Inference of recent admixture using genotype data

PETER PFAFFELHUBER, ELISABETH HUSS, FRANZ BAUMDICKER,
DENISE SYNDERCOMBE COURT, FABIAN STAUBACH

December 2, 2019

TODO

1. What are we doing with the Tvedebrink-tests?
2. Local ancestry with tsinfer;
3. Analyse data from Denise;

Abstract

For inference of individual genetic histories, admixture barplots are being used abundantly in forensic genetics. These plots visualize parameters for individual ancestry (IA), as inferred from the admixture-model, as e.g. implemented in the software STRUCTURE and ADMIXTURE. In this model, it is assumed that every allele in the individual's genome originates in one of several ancestral populations with the same probability. We will showcase that estimates of IA might be inaccurate in cases of recent adxmixture. As a way out, we introduce the recent-admixture model, which makes use of the excess of heterozygotes observed in recently admixed individuals. In this model, we assume that the two homologous copies of each allele originate from the ancestral populations independently for the mother's and the father's copy. Estimates for IA in absence of recent admixture are almost identical for the admixture and recent admixture model. However, they are more accurate for the recent-admixture model in individuals which are in fact recently admixed. Moreover, we develop a likelihood ratio test for recent admixture, which has a high power to find recently admixed individuals. We analyse data from the 1000 genomes dataset with our methods and find some recently admixed individuals.

1 Introduction

Inference of the geographical ancestry of a trace using genetic markers is today a well-established research field in forensic genetics (see e.g. [15, 5, 9]). Either the trace is classified into one of several groups of different origin (e.g. Africa, Europe, East Asia, Native America, and Oceania; see e.g. [14, 12], or it is assumed that it consists of a mixture of ancestral genetic material originating in

several groups. For this task, STRUCTURE [4] has become the de-facto standard to estimate individual ancestry (IA) proportions of a trace among several continental groups. This Bayesian approach using MCMC was complemented by the faster, likelihood based approach, implemented in the software ADMIXTURE [1]; see also [5] for the same model.

When using STRUCTURE or ADMIXTURE for estimating continental (or other scales of geography) IA of a trace, allelic frequencies for all continents from a reference database must be used. A main assumption for estimating IAs is that all allelic states have the same independent chance to share a copy from some continent. When using a (small) set of ancestry informative markers, the large genomic distance makes loci almost independent by recombination, resulting in independence of markers across loci. However, independence of the two copies at the same marker, usually denoted by Hardy-Weinberg equilibrium, may not always hold. For example, consider an individual whose parents have different continental backgrounds, A and B, say, and an allele which separates perfectly between the two continents. Then, the individual will certainly be a heterozygote at this marker. This increase in frequency for heterozygotes under admixture is known for a long time and usually called the Wahlund effect [20]. However, STRUCTURE and ADMIXTURE will estimate that the chance for some allele to come from A and B to be 50%, which is then also be the estimated probability for a heterozygote at the locus (due to the assumption of independence of allelic states in the admixture model).

In general, STRUCTURE (and ADMIXTURE) are able to give accurate estimates of IA also in the case of recent-admixture. As observed in [4], a newly developed Genetic Distance Algorithm preforms better than STRUCTURE for recently admixed individuals up to second generation only if all four grand-parents come from different populations. However, STRUCTURE may lead to wrong conclusions in some situations. E.g., it might happen that a individual, recently admixed from two distant populations, is estimated to come mainly from some intermediate location. xxx Make a good case, e.g. Philippino and Italian located in Afghanistan.

Our goal is to make use of the Wahlund effect for (i) improving estimates of IA and (ii) detecting recent admixture. In particular, we extend the likelihood-model behind STRUCTURE or ADMIXTURE in order to account for recent admixture. In the resulting recent-admixture-model, IA consists of two vectors, one for each parent. We will call this pair of vectors Parental Individual Admixture, PIA. So, in the above example, the recent-admixture-model would estimate father and mother to have IAs as 100% A and 100% B, respectively, due to the excess of heterozygotes. Since we are using only autosomal markers, we cannot distinguish which of the parents comes from A and B, though. Then, with the estimate of PIA, the estimated probability for a heterozygote is 1 in the example.

The excess of heterozygotes was recently used to detect recent admixture using statistical tests. For example, [11] uses a statistical test by simply counting the number of heterozygous positions in an individual in order to infer recent admixture. A different approach is used by [19, 19]. Here, likelihood-ratio tests were developed which test the null-hypothesis of non-admixture and recent (first generation) admixture versus the alternative that the studied sample is not represented in the reference database. We complement this test by observing that the adxmixture model is a special case of the recent-admixture model, if the IA of both parents is the same. This paves the way to give a likelihood-ratio-test for recent admixture, with equality of the parental IAs as null-hypothesis. The alternative

hypothesis, however, is that the studied sample shows recent admixture of populations within the reference database.

2 Materials and Methods

We start by briefly recalling the admixture model, which is the basis for the widely used sofware STRUCTURE [4], ADMIXTURE [1] and FRAPPE [5]. Afterwards, we introduce a new model, called the recent-admixture model. More details on the derivations in the admixture and recent-admiture model can be found in the SI. Moreover, the implementation of our methods can be downloaded from `https://github.com/pfaffelh/recent-admixture`. For both, the adxmiture and recent-admixture model, we assume to have reference database of M bi-allelic markers from K populations. However, from this reference database, we only need to known allele frequencies, i.e. by p_{mk}, the frequency of allele 1 at marker m in population k for all $m = 1, ..., M$ and $k = 1, ..., K$. We have a trace with $G_m \in \{0, 1, 2\}$ copies of allele 1 at marker m for $m = 1, ..., M$. We will assume throughout that the allele frequencies p_{mk} are given and will not be changed by analysing the trace. This is important since in currently used software STRUCTURE, ADMIXUTURE and FRAPPE, mostly in non-forensic use, it is frequently the case that many new individuals are studied, and allele frequencies are updated. For forensic use, when analysing several traces at once, this would imply that the results for the ancestry of trace 1 depend not only on the reference data, but also on the data for traces 2, 3,... which seems inappropriate. Hence, we do not make the computational overload of updating allele frequencies, which would also lead to increased runtimes and take the allele frequencies as given in the reference database. With other words, we will follow the approach of [2] and [3] for analysing our models.

2.1 The admixture model

Assuming that each allele observed in the trace comes from population k with probability q_k, the probability to observe allele 1 at marker m is

$$\beta_m(q) := \sum_k p_{mk} q_k, \tag{1}$$

and the log-likelihood of $q = (q_k)_{k=1,...,K}$ is (see also (S2) in the SI)

$$\ell(q|G) = \sum_{m=1}^{M} \log\Big(\binom{2}{G_m} \beta_m(q)^{G_m} (1 - \beta_m(q))^{2-G_m}\Big). \tag{2}$$

Assuming that all p_{mk}'s are known, this function can be maximized over q by computing $\hat{q} = (\hat{q}_k)_{k=1,...,K}$ such that $\hat{q}_k = f_k(\hat{q})$ for (see also (S4) in the SI)

$$f_k(q) = \frac{1}{2M} \sum_{m=1}^{M} \Big(G_m \frac{p_{mk}}{\beta_m(q)} + (2 - G_m) \frac{1 - p_{mk}}{1 - \beta_m(q)}\Big) q_k, \qquad k = 1, ..., K. \tag{3}$$

This can be done numerically by iterating $q_{n+1} = (f_k(q_n))_{k=1,\dots,K}$ until convergence. (In our implementation, we continue the iteration until $|q_{n+1} - q_n| < 10^{-6}$.) We note that this approach is essentially the same as in the EM-algorithm from [5], but combining the expectation and maximization steps, since we do not update allelefrequencies. In addition, although maximizing (2) could also be handled using a Newton method as in [1], this approach has the advantage that q_n's are positive in all steps, and the sum of all entries in q_n is always 1. Moreover, the iteration is computationally fast if only a small or moderate number of alleles is considered.

2.2 The recent-admixture model

When mother and father of an individual come with their own vectors of admixture proportions, q^M and q^P, the log-likelihood from (2) changes to (see also (S5) in the SI)

$$\ell(q^M, q^P|G) = \sum_{m=1}^{M} \log(\gamma_m(q^M, q^P, G_m)), \tag{4}$$

$$\gamma_m(q^M, q^P, g) = \begin{cases} \beta_m(q^M)\beta_m(q^P), & \text{if } g = 2, \\ (\beta_m(q^M)(1-\beta_m(q^P)) + (1-\beta_m(q^M))\beta_m(q^P)), & \text{if } g = 1, \\ (1-\beta_m(q^M))(1-\beta_m(q^P)), & \text{if } g = 0. \end{cases}$$

As carried out in the SI, this function can be maximized by computing $\hat{q}^M, \hat{q}^P$ such that $\hat{q}^P = f(\hat{q}^M, \hat{q}^P)$ and $\hat{q}^M = f(\hat{q}^P, \hat{q}^M)$ for $f(q, q') = (f_k(q, q'))_{k=1,\dots,K}$ with (see (S7) in the SI)

$$f_k(q, q') := \frac{1}{M} \sum_{m=1}^{M} \delta_k(q, q', G_m) q'_k, \tag{5}$$

$$\delta_k(q^M, q^P, g) = \begin{cases} \dfrac{p_{mk}}{\beta_m(q')}, & \text{if } g = 2, \\ \dfrac{(p_{mk}(1-\beta_m(q)) + (1-p_{mk})\beta_m(q))}{\beta_m(q)(1-\beta_m(q')) + (1-\beta_m(q))\beta_m(q')}, & \text{if } g = 1, \\ \dfrac{(1-p_{mk})}{1-\beta_m(q')}, & \text{if } g = 0. \end{cases}$$

In our implementation, we iteratively compute $q^P_{n+1} = f(q^M_n, q^P_n)$ and $q^M_{n+1} = f(q^P_{n+1}, q^M_n)$ until convergence.

