11 (99%)</td>
<td>0.12 (1%)</td>
<td>3.54</td>
</tr>
<tr>
<td>Waterbuck</td>
<td>393</td>
<td>0.00</td>
<td>0.24 (0%)</td>
<td>0.08 (100%)</td>
<td>0.10 (0%)</td>
<td>3.56</td>
</tr>
<tr>
<td>Waterbuck</td>
<td>414</td>
<td>0.00</td>
<td>0.19 (0%)</td>
<td>0.01 (100%)</td>
<td>0.25 (0%)</td>
<td>4.11</td>
</tr>
<tr>
<td>Waterbuck</td>
<td>427</td>
<td>0.09</td>
<td>0.49 (0%)</td>
<td>0.15 (100%)</td>
<td>0.18 (0%)</td>
<td>4.74</td>
</tr>
<tr>
<td>Waterbuck</td>
<td>429</td>
<td>0.05</td>
<td>0.41 (0%)</td>
<td>0.06 (82%)</td>
<td>0.08 (83%)</td>
<td>5</td>
</tr>
<tr>
<td>Grant’s gazelle</td>
<td>286</td>
<td>0.41</td>
<td>0.00 (95%)</td>
<td>0.06 (0%)</td>
<td>0.10 (5%)</td>
<td>NA</td>
</tr>
<tr>
<td>Grant’s gazelle</td>
<td>283</td>
<td>0.00</td>
<td>0.26 (1%)</td>
<td>0.15 (49%)</td>
<td>0.15 (29%)</td>
<td>3.54</td>
</tr>
<tr>
<td>Grant’s gazelle</td>
<td>3713</td>
<td>0.01</td>
<td>0.00 (64%)</td>
<td>0.10 (0%)</td>
<td>0.10 (36%)</td>
<td>NA</td>
</tr>
<tr>
<td>Grant’s gazelle</td>
<td>3719</td>
<td>0.19</td>
<td>0.00 (65%)</td>
<td>0.03 (21%)</td>
<td>0.24 (14%)</td>
<td>NA</td>
</tr>
<tr>
<td>Grant’s gazelle</td>
<td>3717</td>
<td>0.05</td>
<td>0.01 (40%)</td>
<td>0.11 (0%)</td>
<td>0.20 (60%)</td>
<td>NA</td>
</tr>
<tr>
<td>Grant’s gazelle</td>
<td>3720</td>
<td>0.00</td>
<td>0.15 (37%)</td>
<td>0.08 (63%)</td>
<td>0.19 (0%)</td>
<td>4.82</td>
</tr>
</tbody>
</table>

Note: The inconsistency index indicates how well the ordered paired ancestries for which the pedigree inference is based fit the data. The distance to the different pedigrees is the Jensen–Shannon distance (JSD) between the expected paired ancestries under the proposed pedigree and the paired ancestries estimated by the model. The independent pedigree represents the case of not very recent admixture, while pedigrees 1 and pedigree 2 represent the first and second most compatible recent admixture pedigrees, respectively. Bootstrap support is shown for the three pedigrees, without necessarily summing to 100% since many other pedigrees are evaluated. The best fitting pedigree (with the minimum distance) is highlighted in bold for each sample. For those samples that a recent admixture pedigree fits the data better than the independent pedigree, the admixture index is shown, which indicates the time since admixture if there is a single admixture event (see Methods and Results for details).
3 | DISCUSSION

We have developed methods and summary indices that use the information contained in the paired ancestry proportions to learn whether there is evidence for recent admixture, and to characterize the recent admixture pedigree in case there is. We furthermore implemented two previously described models (Crouch & Weale, 2012; Nohr et al., 2021; Pfaffelhuber et al., 2022; Shastry et al., 2021) to estimate paired ancestry proportions. The use of two models with complementary information allows better utilization of the information contained in the data. Our methods and implementation have minimal requirements in terms of the genomic data used for the analyses, making them applicable to situations with reduced resource availability, such as low-depth sequencing data, analyses based on draft reference genome assemblies or reduced representation sequencing of non-model organisms. At the same time, we present an efficient implementation of the models that allows analysis of genome scale datasets with many samples. We evaluate the parental admixture model using simulated data and data with genotype information from trios. In this way, we are able to compare the inferred parental ancestry from the offspring to that estimated directly from the parents. We also show that our index for consistency was able to indicate violations of the model such as inbreeding and parental populations not represented in the data.

To illustrate potential usages of our methods, we apply it to some very challenging data based on low-depth WGS mapped to a draft assembly, and RADseq data where individuals showed signs of being admixed between two to four different populations. These case studies highlight how the wide applicability of the method can provide insights into biological questions. In this case, we combined...
analyses of two datasets of bovid species from the same geographical area in East Africa. In both cases, our analyses provide the strongest evidence yet presented for recent gene flow between historically isolated lineages in these two species complexes, which has been suggested before (Arctander et al., 1996; García-Erill et al., 2021; Lorenzen et al., 2006; Lorenzen et al., 2008), but never formally tested. Both cases include at least one case of confirmed recent gene flow through a geographical barrier, the Rift Valley, that has historically separated the *N. g. granti* and *N. g. robertsi* subspecies in the Grant’s gazelle, and the *K. e. ellipsiprymnus* and *K. e. defassa* subspecies in the waterbuck. For these two species cases, we are able to identify plausible admixture pedigrees, which provides details regarding the first few generations of hybridization between the different lineages in each case. Such detailed insight into the recent admixture history involving two full or incipient species is extremely valuable both for answering fundamental evolutionary biology questions, for example, related to the speciation process, but also for practical conservation of the involved taxa. With such information in hand, we have tangible evidence concerning the recent migration of individuals in the particular geographical setting, as well as the patterns of backcrossing taking place the first few generations after the initial hybridization event. Even in the cases where our method did not allow for a detailed admixture pedigree inference, it still provides clues regarding which of the modeling assumptions are violated, thereby providing insights into the admixture history. For example, in both the Grant’s gazelles, we identified several admixed individuals where the admixture probably took place >4 generations ago, and in the waterbuck we identified two individuals where the estimated admixture proportions are biased due to a lack of representation of their ancestry sources among the data.

Often there are several possible recent admixture pedigrees that can lead to similar paired ancestries as inferred from the data. To deal with this uncertainty, our method evaluates several plausible pedigrees and prioritizes them based on proposed ad-hoc indexes and JSD distance between the inferred paired ancestry and those of the candidate pedigrees. We do not rely on formal statistical tests of the relative fit between pedigrees, as done in other implementations of the models (Pflaffelhuber et al., 2022), because we specifically aimed our method to be applicable for genome-scale data sets without a chromosome-level reference genome. In this situation, sites are not independent because of linkage disequilibrium (LD) and the absence of reliable genome locations makes it impossible to fully account for the correlation introduced by long admixture tracts in cases of recent admixture. However, we allow the users to perform block bootstrap so that if genomic information is available, bootstrap support for suggested pedigrees can be calculated. It should be said that without chromosomal-level genomic position information it will not be possible to fully account for the biological uncertainty caused by a possibly low number of recombination events since the admixture happened. However, even medium length scaffolds might be long enough to account for the LD, but how long they need to be is hard to establish a priori since the extent of LD depends on the demographic history of the studied populations and the time since admixture.

It has been shown that the accuracy and resolution of recent admixture inference is increased when admixture tracts can be reliably identified, compared to methods based solely on patterns of interancestry heterozygosity, such as ours (Avadhnam & Williams, 2022). Therefore, our method is not intended to supplant or compete with methods based on inferred admixture tract lengths, but rather to make recent admixture history accessible for cases where the reference genome contiguity or the sequencing data precludes the application of such methods. Consequently, our method focuses on developing indices that can aid in identifying, ranking and prioritizing the most compatible recent admixture history, without explicit statistical testing of relative fit.