2.3 Obtaining admixed individuals in silico

In order to test our method, we created admixed individuals from a reference database. (We use the 1000 genomes dataset, but excluding Admixed Americans, AMR; see below.) For example, we obtain an individual admixed from populations k and k' by choosing a genome $\tilde{G} = (\tilde{G}_m)_{m=1,\dots,M}$ from population k and $\bar{G} = (\bar{G}_m)_{m=1,\dots,M}$ from population k' as the parents. Then, $G_m = X_m + Y_m$ where $X_m = 1$ with probability $\tilde{G}_m/2$, $X_m = 0$ with probability $(2 - \tilde{G}_m)/2$ and $Y_m = 1$ with

probability $\bar{G}_m/2$, $Y_m = 0$ with probability $(2 - \bar{G}_m)/2$. When iterating this procedure, we can also model second-order admixed individuals etc. in silico.

Using AFR, EAS, EUR, SAS as population labels (as in the 1000 genomes dataset), all cases for second generation admixed individuals fall into one of seven categories. Writing up the ancestries of the four grand-parents *Mother of mother/father of mother*×*mother of father/father of father*, we have the following distinguishable cases for second generation admixed individuals (the full list of all resulting 55 cases is given in the SI; note that [4] come up with only 35 cases, since they do not distinguish between maternal and paternal ancestry, e.g. they count AFR/AFR×EAS/EAS and AFR/EAS×AFR/EAS as one case):

(A) 4 non-admixed cases, e.g. AFR/AFR×AFR/AFR;

(B) 6 admixed cases with admixture ratio 50:50, where both parents are non-admixed, e.g. AFR/AFR×EAS/EAS;

(C) 6 admixed cases with admixture ratio 50:50, where both parents are admixed, e.g. AFR/EAS×AFR/EAS;

(D) 12 admixed cases with admixture ratio 75:25, e.g. AFR/AFR×AFR/EAS;

(E) 12 admixed cases with admixture ratio 50:25:25, where one parent is non-admixed, e.g. AFR/AFR×EAS/EUR;

(F) 12 admixed with admixture ratio 50:25:25, where both parents are admixed, e.g. AFR/EAS×AFR/EUR;

(G) 3 admixed with admixture ratio 25:25:25:25, e.g. AFR/EAS×EUR/SAS;

For each of the other 55 cases, we simulated 500 individuals in silico by picking four grand-parents at random from the populations, creating mother and father from the grand-parents, and creating a new individual from the parents, as described above.

2.4 Comparing results from admixture and recent-admixture

For a reference database from which we compute (or estimate) allele frequencies p_{mk} (which is the allele frequency of allele 1 at marker m in population k), we can estimate q from the admixture model as well as q^M, q^P from the recent-admixture model, as described in (3) and (5). In order to compare the results from the admixture and recent-admixture model, we compute $q_k^{MP} := \frac{1}{2}(q_k^M + q_k^P)$ for $k = 1, ..., K$, which give the fractions of the genome coming from populations $1, ..., K$. Then, for a non-admixed individual, we have $q_k^{\text{TRUE}} = 1$ for some k, and for an admixed individual with parents from populations k and k' we have $q_k^{\text{TRUE}} = q_{k'}^{\text{TRUE}} = 0.5$, and similarly for individuals with grandparents from two up to four different populations. Computing the estimation error, i.e. the *distance to the true IA* for the admixture model, results in

$$\sum_k |q_k - q_k^{\text{TRUE}}| \text{ and } \sum_k |q_k^{MP} - q_k^{\text{TRUE}}| \tag{6}$$

for the recent-admixture model. We stress that in the recent-admixture model, we in fact obtain results for q^M and q^P separately, such that even more information than q^{MP} is contained in the estimates for this model.

2.5 A likelihood-ratio test for recent admixture

We want to test if data $G = (G_m)_{m=1,\ldots,M}$ from a new trace fits significantly better to the recent-admixture model than to the admixture model. Since the admixutre model is identical to the recent-admixture model for $q^M = q^P = q$, we are testing the hypothesis $H_0 : q^M = q^P$ against $H_1 : q^M \neq q^P$. For this, we take the estimators $\hat{q}$ of q from iteration of (3), and $\hat{q}^M, \hat{q}^P$ of q^M and q^P from iteration of (5) and compute

$$\Delta\ell := \ell(\hat{q}^M, \hat{q}^P|G) - \ell(\hat{q}|G) \tag{7}$$

with $\ell(q^M, q^P|G)$ from (4) and $\ell(q|G)$ from (2). If $\Delta\ell > x$ for some x (which needs to be specified), the recent-admixture model fits significantly better and we reject H_0. If $\Delta\ell \leq x$, we accept H_0. In order to find x, we fix a p-value (0.01, say), and take x to be the p-quantile of the empirical distribution for data in the reference database. (This means, if $p = 0.01$ and the reference dataset contains 1000 samples, we compute all values for $\Delta\ell$ for all samples, and set x to be the 10th-smallest value we obtained.)

2.6 Human data

We downloaded 1000 Genomes data (phase 3) from `ftp://ftp.1000genomes.ebi.ac.uk/vol1/ftp/release/20130502/`, as well as information on the sampling locations from `ftp://ftp.1000genomes.ebi.ac.uk/vol1/ftp/release/20130502/integrated_call_samples_v3.20130502.ALL.panel` [1]. This is data from 661 individuals from Africa (AFR), 347 Admixed Americans (AMR), 504 East Asians (EAS), 503 Europeans (EUR) and 489 South Asians (SAS). The dataset comes with approximately 80 million SNPs. However, we use only a few of them known as the EUROFORGEN AIMset [13] and Kidd AIMset [10], respectively. The former comes with 128 SNPs, and we ignore seven tri-allelic SNPs (rs17287498, rs2069945, rs2184030, rs433342, rs4540055, rs5030240, rs12402499), since our methods currently rely on bi-allelic SNPs. It was designed to distinguish Africa, Europe, East Asia, Native America, and Oceania, but was shown to perform well on the 1000 genomes dataset, also for distinguishing South Asia, even when ignoring the tri-allelic SNPs [12]. The latter comes with 55 bi-allelic SNPs and was introduced as a global AIMset differentiating between 73 populations. We note that this AIMset is part of the Verogen MiSeq FGx™ Forensic Genomics Solution.

3 Results

3.1 Some showcases from simulations

We simulated genome-wide data from a sample, taken from a population genetic model with three islands, A, B and C, using [8]. Migration is such that only A, B and B, C are connected, but not A, C. We used a migration rate of 10 diploid individuals per connected islands (in both directions) per generation. More precisely, we simulate a sample of 400 individuals per island, each with 20 recombining chromosomes, each with about $2.5{\cdot}10^4$ SNPs. From these $\sim 5 \cdot 10^5$ SNPs, we use the step-wise approach from [12] to look for 10 Ancestry Informative Markers (AIMs). When using a naive Bayes approach as in SNIPPER [14], this AIMset gives a vanishing misclassification error for the task of classifying the $3{\times}400$ simulated, non-admixed individuals.

Subsequently, we used the admixture and recent-admixture model to estimate IA (and PIA) for both, non-admixed and recently (first generation) admixed individuals. The latter were obtained as described in Section 2.3. We observe that the admixture model fails to give accurate estimates for IA in $A{\times}C$-recently-admixed individuals for two reasons. First, it correctly predicts that the individual is $A \times C$-admixed, but overestimates one of the two ancestral proportions. Second, and more severely, it confounds the signal for recent admixture with an ancestral proportion from island B. In both cases, the recent-admixture model frequently gives better estimates for IA. Figure 1 shows examples for these two cases, which are taken from all $A \times C$-admixed individuals as displayed in Figure S1.

To get an overall picture, we computed errors for estimating IA as given in (6) for the admixture and recent-admixture model. As described in MM, we average estimates $\hat{q}^M$ and $\hat{q}^P$ from the recent-admixture model, in order to compare to the true IAs. Figure S2 displays these errors in all cases including non-admixed individuals, and all three cases of recent-admixture. Interestingly, binomial tests with the alternative that the recent-admixture model gives smaller errors show significant results in all but one case ($A, C, A \times C$: $p < 0.001$, $B : p = 0.001$, $A \times B : p = 0.004$, $B \times C$: $p = 0.23$). This shows that on average the recent-admixture model performs better than the admixture model, even on non-admixed samples.

3.2 Estimation accuracy

For comparing the accuracy of the admixture and recent-admixture model, we extended our analysis of the simuated data to the 1000 genomes dataset. We excluded all Admixed Americans (AMRs) since they are known to have an admixed background [5, 12] and do not form a well-defined own group. As the true IA, we use the continental origins as described in the dataset, i.e. we have AFR (African), EAS (East-Asia), EUR (European) and SAS (South Asian) samples. This means e.g. that we set $q^{\text{TRUE}}_{\text{EUR}} = 1$ for a European sample.

We ran three kinds of analyses. First, on the non-admixed samples, i.e. the original 1000 genomes data (denoted AFR,...). Second, we produced in silico recently admixed individuals with parents from the non-admixed samples (denoted AFR$\times$EAS etc.) and ran the analysis on these samples. Third, the analysis was performed on second-generation admixed samples, i.e. grand-parents were taken from

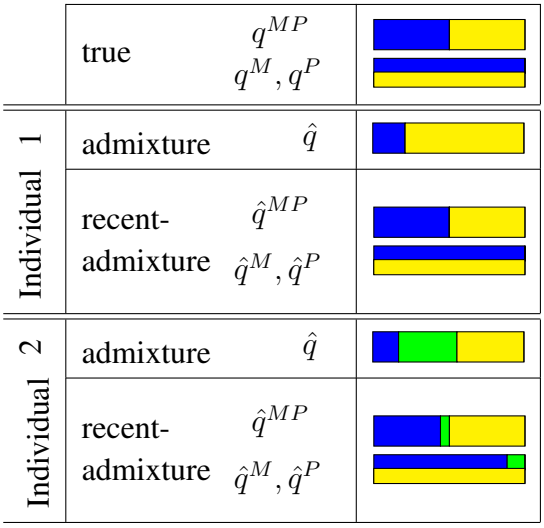


Figure 1: Colors indicate IA from islands A ■, B ■ and C ■. Here, in the model where only $A \leftrightarrow B$ and $B \leftrightarrow C$ exchange migrants (but not $A \not\leftrightarrow C$), admixture has two sources of errors when estimating IA for a recently admixed individual from $A \times C$. In individual 1, C is overrepresented in the IA-vector q. In individual 2, admixture gives the greatest fraction of ancestry to island B, since allele frequencies in B are between A and C. In both cases, the IA in terms of the combined vector $q^{MP} = (q^M + q^P)/2$ from the recent-admixture model gives more accurate estimates.

the non-admixed samples (denoted AFR/EAS×EUR/SAS etc). In the first case, Figure S3 shows that the resulting errors for the admixture and recent-admixture model are almost identical. Overall, recent-admixture has a smaller error in 1364 out of 2157 cases, i.e. th hypothesis that the error for recent-admixture is at least as large as for admixture can be rejected (binomial test, $p < 0.001$. In the second case, Figure 2(A) shows clearly that errors for recent-admixture are smaller for all pairs of continents. More precisely, in 2279 out of 3000 individuals, recent-admixture is more accurate ($p < 0.001$). Third, for second-generation admixed individuals, Figure 2(B) displays errors in the cases (A)–(G), and shows that again, recent-admixture is more accurate. Here, recent-admixture outperforms admixture in 15761 out of 27500 cases (resulting in $p < 0.001$) A full list of 55 cases is displayed in Figure S4 in the SI. The corresponding results for the Kidd AIMset are similar and found in the SI.

3.3 Power of the Likelihood-ratio test for recent admixture

When fixing the maximal p-value for significance of the likelihood-ration test for recent admixture (as described in MM), we obtain the power of the test for all cases of recent admixture. Displaying the false positives (i.e. positively tested non-admixed) against true positives (i.e. positively tested admixed) in cases (B)–(G) for all possible values of p, we obtain the Receiver-Operation-Characteristic (ROC) curve [6]. The optimal curve nearly hits 0 false positives with 100% true positives and has an AUC (Area Under the Curve) of 1. As we see in Figure 3, the power of the test differs with the kind of admixture. For first generation admixed (case (B)), one non-admixed parent (case (E))

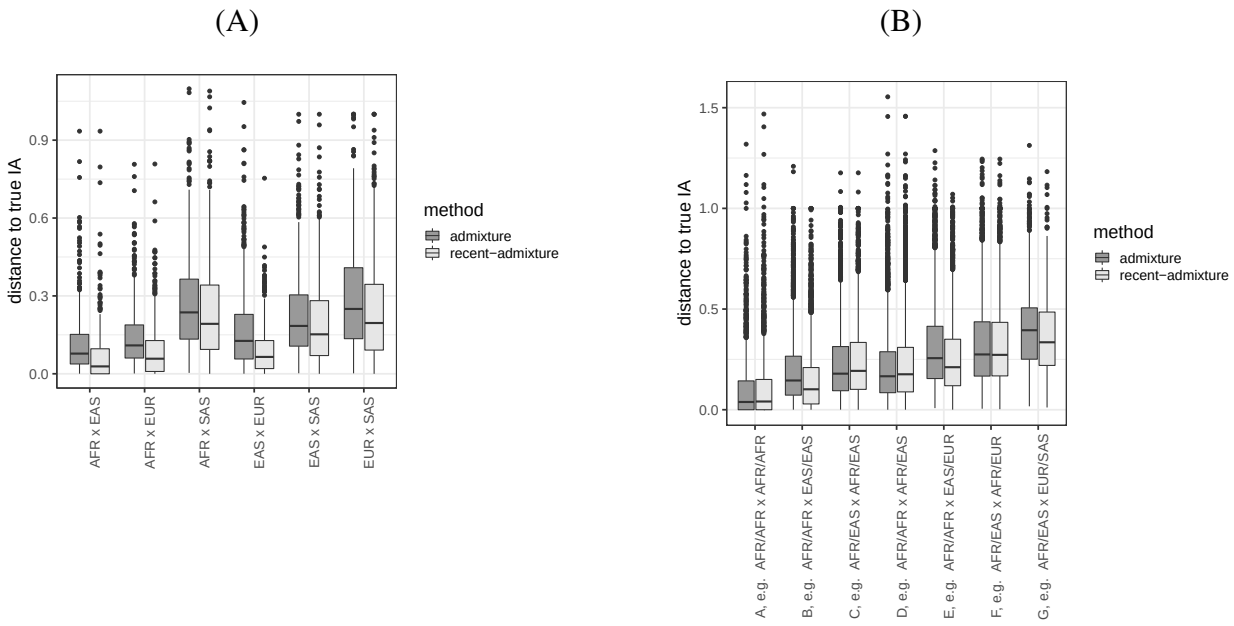


Figure 2: For all first generation admixed samples (A) and second generation admixed samples (B), we computed IA from the admixture and recent-admixture model. The distance to the true IA is computed as in (6). The cases in (B) are as described above.

Case	B	C	D	E	F	G
AUC	0.99	0.637	0.872	0.983	0.928	0.99
Power at $p = 0.01$	0.94	0.23	0.51	0.9	0.62	0.9

Table 1: Using the same data as in Figure 3, we e.g. see that the test for recent admixture turns out to have a p-value below 1% in 94% cases of first generation admixed individuals.

and all grand-parents from different continents (case (G)), the test is nearly perfect in distinguishing recent-admixture from admixture. If only half of the genome has two different ancestries (cases (D) and (F)), the power is reduced. If the individual is not recently-admixed in first generation, but both parents are (case (C)), power drops even more. In fact, the latter case is not recent-admixture as in our definition, since $q^M = q^P$ should technically hold. For the overall performance of the test, we give some examples in Table 1, in particular the power at $p = 0.001$ and AUC in all cases. Results for the EUROFORGEN AIMset are marginally better than for the Kidd AIMset, which are given in the SI in Figure S8 and Table S1.

3.4 Recent admixture in the 1000 genomes dataset

From the 1000 genomes dataset, we highlight individuals which give significant results for the test of recent admixture for both AIMsets. We exclude AMRs from our analysis, and give in Table 4 the eight most extreme cases, which show highly significant results for recent admixture for both

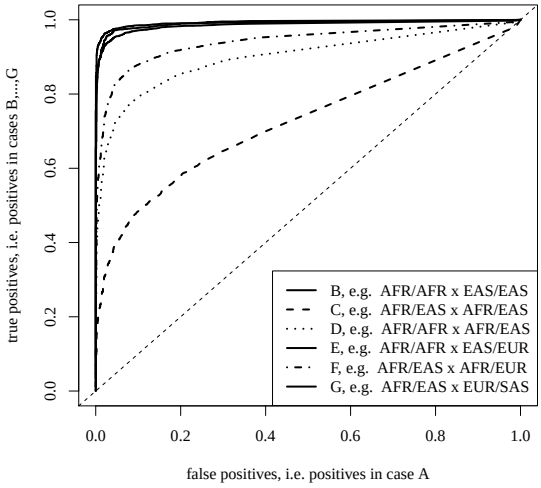


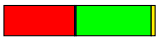
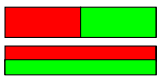
Figure 3: Using the EUROFORGEN AIMset, we plot false positives (i.e. positive non-admixed individuals, as in case (A) above, againts positives for all cases of admixture in second generation.

AIMsets. There are six Africans from population ASW (Americans from Southwest USA), and two from South Asia, one from GIH (Gujarati Indian from Houston, Texas) and one from BEB (Bengali from Bangladesh). We note that it is known that the ASW population is admixed [5], but until now, it has not been tested if admixture is recent.