In conclusion, we present a simple and user-friendly set of methods to estimate patterns of paired ancestry proportions from genetic data. These methods facilitate the identification of recently admixed individuals and characterization of their admixture history, which is done through evaluation of recent admixture pedigrees. The methods demand minimal requirements for the amount of information present in the sequencing data, which makes them widely applicable in studies of non-model organisms with limited availability of genomic resources.

## 4 METHODS

### 4.1 Models of paired ancestries

We use ‘paired ancestries’ to model the patterns of interancestry heterozygosity. A paired ancestry is a vector $A = (a_1, a_2)$ that specifies the ancestry of each allele pair at a locus (for a diploid organism). The genome-wide paired ancestry proportions can be estimated with different models of ancestry. The models we use share the basic form of the standard admixture model from the widely used software ADMIXTURE and structure (Alexander et al., 2009; Pritchard et al., 2000), but they differ in how paired ancestries are modelled. For simplicity, we assume that the population allele frequencies $F$ for each of the $K$ populations are known. These can be obtained from ADMIXTURE if called genotypes are available, or from NGSadmixture (Skotte et al., 2013), when working with low-depth sequencing data. Neither method use information from the genomic position and can therefore work for RADseq data, or when working with draft genome assemblies as reference. We denote the genetic data for the focal individual as $X$, which can either be genotype data or genotype likelihoods for $M$ sites. For both the independent ancestry model used in ADMIXTURE, the paired ancestry model and the parental ancestry model we can write the likelihood as

$$P(X | F, \Theta) = \prod_{j=1}^{M} \sum_{a_1}^{K} \sum_{a_2}^{K} P(X_j | F_j, A = (a_1, a_2)) P(A = (a_1, a_2) | \Theta).$$  (1)
where $A = (a_1, a_2)$ indicates the paired ancestries, and $\theta$ indicates a un-specific parameter that will vary in each model, which are described in the following sections. $F_i = (f_{i1}, f_{i2}, \ldots, f_{iK})$ are the $K$ ancestral population allele frequencies for site $i$. The calculation of $P(X_j | A = (a_1, a_2), F_j)$ depends on whether we have called genotype data or genotype likelihood data. With genotype data $X$ is an $M$ length vector where $X_j \in \{0,1,2\}$ for site $j$, and we can calculate the probability as

$$P(X_j = g_j | A = (a_1, a_2), F_j) = \begin{cases} (1-f_{a_1})(1-f_{a_2}) & \text{if } g_j = 0 \\ f_{a_1}(1-f_{a_2}) + (1-f_{a_1})f_{a_2} & \text{if } g_j = 1 \\ f_{a_1}f_{a_2} & \text{if } g_j = 2 \end{cases} \quad (2)$$

When working with genotype likelihoods, $X$ is the sequencing data for $M$ sites. In this case, the probability of the data is given by

$$P(X | A = (a_1, a_2), F) = \sum_{g' \in \{0,1,2\}} P(G_j = g' | A = (a_1, a_2), F_j) P(X_j | G_j = g'). \quad (3)$$

where $P(X | G_j = g')$ is the genotype likelihood for genotype $g'$ at site $j$. They can be estimated from aligned sequencing data with software such as GATK (McKenna et al., 2010), ANGS (Korneliussen et al., 2014) or samtools (Li, 2011).

Different models of paired ancestries will differ in the parameterization of $P(A = (a_1, a_2) | \theta)$ with $\theta$ being a different set of parameters depending on whether the method assumes a independent ancestry model, a parental ancestry model or a paired ancestry model. In the following, we briefly describe how paired ancestries are modelled in the standard admixture model and in the two explicit models of paired ancestries we implement.

### 4.1.1 Standard admixture model

In the standard admixture model (Alexander et al., 2009; Pritchard et al., 2000), $\theta$ corresponds to the admixture proportions $Q = (q_1, q_2, \ldots, q_K)$. Under the admixture model, the population of origin of a pair of alleles is assumed to be independent given the admixture proportions, since

$$P(A = (a_1, a_2) | Q) = q_{a_1}q_{a_2}. \quad (4)$$

This assumption does not hold for inbred individuals or individuals who are very recently admixed. Both of the other models that we use to infer recent admixture avoid this latter assumption, and use the patterns of paired ancestry to detect and characterize cases of recent admixture.