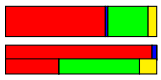
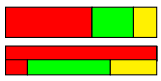
- NA20278: Giving the most significant results for both datasets, this male has most probably parents of African and European ancestry. Note also that q^{MP} and q are very similar for both AIMsets.
- NA20342, NA19625, NA20355: Clearly, one parent has African ancestry. The other parent is most likely partly European.
- NA20274: Our test indicates two parents of different ancestry, one mostly African, the other mostly East-Asian.
- NA20299: Interestingly, the results for both AIMsets differ in this example. One parent has most likely African ancestry, the other is European according to the EUROFORGEN AIMset and South Asian according to the Kidd AIMset.
- HG03803: Most likely, one parent is of South Asian, the other has East-Esian ancestry.
- NA20864: Most likely, one parent is of South Asian, the other has European ancestry.

In order to get a clearer picture about *true* ancestries in these individuals, we examined local ancestry in the genome. xxx

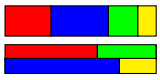
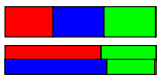
NA20278, AFR, ASW

EURO	ad	q	
$\Delta\ell = 6.69$	r-ad	q^{MP}	
		q^M, q^P	
Kidd	ad	q	
$\Delta\ell = 9.98$	r-ad	q^{MP}	
		q^M, q^P	

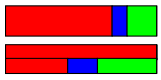
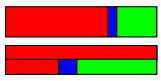
NA20342, AFR, ASW

EURO	ad	q	
$\Delta\ell = 1.82$	r-ad	q^{MP}	
		q^M, q^P	
Kidd	ad	q	
$\Delta\ell = 3.66$	r-ad	q^{MP}	
		q^M, q^P	

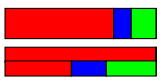
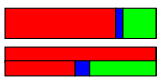
NA20274, AFR, ASW

EURO	ad	q	
$\Delta\ell = 6.56$	r-ad	q^{MP}	
		q^M, q^P	
Kidd	ad	q	
$\Delta\ell = 2.68$	r-ad	q^{MP}	
		q^M, q^P	

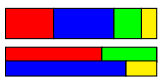
NA19625, AFR, ASW

EURO	ad	q	
$\Delta\ell = 1.52$	r-ad	q^{MP}	
		q^M, q^P	
Kidd	ad	q	
$\Delta\ell = 1.6$	r-ad	q^{MP}	
		q^M, q^P	

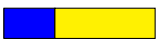
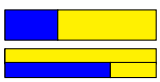
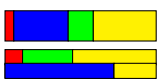
NA20355, AFR, ASW

EURO	ad	q	
$\Delta\ell = 1.11$	r-ad	q^{MP}	
		q^M, q^P	
Kidd	ad	q	
$\Delta\ell = 1.55$	r-ad	q^{MP}	
		q^M, q^P	

NA20299, AFR, ASW

EURO	ad	q	
$\Delta\ell = 6.31$	r-ad	q^{MP}	
		q^M, q^P	
Kidd	ad	q	
$\Delta\ell = 1.52$	r-ad	q^{MP}	
		q^M, q^P	

HG03803, SAS, BEB

EURO	ad	q	
$\Delta\ell = 1.01$	r-ad	q^{MP}	
		q^M, q^P	
Kidd	ad	q	
$\Delta\ell = 1.12$	r-ad	q^{MP}	
		q^M, q^P	

NA20864, SAS, GIH

EURO	ad	q	
$\Delta\ell = 0.87$	r-ad	q^{MP}	
		q^M, q^P	
Kidd	ad	q	
$\Delta\ell = 0.95$	r-ad	q^{MP}	
		q^M, q^P	

Figure 4: The most extreme individuals in the 1000 genomes dataset in terms of a signal for recent admixture. For all individuals, we give IA from the admixture model (ad), given by q, and the recent-admixture model (r-ad), given by $q^{MP} = \frac{1}{2}(q^M + q^P)$. In this case, we also give PIA, given by q^M, q^P. We use both, the EUROFORGEN and Kidd AIMset. Difference in log-likelihoods for both models is given by $\Delta\ell$; see (7). Colors are AFR, EAS, EUR, SAS.

3.5 Runtimes

The result is that the analysis is fast. In particular, it scales linearly with the number of analysed traces. E.g. once allele frequencies for all AIMs from the reference dataset are given,the $500 \cdot 55 = 27500$ individuals which need to be analysed for Figure 2, and which are created in silico, takes about 1.5 seconds per individuals on a standard laptop computer using the statistical language R.

4 Discussion

xxx add references, write some more paragraphs

xxx Mention [16].

We only need allele frequencies. These could e.g. be looked up in Frog-kb. No full reference dataset is necessary.

We address the inference of recent admixture in estimating individual admixture (IA), when data from Ancestry Informative Markers is available. The well-known admixture model assumes that in each individual for some $q = (q_k)_{k=1,\dots,K}$, every allele has ancestry within population k with probability q_k. We extend this model by distinguishing between maternally and paternally inherited alleles (but still consider only autosomes). The maternally inherited alleles come with IA q^M, and the paternally inherited alleles with q^P, which extends the admixture model if $q^M \neq q^P$. Within this model, q^P and q^M can be estimated by the same methods as in the admixture model. Taking the average of q^M and q^P, we obtain IA which is highly similar to q in the admixture model in non-admixed and admixed individuals. In addition, the IA estimated from the recent-admixture model in recently admixed individuals, in particular if the two parents come from different populations, is more accurate than for the admixture model. Moreover, and in contrast to the admixture model, we also estimate the form of recent admixture. A resulting likelihood ratio test for recent admixture has high power to detect recent admixture in many relevant scenarios, in particular if the two parents come from different populations.

In many applications, STRUCTURE still seems to be the gold-standard for estimating the IA of traces. STRUCTURE uses the same admixture model as we do, it is computationally much more demanding. The same holds for ADMIXTURE and FRAPPE. The reason is that these programs aim to simultaneously estimate the IA of the new trace, as well as allele frequencies in ancestral populations. In other words, in these programs, the new trace changes allelic frequencies in ancestral populations during the runtime. However, as in forensics we mostly have a large reference database and only one (or a few) new traces, the impact on the allelic frequencies are negligible. Therefore, we take the computationally easier way and take allele frequencies for ancestral populations as fixed, and only change the IAs of the new traces during the runtime. The result is that the analysis is fast. E.g. the $500 \cdot 55 = 27500$ individuals which need to be analysed for Figure 2, and which are created in silico, takes about five hours on a standard laptop computer using the statistical language R.

Not included the GDA from Cheung 2019. In addition, we also only need frequencies from the reference database (or some external source, e.g. from Frog-kb), although we use the same model as structure. Reason is that we do not update frequencies. Updating frequencies would mean that the

trace can alter ancestral allelic frequencies.

References

[1] 1000 Genomes Project Consortium, Adam Auton, Lisa D. Brooks, Richard M. Durbin, Erik P. Garrison, Hyun Min Kang, Jan O. Korbel, Jonathan L. Marchini, Shane McCarthy, Gil A. McVean, and Gonçalo R. Abecasis. A global reference for human genetic variation. *Nature*, 526(7571):68–74, 2015.

[2] D. H. Alexander, J. Novembre, and K. Lange. Fast model-based estimation of ancestry in unrelated individuals. *Genome Research*, 19:1655–1664, 2009.

[3] R. Chakraborty. Gene Admixture in Human Populations: Models and Predictions. *Yearbook of Physical Anthropology*, 29:1–43, 1986.

[4] Elaine Y. Y. Cheung, Michelle Elizabeth Gahan, and Dennis McNevin. Prediction of biogeographical ancestry in admixed individuals. *Forensic Science International. Genetics*, 36:104–111, 2018.

[5] M. Eduardoff, T. E. Gross, C. Santos, M. de la Puente, D. Ballard, C. Strobl, C. Børsting, N. Morling, L. Fusco, C. Hussing, B. Egyed, L. Souto, J. Uacyisrael, D-Syndercombe Court, A Carracedo, M. V. Lareu, P. M. Schneider, W. Parson, C. Phillips, EUROFORGEN-NoE Consortium, W. Parson, and C. Phillips. Inter-laboratory evaluation of the EUROFORGEN Global ancestry-informative SNP panel by massively parallel sequencing using the Ion PGM™. *Forensic Science International. Genetics*, 23:178–189, 2016.

[6] T. Fawcett. An introduction to ROC analysis. *Pattern Recognition Letters*, 27:861–874, 2006.

[7] C. L. Hanis, R. Chakraborty, R. E. Ferrell, and W. J. Schull. Individual Admixture Estimates: Disease Associations and Individual Risk of Diabetes and Gallbladder Disease Among Mexican-Americans in Starr County, Texas. *American Journal of Physical Anthropology*, 70:433–441, 1986.

[8] Jerome Kelleher, Alison M. Etheridge, and Gilean McVean. Efficient coalescent simulation and genealogical analysis for large sample sizes. *PLoS Computational Biology*, 12(5):e1004842, 2016.

[9] K. Kidd, U. Soundararajana, H. Rajeevana, A. J. Pakstisa, K. N. Moorec, and J. D. Ropero-Millerc. The redesigned forensic research/reference on genetics-knowledge base, frog-kb. *Forensic Science International. Genetics*, 33:33–37, 2017.

[10] Kenneth K. Kidd, William C. Speed, Andrew J. Pakstis, Manohar R. Furtado, Rixun Fang, Abeer Madbouly, Martin Maiers, Mridu Middha, Françoise R. Friedlaender, and Judith R. Kidd. Progress toward an efficient panel of SNPs for ancestry inference. *Forensic Science International. Genetics*, 10:23–32, 2014.