### 4.1.2 Paired ancestry model

In the paired ancestry model (Gompert et al., 2014; Nohr et al., 2021), we use a vector $\Phi$ containing $\frac{K(K+1)}{2}$ parameters, with $\Phi = (\phi_{11}, \phi_{12}, \ldots, \phi_{K-1,K})$, where $\phi_{ik}$ denote the proportion of sites from the individual’s genome that carry one allele from population $a_i$ and the other from population $a_k$. The probability of a paired ancestry is therefore directly specified by $\Phi$

$$P(A = (a_1, a_2) | \Phi) = \phi_{a_1a_2}. \quad (5)$$

This model does not distinguish between the two possible ordering of paired ancestries, $P(A = (a_1, a_2)) = P(A = (a_2, a_1)) = \phi_{a_1a_2}$. It therefore only allows estimating ‘unordered paired ancestries’, as introduced above.

### 4.1.3 Parental admixture model

In the parental admixture model (Crouch & Weale, 2012; Pfaffelhuber et al., 2022), we model the ancestry of the sample using its two parental admixture proportions. For $K$ ancestral populations, we have $2K-2$ free parameters, with two $K$ length vectors giving the admixture proportions of each parent, $Q^P = (q_{11}^P, q_{21}^P, \ldots, q_{K1}^P)$ and $Q^{P2} = (q_{12}^P, q_{22}^P, \ldots, q_{K2}^P)$. The probability of a paired ancestry proportion is then given by

$$P(A = (a_1, a_2) | Q^P, Q^{P2}) = q_{a_1}^Pq_{a_2}^{P2}. \quad (6)$$

This model allows distinguishing the ordering of paired ancestries, since under this model $P(A = (a_1, a_2)) \neq P(A = (a_2, a_1))$. Therefore, from this model, we can calculate both unordered and ordered paired ancestries (Figure 1). Note that the method will have two equal solutions where the $P_1$ and $P_2$ labels are switched. For most downstream inference, we use the ordered paired ancestry estimates from this parental ancestry model, since its estimate relates directly to the recent admixture history of the individual.

### 4.2 Summarizing the information in the paired ancestry proportions

Given an estimate of the parental admixture proportions of an admixed individual, we can find the admixture pedigrees most compatible with the estimated proportions. We need to assume that the individual is the result of recent admixture, and that $g$ generations back all its ancestors are unadmixed. We can define an admixture pedigree, going back $g$ generations ago and involving $K$ ancestries, as two $K$ length vectors $N^P = (n_{11}^g, n_{21}^g, \ldots, n_{K1}^g)$ and $N^{P2} = (n_{12}^g, n_{22}^g, \ldots, n_{K2}^g)$. Each $n_{ik}^g$ gives the number of ancestors of parent $i=1, 2$ that are unadmixed from ancestry $k$, such that $\sum_{k=1}^{K} n_{ik}^g = 2^{g-1}$. The parental admixture proportions have no information to distinguish the ordering of ancestors of each parent in the pedigree, so any permutation of the set of ancestors is equivalent. Because the space of potential recent admixture pedigrees under this conditions is relatively limited, we chose to explore all compatible recent admixture pedigrees. The only constraint is that we only consider recent admixture pedigrees.
with admixture proportions in the vicinity of the estimated parental admixture proportions. Furthermore, we also consider the ‘independent pedigree’, that corresponds to the pedigree where all ancestors have the same admixture proportions, such that it is compatible with a case of not so recent admixture. This pedigree corresponds to the one expected under the standard admixture model used by ADMIXTURE.

Our approach of exploring and comparing multiple recent admixture pedigrees requires some way of ranking them by how well they explain the paired ancestry estimates. For this, we use the JSD between the ordered paired ancestry proportions estimated with the parental admixture model, and the ordered paired ancestry proportions expected under each of the considered admixture pedigrees. The JSD is a symmetrical measure of the distance between two probability distributions, and ranges from 0 to 1. In the context of comparing two paired ancestry proportions $P_1$ and $P_2$, the JSD is given by

$$JSD(P_1 \parallel P_2) = \sqrt{\frac{H(P_1 + P_2)}{2} - \frac{H(P_1) + H(P_2)}{2}},$$

where $H(P_x)$ gives the entropy of the paired ancestry distribution, which in this context can be expressed as

$$H(P_x) = -\sum_{a_1=1}^{K} \sum_{a_2=1}^{K} P(A = (a_1, a_2) | P_x) \log P(A = (a_1, a_2) | P_x),$$

for $x = 1, 2$, for ordered paired ancestries, and

$$H(P_x) = -\sum_{a_1=1}^{K} \sum_{a_2=0}^{K} P(A = (a_1, a_2) | P_x) \log P(A = (a_1, a_2) | P_x)$$

for $x = 1, 2$, for unordered paired ancestries.

In cases of complex admixture or when analysing many individuals, it can be difficult to interpret the results based on the paired ancestry proportions estimates and their fit to proposed admixture pedigrees. For this reason, we introduce two complementary indices that can be used as summary statistics to guide and prioritize the interpretation.

- **Inconsistency index**: this index is intended as an indicator of the reliability of the paired ancestry estimates. It is based on assessing the consistency between the paired ancestry estimates from the paired ancestry model, $P_{paired}$, and the ones from the parental admixture model, $P_{parental}$. We measure the consistency using the JSD between unordered paired ancestries probabilities defined by each model, such that a value of 0 indicates maximum consistency,

$$\kappa = JSD(P_{paired} \parallel P_{parental}).$$

This index reflects violations of the two main model assumptions in the parental ancestry model. The first one is the assumption of no inbreeding, which results in a higher proportion of homozygous paired ancestries than expected from a bifurcating pedigree. The paired ancestry model allows for excess of homozygous ancestries and thus does not make this assumption, which leads to the two models having different estimates in the case of inbreeding. The second assumption is that of having good estimates of allele frequencies for the populations from which the individual derives its ancestry. Many common situations, like admixture from ghost populations or continuous population structure, can lead to spurious admixture model results (Garcia-Erill & Albrechtsen, 2020; Lawson et al., 2018). While erroneous estimates of ancestral allele frequencies from such cases violate the assumptions of both models, it also results in different estimates between the two models. Therefore, the inconsistency index will also be elevated in the presence of this violation, making the index informative about this complication as well (see results).

- **Admixture index**: if an individual is recently admixed, the proportion of heterogeneous ancestry contains information on the admixture time and the number of admixture events. This index aims to reflect this information and is defined as

$$r = -\log_2(H_{min}) + 1,$$

where $H_{min}$ is the lowest non-zero proportion of heterozygous ancestry. In the simplest case with only one recent admixture event between two ancestries, $r$ is equal to the admixture time in generations. In more complex cases involving multiple ancestries, if we assume each ancestry is different from the others and only admixes once, $r$ is the sum of all admixture times in generations, making it still informative on the admixture process. In any case, the index is a lower bound on the sum of admixture times.

This index only makes sense if the individual is recently admixed and the estimated ordered paired ancestries are accurate. The value of $r$ therefore should not be considered unless the other metrics support this assumption, that is, it is meaningless if the best fitting pedigree is the independent ancestries pedigree, or if an elevated inconsistency index suggests the ordered paired ancestry estimates are not accurate. For example, if an individual is admixed from an ancestry that is not well represented in our inferred ancestral population frequencies, the model would lead to estimates that would not be reflective of the true recent admixture history even if it is recently admixed.