[11] D. McNevin. Forensic inference of biogeographical ancestry from genotype: The genetic ancestry lab. *WIREs Forensic Science*, e1356:1–26, 2019.

[12] P. Pfaffelhuber, F. Grundner-Culemann, V. Lipphardt, and F. Baumdicker. How to choose sets of ancestry informative markers: A supervised feature selection approach. *Forensic Science International. Genetics*, page minor revision, 2019.

[13] C. Phillips, W. Parson, B. Lundsberg, C. Santos, A. Freire-Aradas, M. Torres, M. Eduardoff, C. Børsting, P. Johansen, M. Fondevila, N. Morling, P. Schneider, EUROFORGEN-NoE Consortium, A. Carracedo, and M. V. Lareu. Building a forensic ancestry panel from the ground up: The EUROFORGEN Global AIM-SNP set. *Forensic Science International. Genetics*, 11:13–25, 2014.

[14] C. Phillips, A. Salas, J. J. Sánchez, M. Fondevila, A. Gómez-Tato, J. Álvarez Dios, M. Calaza, M. Casares de Cal, D. Ballard, M. V. Lareu, A.. Carracedo, and The SNPforID Consortium. Inferring ancestral origin using a single multiplex assay of ancestry-informative marker SNPs. *Forensic Science International. Genetics*, 1:273–280, 2007.

[15] Chris Phillips, Carla Santos, Manuel Fondevila, Ángel Carracedo, and Maria Victoria Lareu. Inference of Ancestry in Forensic Analysis I: Autosomal Ancestry-Informative Marker Sets. In *Forensic DNA Typing Protocols*, volume 1420 of *Methods in Molecular Biology*, pages 233–253. Springer, New York, 2016.

[16] P. R. Prestes, R. J. Mitchell, R. Daniel, J. J. Sanchez, and R. A.H. van Oorschot. Predicting biogeographical ancestry in admixed individuals – values and limitations of using uniparental and autosomal markers. *Australian Journal of Forensic Sciences*, 48:10–23, 2015.

[17] J. Pritchard, M. Stephens, and P. Donnelly. Inference of population structure using multilocus genotype data. *Genetics*, 155:945–954, 2000.

[18] H. Tang, J. Peng, P. Wang, and N. Risch. Estimation of individual admixture: Analytical and study design considerations. *Genet Epidemiol.*, 28:289–301, 2005.

[19] T. Tvedebrink, P. S. Eriksen, H. S. Mogensen, and N. Morling. Genogeographer – a tool for genogeographic inference. *Forensic Science International: Genetics Supplement Series*, 6:e463–e465, 2018a.

[20] S. Wahlund. Zusammensetzung von populationen und korrelationserscheinungen vom standpunkt der vererbungslehre aus betrachtet. *Hereditas*, 11:65–106, 1928.

Supporting Information: Inference of recent admixture using genotype data

PETER PFAFFELHUBER, ELISABETH HUSS, FRANZ BAUMDICKER,
DENISE SYNDERCOMBE COURT, FABIAN STAUBACH

December 2, 2019

S1 Theory

We write down the admixture model and derive a method to estimate Individual Admixture (IA) in the case when allele frequencies within populations are not updated. In addition, we give the recent-admixture model, where an individual is allowed to have parents with different admixture proportions. We take the following notation for the reference database:

K : number of ancestral populations,

M : number of markers,

p_{mk} : frequency of allele 1 at (bi-allelic) marker m in population k.

In addition, we consider one additional diploid genome (called the trace) $(G_{m1}, G_{m2})_{m=1,\dots,M}$, or $(G_m)_{m=1,\dots,M}$ with $G_m = G_{m1} + G_{m2}$ if phase is not known. The goal is to estimate admixture proportions $(q_k)_{k=1,\dots,K}$ (or $(q_k^M)_{k=1,\dots,K}, (q_k^P)_{k=1,\dots,K}$) of this additional genome. Recall that we take the approach from [2] and [3] and do not update p_{mk}'s during the analysis.

S1.1 The admixture model

In [1], [4], [5] and elsewhere, the main goal is to maximize the log-likelihood (see also (1) and (2) of [5])

$$\ell(q|G) = \sum_{m=1}^{M} \sum_{a=1,2} \log\Big(\sum_{k=1}^{K} \alpha_{mak} q_k\Big),$$

where

$$\alpha_{mak} := \begin{cases} p_{mk}, & \text{if } G_{ma} = 1, \\ 1 - p_{mk}, & \text{if } G_{ma} = 0. \end{cases}$$

is the frequency of the observed allele in copy a of marker m in population k. (Note that $e^{\ell(q|G)}$ is the probabiltiy of observing $(G_{ma})_{m=1,\dots,M;a=1,2}$, if every allele is picked independently from population $k$with probability q_k.) Assuming that phase is not known, and with

$$\alpha_{mkl} := \alpha_{m1k}\alpha_{m2l} = \begin{cases} p_{mk}p_{ml}, & \text{if } G_{m1} + G_{m2} = 2, \\ p_{mk}(1-p_{ml}) + (1-p_{mk})p_{ml}, & \text{if } G_{m1} + G_{m2} = 1, \\ (1-p_{mk})(1-p_{ml}), & \text{if } G_{m1} + G_{m2} = 0, \end{cases} \tag{S1}$$

note that the log-likelihood can as well be written as

$$\ell(q|G) = \sum_{m=1}^{M} \log\Big(\sum_{k,l=1}^{K} \alpha_{mkl} q_k q_l \Big).$$

We set $\beta_m(q) := \sum_{k=1}^{K} p_{mk} q_k$ and analyse the last sum by distinguishing the case $G_m = 2$, where

$$\sum_{k,l=1}^{K} \alpha_{mkl} q_k q_l = \Big(\sum_{k=1}^{k} p_{mk} q_k \Big)^2 = \beta_m(q)^2$$

while for $G_m = 1$ we find

$$\sum_{k,l=1}^{K} \alpha_{mkl} q_k q_l = 2\Big(\sum_{k=1}^{k} p_{mk} q_k \Big)\Big(1 - \sum_{k=1}^{k} p_{mk} q_k \Big) = 2\beta_m(q)(1 - \beta_m(q))$$

and for $G_m = 0$

$$\sum_{k,l=1}^{K} \alpha_{mkl} q_k q_l = \Big(1 - \sum_{k=1}^{k} p_{mk} q_k \Big)^2 = (1 - \beta_m(q))^2,$$

such that

$$\ell(q|G) = \sum_{m=1}^{M} \log\Big(\binom{2}{G_m} \beta_m(q)^{G_m} (1 - \beta_m(q))^{2-G_m} \Big) \tag{S2}$$

Lemma S1.1. *The maximum of $q \mapsto \ell(q|G)$ under the constraint $\sum_{k=1}^{K} q_k = 1$ solves*

$$\frac{1}{M} \sum_{m=1}^{M} \sum_{l=1}^{K} \frac{\alpha_{mkl} q_l}{\sum_{k',l'=1}^{K} \alpha_{mk'l'} q_{k'} q_{l'}} = 1, \qquad k = 1, ..., K. \tag{S3}$$

Remark S1.1. 1. Assume we add a set of *Ancestry Uninformative Markers*, i.e. a set of markers $\tilde{m} = 1, ..., \tilde{M}$, with frequencies not depending on the population, i.e. $\alpha_{\tilde{m}kl} = \alpha_{\tilde{m}k'l'}$ for $k, l, k, l'' = 1, ..., K$. For these markers,

$$\sum_{\tilde{m}=1}^{\tilde{M}} \sum_{l=1}^{K} \frac{\alpha_{\tilde{m}kl} q_l}{\sum_{k',l'=1}^{K} \alpha_{\tilde{m}k'l'} q_{k'} q_{l'}} = \tilde{M}.$$

This implies that q is a solution of (S3) without these markers iff q is a solution of (S3) if these markers are included. This might be reassuring.

2. Let us have a closer look at the left hand side of (S3). We set $\beta_m := \sum_{k=1}^{K} p_{mk} q_k$. For $G_m = 2$,

$$\sum_{l=1}^{K} \frac{\alpha_{mkl} q_l}{\sum_{k',l'} \alpha_{mk'l'} q_{k'} q_{l'}} = \sum_{l=1}^{K} \frac{p_{mk} p_{ml} q_l}{\sum_{k',l'} p_{mk'} p_{ml'} q_{k'} q_{l'}} = \frac{p_{mk}}{\beta_m}.$$

For $G_m = 1$, we have

$$\sum_{l=1}^{K} \frac{\alpha_{mkl} q_l}{\sum_{k',l'} \alpha_{mk'l'} q_{k'} q_{l'}} = \sum_{l=1}^{K} \frac{(p_{mk}(1-p_{ml}) + (1-p_{mk})p_{ml})q_l}{\sum_{k',l'}(p_{mk'}(1-p_{ml'}) + (1-p_{mk'})p_{ml'})q_{k'}q_{l'}}$$
$$= \frac{p_{mk}(1-\beta_m) + (1-p_{mk})\beta_m}{2\beta_m(1-\beta_m)} = \frac{1}{2}\Big(\frac{p_{mk}}{\beta_m} + \frac{1-p_{mk}}{1-\beta_m}\Big)$$

and for $G_m = 0$

$$\sum_{l=1}^{K} \frac{\alpha_{mkl} q_l}{\sum_{k',l'} \alpha_{mk'l'} q_{k'} q_{l'}} = \sum_{l=1}^{K} \frac{(1-p_{mk})(1-p_{ml})q_l}{\sum_{k',l'}(1-p_{mk'})(1-p_{ml'})q_{k'}q_{l'}} = \frac{1-p_{mk}}{1-\beta_m}.$$

In total, this gives that q needs to solve

$$\frac{1}{2M}\sum_{m=1}^{M}\Big(G_m\frac{p_{mk}}{\beta_m} + (2-G_m)\frac{1-p_{mk}}{1-\beta_m}\Big) = 1$$

In order to find a solution, we reformulate as the fixed point equation

$$q_k = f_k(q) \text{ for } f_k(q) = \frac{1}{2M}\sum_{m=1}^{M}\Big(G_m\frac{p_{mk}}{\beta_m} + (2-G_m)\frac{1-p_{mk}}{1-\beta_m}\Big)q_k, \qquad k = 1, ..., K. \quad \text{(S4)}$$

A solution can be computed by iteratively computing $q' = (f_k(q))_{k=1,...,K}$.