### 4.3 Estimating uncertainty with bootstrap

Our method for suggesting possible admixture pedigree is based on point estimates of paired and parental ancestry. However, there are several sources of uncertainty that are important to consider. There is statistical uncertainty from the fact that we only have data for a finite number of sites, and we might also have genotype uncertainty from low-depth sequencing data. There is also biological
uncertainty that arises from a finite number of recombination events which cause genotype correlations from LD and from ancestry tracts. We try to quantify these sources of uncertainty based on a bootstrap procedure using genomic blocks of data. However, depending on the data types it might not be possible to accommodate all of the biological uncertainty. If genomic position information is not available, like for RADseq data, we cannot define blocks of contiguous sites with data located next to each other on the genome. In this case, the estimated uncertainty from the bootstrap will only account for the uncertainty from having a finite number of sites and to low-depth sequencing, but it will not be able to account for the biological uncertainty caused by correlation of genotypes due to ancestry tracts and LD. If there is some genomic information available like for medium-sized scaffolds, block bootstrap will capture some of the biological uncertainty. But if the size of the genomic blocks is smaller than the extent of LD and admixture tracts it will not capture all of the uncertainty. We allow for the user to define a block size as the number of SNPs, similar to other the bootstrap used in other software (Pickrell & Pritchard, 2012). We perform 100 bootstrap samples where the SNP blocks are sampled with replacement. Then we obtain the bootstrap support for a given admixture pedigree as the proportion of bootstrap replicates for which the expected proportions for that pedigree is the closest to the estimates based on the JS distance.

4.4 | Data analyses

4.4.1 | Simulations

We used the R package ‘bnpsd’ (Ochoa & Storey, 2021) to simulate allele frequencies for 100,000 independent sites in two populations with varying degrees of divergence. From these allele frequencies, we sampled genotypes for 20 unadmixed individuals for each of the two populations, and for 10 recently admixed individuals for each of 5 different recent admixture classes. Each class consisted of a single admixture event between the two populations, and varied by the number of generations that has passed since admixture, from one generation (resulting in a F1 admixed individual) up to five generations (quadruple backcross). For each simulation and admixture classes, we ran ADMIXTURE assuming K = 2, and used the estimated population allele frequencies as input to estimate parental admixture proportions with our new implementation in NGSremix (Nohr et al., 2021).

4.5 | Admixed family trios from the 1000 genomes project

We tested our implementation of the parental admixture model using the family trios from the 1000G data set (Byrksa-Bishop et al., 2022). We used the admixed populations from the Americas that have Native American ancestry: Peruvian in Lima Peru (PEL), Mexican Ancestry in Los Angeles, CA, USA (MXL), Colombian in Medellin, Colombia (CLM) and Puerto Rican in Puerto Rico (PUR), and the two groups with African ancestry, African ancestry in SW USA (ASW), and African Caribbean in Barbados (ACB). We downloaded the publicly available genotype calls for all autosomes, and used bcftools v. 1.10 (Danecek et al., 2021) to keep only diallelic SNPs. We then used plink v.1.9 (Purcell et al., 2007) to remove related samples from these 6 admixed populations, and from the African population Yoruba in Ibadan, Nigeria (YRI) and the European population Utah residents with Northern and Western European ancestry (CEU). After filtering SNPs with minor allele frequency below 0.05, we kept 7,337,603 SNPs for 711 unrelated individuals. We used that data as input for ADMIXTURE assuming K = 3 to estimate the individuals admixture proportions and population allele frequencies. We did three independent optimization runs and, after assessing convergence by observing similar log likelihoods across runs, selected the results with the maximum final log likelihood.

We then used our implementation of the parental admixture model to estimate parental admixture proportions for the offspring of the trios from admixed populations. Furthermore, we also estimated parental admixture proportions for the offspring of the South Asian population Punjabi in Lahore, Pakistan (PJL), which had not been included in the estimation of ancestral population allele frequencies with ADMIXTURE. The aim in this case is to explore the performance of the method in cases of spurious admixture due to individuals with incorrect population allele frequencies. To assess the relationship between inbreeding and the inconsistency index, we estimated inbreeding coefficients for the trio offspring for which we had estimated paired ancestry proportions. We used PCAngsd (Meisner & Albrechtsen, 2018) implementation of the ngsF maximum likelihood method to estimate inbreeding (Meisner & Albrechtsen, 2019; Vieira et al., 2013), using the individual allele frequencies estimated using three principal components to control for population structure.

4.6 | Low-depth WGS data set (waterbuck)

We got access to low coverage (3X) sequencing data of waterbucks (Kobus ellipsiprymnus) mapped to the defassa waterbuck (K. e. defassa) draft genome (DFW) and a list of quality controlled sites (Wang et al., 2022). From that, we estimated genotype likelihoods with ANGSD (Korneliussen et al., 2014) using the GATK genotype likelihood model (McKenna et al., 2010). After keeping only sites with minor allele frequency 0.05, we had genotype likelihoods for 9,581,116 SNPs from 38 individuals, from which we estimated admixture proportions and ancestral population allele frequencies with NGSadmix (Skotte et al., 2013), assuming K=4. We did 20 independent runs and, after assessing convergence, selected the run with the maximum final log likelihood. We then estimated paired ancestry proportions with the two models for admixed individuals, defined as those whose major ancestry proportion was below 97% in the NGSadmix results.
4.7 | RADseq data set (Grant’s gazelle)

We obtained the publicly available alignment files of Grant’s gazelle (Nanger ssp.) RADseq data mapped to the domestic cow bosTau8 (Bos taurus) reference genome, and a list of quality controlled sites (Garcia-Erill et al., 2020). We estimated genotype likelihoods using ANGSD with the GATK model, for 70 individuals from five populations. We estimated genotype likelihoods for a total of 33,656 SNPs with minor allele frequency higher than 0.05 and used these as input in NGSadmix assuming K = 5, with 20 independent runs and selecting the run with the maximum final log likelihood. We then estimated paired ancestry proportions with the two models for admixed individuals, defined as those whose major ancestry proportion was below 0.97 in the NGSadmix results.

ACKNOWLEDGEMENTS
We thank Hans Redlef Siegismund, Laura Bertola, Malthe Sebro Rasmussen, Ida Moltke and all the members of the Population and Statistical Genetics group at the University of Copenhagen for their input in previous versions of the manuscript. We also thank Xi Wang and Mikkel Schubert for their contribution in processing the waterbuck data set. GGE and RH are supported by a Danmarks Frie Forskningsfond Sapere Aude research grant (DFFF8049-00098B). RH is further supported by an European Research Council Starting Grant (No. 853442), and CW and AA are supported by the Independent Research Fund Denmark (grant number: 8021-00360B) and the University of Copenhagen through the Data+ initiative.

CONFLICT OF INTEREST STATEMENT
The authors declare that they have no conflict of interest.

OPEN RESEARCH BADGES

This article has earned an Open Data badge for making publicly available the digitally-shareable data necessary to reproduce the reported results. The data is available at SRA https://www.ncbi.nlm.nih.gov/sra/ with BioProject IDs PRJNA673069 and PRJNA985556.

DATA AVAILABILITY STATEMENT
No new data were generated as part of this study. The variant calls of the 1000G project data can be accessed in http://ftp.1000genomes.ebi.ac.uk/vol1/ftp/data_collections/1000G_2504_high_coverage/working/20201028_3202_phased/, and the raw sequencing data for the Grant’s gazelle RADseq data set and the waterbuck low-depth sequencing data sets can be accessed in SRA BioProject IDs PRJNA673069 and PRJNA985556, respectively.

ORCID
Genis Garcia-Erill  https://orcid.org/0000-0003-3150-1708
Rasmus Heller  https://orcid.org/0000-0001-6583-6923

REFERENCES


Nohr, A. K., Hanghoj, K., Garcia-Erill, G., Li, Z., Moltek, I., & Albrechtsen, A. (2021). NGRemix: A software tool for estimating pairwise relatedness between admixed individuals from next-generation sequencing data. G3 (Bethesda), 11(8), jkab174.


SUPPORTING INFORMATION

Additional supporting information can be found online in the Supporting Information section at the end of this article.