Proof of Lemma S1.1. We use the theory of Lagrange multipliers, since we need to maximize ℓ over q under the constraint $\sum_k q_k = 1$. Since

$$\frac{\partial \ell(q|G)}{\partial q_k} = \sum_{m=1}^{M}\sum_{l=1}^{K} \frac{\alpha_{mkl} q_l}{\sum_{k',l'=1}^{K} \alpha_{mk'l'} q_{k'} q_{l'}},$$

we have to solve the system of equations

$$\lambda = \sum_{m=1}^{M}\sum_{l=1}^{K} \frac{\alpha_{mkl} q_l}{\sum_{k',l'=1}^{K} \alpha_{mk'l'} q_{k'} q_{l'}}, \qquad k = 1, ..., K,$$
$$1 = \sum_{k=1}^{K} q_k.$$

It is easy to eliminate λ, since the last equation gives

$$\lambda = \lambda\sum_{k=1}^{K} q_k = \sum_{m=1}^{M}\sum_{k,l=1}^{K} \frac{\alpha_{mkl} q_k q_l}{\sum_{k',l'=1}^{K} \alpha_{mk'l'} q_{k'} q_{l'}} = M.$$

So, we are left with finding $q = (q_k)_{k=1,...,K}$ such that

$$\frac{1}{M}\sum_{m=1}^{M}\sum_{l=1}^{K} \frac{\alpha_{mkl} q_l}{\sum_{k',l'=1}^{K} \alpha_{mk'l'} q_{k'} q_{l'}} = 1, \qquad k = 1, ..., K.$$

□

S1.2 The recent-admixture model

For the recent-admixture-version, we want to estimate $q^M = (q_k^M)_{k=1,...,K}, q^P = (q_k^P)_{k=1,...,K}$, where q_k^M and q_k^P are the fractions of the genomes of the mother and father, respectively, which come from population k. This assumption implies that the log-likelihood is

$$\ell(q^M, q^P|G) = \sum_{m=1}^{M} \log\Big(\sum_{k,l=1}^{K} \alpha_{mkl} q_k^M q_l^P\Big),$$

where α_{mkl} is given as in (S1). As is apparent from the log-Likelihood function, the recent-admixture model generalizes the admixture model. Put differently, choosing $q^M = q^P = q$ in the recent-admixture model gives the admixture model. We note that again, the log-likelihood can be written differently,

$$\begin{aligned}\ell(q^M, q^P|G) = \sum_{m=1}^{M} \log\Big(&1_{G_m=2}\beta_m(q^M)\beta_m(q^P) \\ &+ 1_{G_m=1}(\beta_m(q^M)(1-\beta_m(q^P)) + (1-\beta_m(q^M))\beta_m(q^P)) \\ &+ 1_{G_m=0}(1-\beta_m(q^M))(1-\beta_m(q^P))\Big).\end{aligned} \tag{S5}$$

Lemma S1.2. *The maximum of* $(q^M, q^P) \mapsto \ell(q^M, q^P, G)$ *under the constraint* $\sum_{k=1}^K q_k^M = \sum_{k=1}^K q_k^P = 1$ *solves*

$$\frac{1}{M}\sum_{m=1}^{M}\sum_{l=1}^{K} \frac{\alpha_{mkl} q_l^M}{\sum_{k',l'=1}^K \alpha_{mk'l'} q_{k'}^M q_{l'}^P} = \frac{1}{M}\sum_{m=1}^{M}\sum_{l=1}^{K} \frac{\alpha_{mkl} q_l^P}{\sum_{k',l'=1}^K \alpha_{mk'l'} q_{k'}^P q_{l'}^M} = 1, \qquad k = 1, ..., K. \tag{S6}$$

Remark S1.2.

1. Note that (S6) is symmetric in q^M and q^P, i.e. if (q^M, q^P) solve (S6), another solution is given by (q^P, q^M).

2. Again, we can turn (S6) into fixed point equations. To derive it, we again have a closer look at the left hand side of (S6). For $\beta_m(q) := \sum_k p_{mk} q_k$, we have for $G_m = 2$

$$\sum_{l=1}^{K} \frac{\alpha_{mkl} q_l^M}{\sum_{k',l'=1}^K \alpha_{mk'l'} q_{k'}^M q_{l'}^P} = \frac{p_{mk}\beta_m(q^M)}{\beta_m(q^M)\beta_m(q^P)} = \frac{p_{mk}}{\beta_m(q^P)},$$

for $G_m = 1$

$$\sum_{l=1}^{K} \frac{\alpha_{mkl} q_l^M}{\sum_{k',l'=1}^K \alpha_{mk'l'} q_{k'}^M q_{l'}^P} = \frac{p_{mk}(1-\beta_m(q^M)) + (1-p_{mk})\beta_m(q^M)}{\beta_m(q^M)(1-\beta_m(q^P)) + (1-\beta_m(q^M))\beta_m(q^P)}$$

and for $G_m = 0$

$$\sum_{l=1}^{K} \frac{\alpha_{mkl} q_l^M}{\sum_{k',l'=1}^K \alpha_{mk'l'} q_{k'}^M q_{l'}^P} = \frac{(1-p_{mk})(1-\beta_m(q^M))}{(1-\beta_m(q^M))(1-\beta_m(q^P))} = \frac{1-p_{mk}}{1-\beta_m(q^P)}.$$

So, we suggest to iteratively compute

$$\tilde{q}^P = f(q^M, q^P) \text{ and } \tilde{q}^M = f(\tilde{q}^P, q^M) \tag{S7}$$

for $f(q, q') = (f_k(q, q'))_{k=1,...,K}$ with

$$f_k(q, q') := \frac{1}{M} \sum_{m=1}^{M} \Big(1_{G_m=2} \frac{p_{mk}}{\beta_m(q')} + 1_{G_m=1} \frac{(p_{mk}(1-\beta_m(q)) + (1-p_{mk})\beta_m(q))}{\beta_m(q)(1-\beta_m(q')) + (1-\beta_m(q))\beta_m(q')} + 1_{G_m=0} \frac{(1-p_{mk})}{1-\beta_m(q')} \Big) q'_k.$$

Proof of Lemma S1.2. Again, we use Lagrange multipliers. Since

$$\frac{\partial \ell(q^M, q^P|G)}{\partial q_k^P} = \sum_{m=1}^{M} \sum_{l} \frac{\alpha_{mak} q_l^M}{\sum_{k',l'} \alpha_{mk'l'} q_{k'}^P q_{k'}^M},$$
$$\frac{\partial \ell(q^M, q^P|G)}{\partial q_k^M} = \sum_{m=1}^{M} \sum_{l} \frac{\alpha_{mak} q_l^P}{\sum_{k',l'} \alpha_{mk'l'} q_{k'}^P q_{k'}^M},$$

we have to solve the system of equations

$$\lambda = \sum_{m=1}^{M} \sum_{l} \frac{\alpha_{mak} q_l^M}{\sum_{k',l'} \alpha_{mk'l'} q_{k'}^P q_{k'}^M}, \qquad k = 1, ..., K,$$
$$\rho = \sum_{m=1}^{M} \sum_{l} \frac{\alpha_{mak} q_l^P}{\sum_{k',l'} \alpha_{mk'l'} q_{k'}^P q_{k'}^M}, \qquad k = 1, ..., K,$$
$$1 = \sum_{k=1}^{K} q_k^P = \sum_{k=1}^{K} q_k^M.$$

It is easy to eliminate λ and ρ, since

$$\lambda = \lambda \sum_{k=1}^{K} q_k^P = \sum_{m=1}^{M} \sum_{k,l} \frac{\alpha_{mkl} q_k^P q_l^M}{\sum_{k',l'} \alpha_{mk'l'} q_{k'}^P q_{l'}^M} = M,$$
$$\rho = \rho \sum_{k=1}^{K} q_k^M = \sum_{m=1}^{M} \sum_{k,l} \frac{\alpha_{mkl} q_k^P q_l^M}{\sum_{k',l'} \alpha_{mk'l'} q_{k'}^P q_{l'}^M} = M.$$

So, we are left with finding q^P and q^M such that

$$\frac{1}{M} \sum_{m=1}^{M} \frac{\alpha_{mkl} q_l^P}{\sum_{k',l'}^{K} \alpha_{mk'l'} q_{k'}^P q_{l'}^M} = 1, \qquad k = 1, ..., K,$$
$$\frac{1}{M} \sum_{m=1}^{M} \frac{\alpha_{mkl} q_l^M}{\sum_{k',l'}^{K} \alpha_{mk'l'} q_{k'}^P q_{l'}^M} = 1, \qquad k = 1, ..., K.$$

□

S2 Additional results

S2.1 Some showcases from simulations

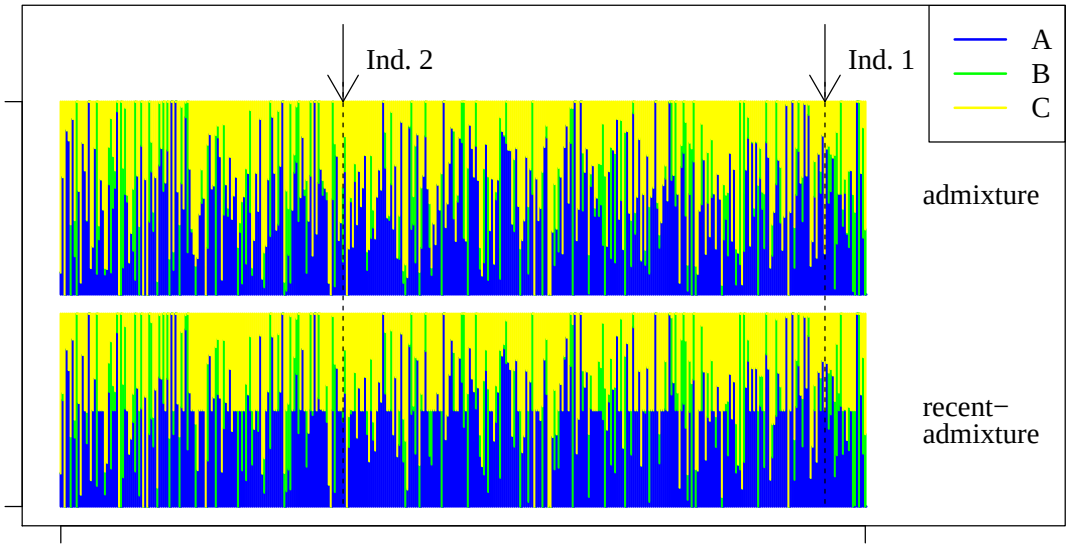


Figure S1: Structure barplot for all 400 recently $A \times C$-admixed individuals. Arrows indicate the two individuals from Figure 1.

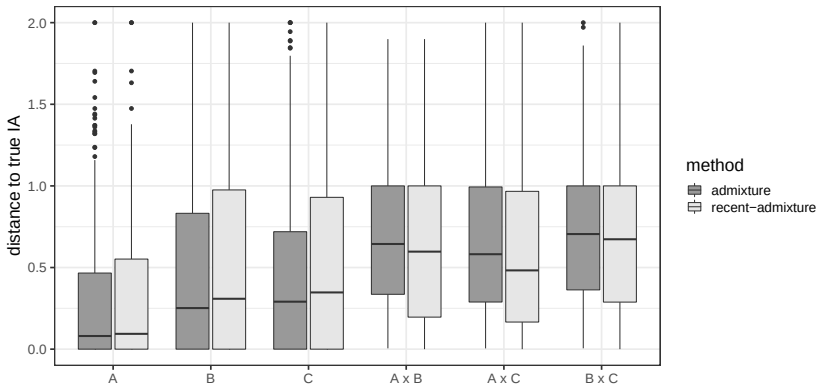


Figure S2: For all 3×400 non-admixed samples, and 3×400 admixed samples, we compute the errors by the distance to the true IA as given in (6).

S2.2 Estimation accuracy

For first generation admixed individuals, all cases (non-admixed and parents come from different continents) are described in the main text. For second generation admixed individuals, we have several cases, depending on the origin of the grand-parents. When data from four populations (AFR, EAS, EUR, SAS; see below) is available, we have the following cases:

(A) 4 non-admixed cases with IA 100:0:
AFR/AFR×AFR/AFR, EAS/EAS×EAS/EAS, EUR/EUR×EUR/EUR, SAS/SAS×SAS/SAS;

(B) 6 admixed cases with admixture ratio 50:50, where both parents are non-admixed:
AFR/AFR×EAS/EAS, AFR/AFR×EUR/EUR, AFR/AFR×SAS/SAS, EAS/EAS×EUR/EUR, EAS/EAS×SAS/SAS, EUR/EUR×SAS/SAS;

(C) 6 admixed cases with admixture ratio 50:50, where both parents are admixed:
AFR/EAS×AFR/EAS, AFR/EUR×AFR/EUR, AFR/SAS×AFR/SAS, EAS/EUR×EAS/EUR, EAS/SAS×EAS/SAS, EUR/SAS×EUR/SAS;

(D) 12 admixed cases with admixture ratio 75:25:
AFR/AFR×AFR/EAS, AFR/AFR×AFR/EUR, AFR/AFR×AFR/SAS, EAS/EAS×EAS/AFR, EAS/EAS×EAS/EUR, EAS/EAS×EAS/SAS, EUR/EUR×EUR/AFR, EUR/EUR×EUR/EAS, EUR/EUR×EUR/SAS, SAS/SAS×SAS/AFR, SAS/SAS×SAS/EAS, SAS/SAS×SAS/EUR;

(E) 12 second generation admixed with admixture ratio 50:25:25, where one parent is non-admixed:
AFR/AFR×EAS/EUR, AFR/AFR×EAS/SAS, AFR/AFR×EUR/SAS, EAS/EAS×AFR/EUR, EAS/EAS×AFR/SAS, EAS/EAS×EUR/SAS, EUR/EUR×AFR/EAS, EUR/EUR×AFR/SAS, EUR/EUR×EAS/SAS, SAS/SAS×AFR/EAS, SAS/SAS×AFR/EUR, SAS/SAS×EAS/EUR;

(F) 12 second generation admixed with admixture ratio 50:25:25, where both parents are admixed:
AFR/EAS×AFR/EUR, AFR/EAS×AFR/SAS, AFR/EUR×AFR/SAS, EAS/AFR×EAS/EUR, EAS/AFR×EAS/SAS, EAS/EUR×EAS/SAS, EUR/AFR×EUR/EAS, EUR/AFR×EUR/SAS, EUR/EAS×EUR/SAS, SAS/AFR×SAS/EAS, SAS/AFR×SAS/EUR, SAS/EAS×SAS/EUR;

(G) 3 second generation admixed with admixture ratio 25:25:25:25:
AFR/EAS×EUR/SAS, AFR/EUR×EAS/SAS, AFR/SAS×EAS/EUR;

As can be seen in Figure S4, the recent-admixture model never gives worse estimates than the admixture model, and outperforms the admixture-model in several cases; see also Figure 2 in the main text. For the Kidd AIMset, see Figures S6 and S7.

EUROFORGEN AIMset

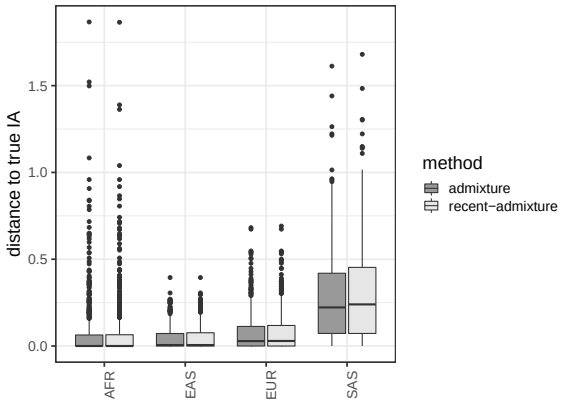


Figure S3: For all non-admixed samples from the 1000 genomes dataset, we computed IA from the admixture and recent-admixture model. The distance to the true IA is computed as in (6).

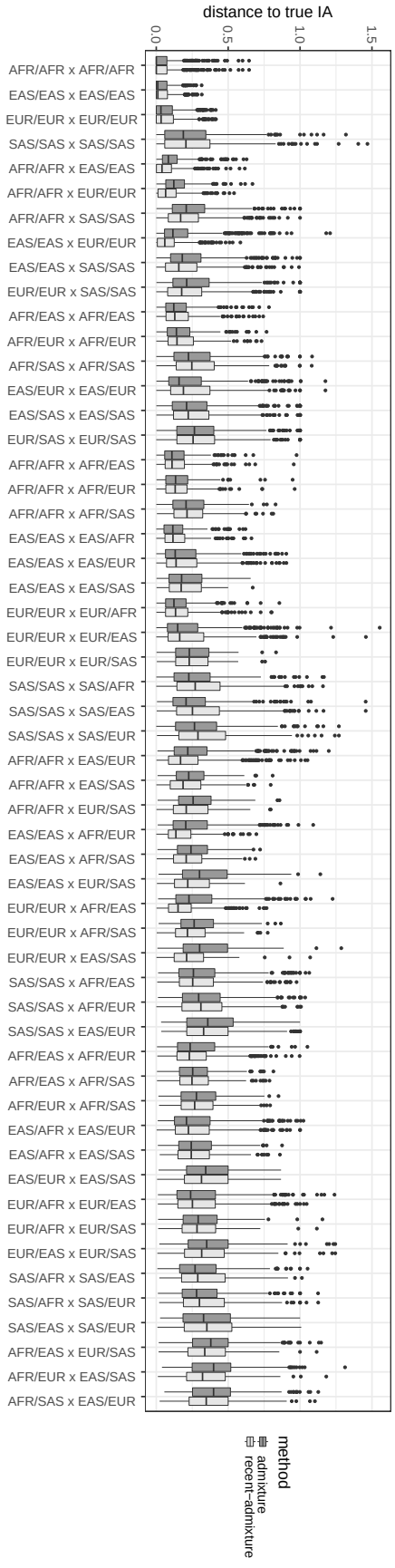


Figure S4: For all cases of second generation admixed individuals, we compare estimation accuracy of IA using the EUROFORGEN AIMset.

Kidd AIMset

For the Kidd AIMset, recent-admixture produces an error in the non-admixed samples (see also Figure S5), which is smaller than the error for the admixture-model in 1394 out of 2157 cases, so the hypothesis that the error is worse for recent-admixture can be rejected ($p < 0.001$). For first order admixed samples, 2381 out of 3000 cases have a smaller error under recent-admixture (which gives $p < 0.001$; see also Figure S6.A. Last, for second-generation admixed individuals, recent-admixture is more accurate in 16116 out of 27500 cases ($p < 0.001$); see also Figures S6.B and S7.

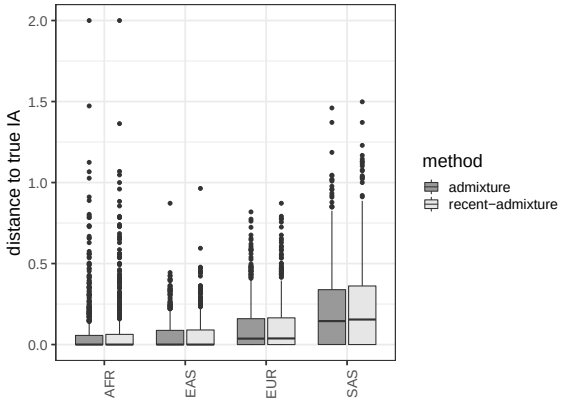


Figure S5: Same as in Figure S3, but for the Kidd AIMset.

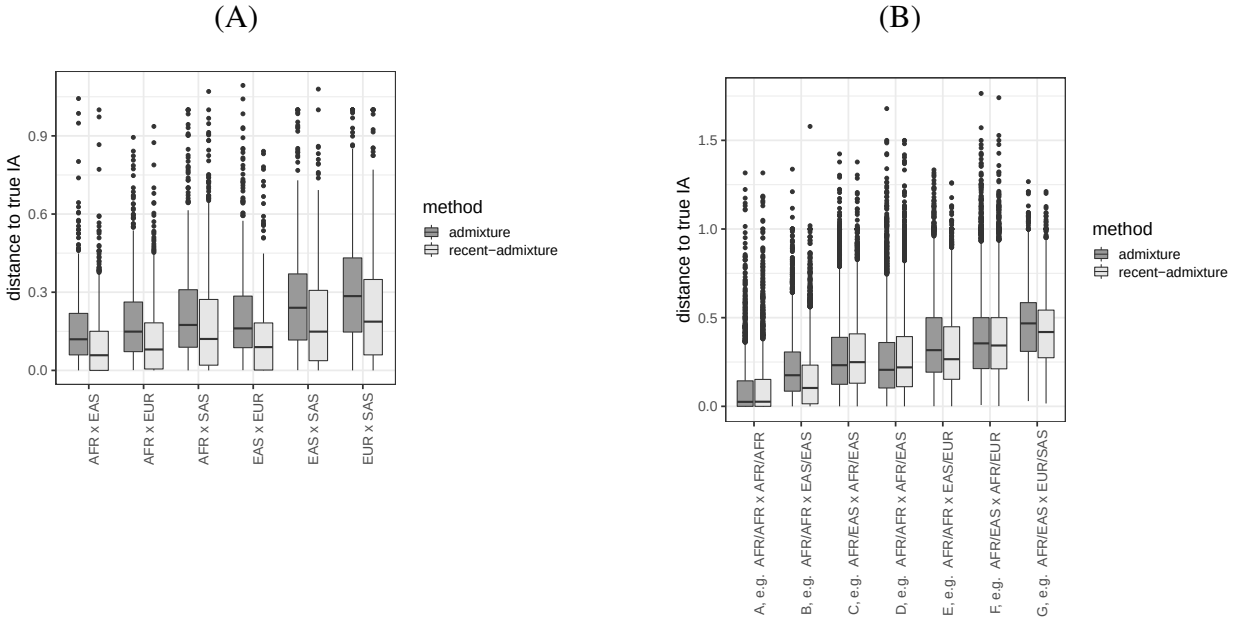


Figure S6: Same as in Figure 2, but using the Kidd AIMset.

Case	B	C	D	E	F	G
AUC	0.982	0.655	0.855	0.983	0.921	0.982
Power at $p = 0.01$	0.92	0.29	0.5	0.89	0.65	0.89

Table S1: Same as Table 1, but using the Kidd AIMset.

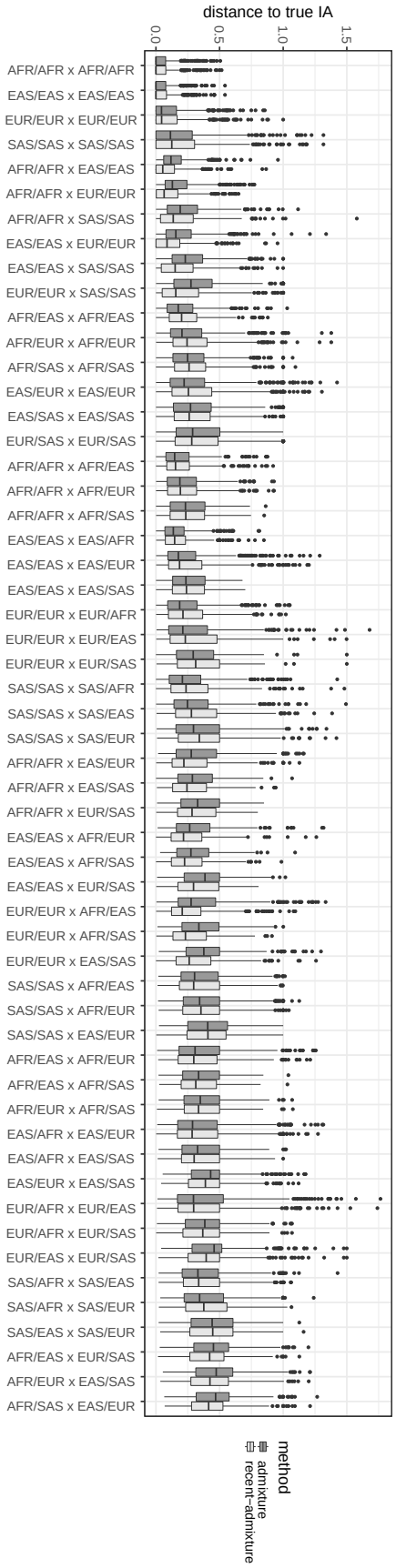


Figure S7: Same as in Figure S4, but using the Kidd AIMset.

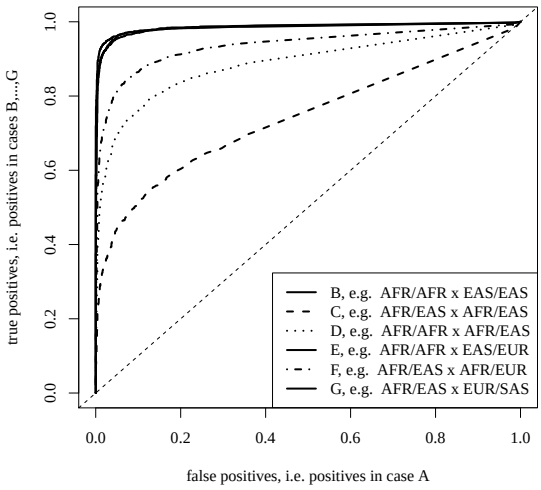


Figure S8: Same as in Figure 3, but using the Kidd AIMset.

References

[1] D. H. Alexander, J. Novembre, and K. Lange. Fast model-based estimation of ancestry in unrelated individuals. *Genome Research*, 19:1655–1664, 2009.

[2] R. Chakraborty. Gene Admixture in Human Populations: Models and Predictions. *Yearbook of Physical Anthropology*, 29:1–43, 1986.

[3] C. L. Hanis, R. Chakraborty, R. E. Ferrell, and W. J. Schull. Individual Admixture Estimates: Disease Associations and Individual Risk of Diabetes and Gallbladder Disease Among Mexican-Americans in Starr County, Texas. *American Journal of Physical Anthropology*, 70:433–441, 1986.

[4] J. Pritchard, M. Stephens, and P. Donnelly. Inference of population structure using multilocus genotype data. *Genetics*, 155:945–954, 2000.

[5] H. Tang, J. Peng, P. Wang, and N. Risch. Estimation of individual admixture: Analytical and study design considerations. *Genet Epidemiol.*, 28:289–301, 2005.