Inference of recent admixture using genotype data

PETER PFAFFELHUBER, ELISABETH SESTER-HUSS, FRANZ BAUMDICKER, JANA NAUE, SABINE LUTZ-BONENGEL, FABIAN STAUBACH

September 16, 2020

Abstract

The inference of biogeographic ancestry (BGA) has become a focus of forensic genetics. Misinference of BGA can have profound unwanted consequences for investigations and society. We show that recent admixture can lead to misclassification and erroneous inference of ancestry proportions, using state of the art analysis tools with (i) simulations, (ii) 1000 genomes project data, and (iii) two individuals analyzed using the ForenSeq DNA Signature Prep Kit. Subsequently, we extend existing tools for estimation of individual ancestry (IA) by allowing for different IA in both parents, leading to estimates of parental individual ancestry (PIA), and a statistical test for recent admixture. Estimation of PIA outperforms IA in most scenarios of recent admixture. Furthermore, additional information about parental ancestry can be acquired with PIA that may guide casework.

1 Introduction

Inference of the biogeographical ancestry of a trace or an unknown body, using genetic markers, is a focus of recent forensic genetics research (see e.g. [1, 2, 3]). Misclassification and erroneous inference of ancestry proportions can mislead casework and result in unwanted societal consequences [4]. Therefore, discovering potential error sources and improving methods for BGA is essential for successful and responsible application of the technology.

For inference of BGA, either the trace is classified into one of several groups of different origin (e.g. Africa, Europe, East Asia, Native America, South-East Asia and Oceania; see e.g. [5, 6]), or it is assumed that it consists of a mixture of ancestral genetic material originating in several groups. For the inference of such admixture proportions of individual ancestry (IA), STRUCTURE [4] and the faster ADMIXTURE have become the de-facto standards [3]; see also [5] for the same model.

To understand how recent admixture can cause errors in BGA inference, consider a recently admixed individual, i.e. the continental BGA of both parents differs. Since methods used in BGA classification [5, 6] can only result in single population/class label, the possibility of recent admixture is usually not even implemented by such methods, making results on classification of recently admixed individuals hard to interpret. In mixed membership models as implemented in STRUCTURE

or ADMIXTURE for estimating continental (or other scales of geography) IA of a trace, the genome is thought to be a mosaic of stones of different (continental) origin. The geographic distribution of the mosaic stones (alleles) is given by the IA. Here, a main assumption is that all alleles have the same chance (given by the IA) to stem from one of the continents. However, if the BGA of the two parents differs, the chance to encounter two different alleles at a locus increases due to population differentiation, e.g. between continents [10]. This is expected to lead to a genome wide increase in heterozygosity. The violation of the assumption of equal chances for homologous allelic states could potentially lead to misinferences. Given that recent admixture is a common issue in the light of increased human mobility, currently and in the past decades, such misinference could be a common error source in forensic genetic analyses. Therefore, a more comprehensive understanding of the consequences and frequency of recent admixture in forensic analyses is needed.

For forensic applications the approaches taken by Zou et al. (2015) [11] and Pei et al. (2020) [12] to infer recent admixture are often not feasible because they rely on the inference of phase along the chromosome with dense marker sets. Furthermore, these approaches are computationally demanding. Crouch and Weale (2012) developed the LEAPFrOG algorithm that can be used with the limited marker density of most forensic applications [13]. These authors inferred the parental IA with a maximum likelihood approach and applied their method to simulated and forensically relevant datasets with a focus on European/African admixture. To identify recently admixed individuals in forensic samples, an excess of heterozygous sites was used in statistical tests [14]. Tvedebrink et al. (2018) and Tvedebrink and Eriksen (2019) developed likelihood-ratio tests for the null-hypothesis of non-admixture and recent (first generation) admixture, respectively, versus the alternative that the studied sample is not represented in the reference database [15, 15].

What is currently missing is an approach to test for recent admixture in a trace, where the null-hypothesis is that both parents have the same IA. The alternative hypothesis, called the recent-admixture model below, would be that the studied sample has parents of different IA and therefore shows recent-admixture of populations within the reference database. If such an approach also identified the IAs of the parents of an admixed individual, this might inform casework. Moreover, for a better understanding of the potential misinference in forensic applications, simulations should be based on the most realistic human population genetic models and include the analysis of recent methods and marker sets. Given global mobility of humans, global sample collections should be included in the analyses.

Our goals were to identify and quantify potential errors that result from recent admixture in standard methods for BGA inference and to develop a statistical test for recent admixture vs the null hypothesis of non-admixture. To this end, we developed a method for the inference of admixture proportions of both parents (parental individual ancestry, PIA). We use this method to (i) assign IA to the parents, (ii) improve current methods of ancestry inference, (iii) perform a likelihood ratio test to identify recently admixed individuals. For assessing our method and also to assess the misinference of BGA with standard methods in the context of recent admixture, we leverage population genetic simulations including the most realistic human population genetic scenarios. Furthermore, we leverage a global dataset from the 1000 genomes project, and two samples analysed using the ForenSeq Signature Prep Kit on a MiSeq FGx (Verogen).

2 Materials and Methods

We start by briefly recalling the admixture model, which is the basis for the widely used sofware STRUCTURE [4], ADMIXTURE [3] and FRAPPE [5]. Afterwards, we introduce a new model, called the recent-admixture model. More details on the derivations in the admixture and recent-admixture 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 admixture and recent-admixture model, we assume to have a reference database of M bi-allelic markers from K populations. However, from this reference database, we only need to know 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. Instead, we take the allele frequencies as given in the reference database. This approach has also been used in early papers such as [1] and [2].

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)

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

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)

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

This can be done numerically by iterating $q_{n+1} = (f_k(q_n))_{k=1,...,K}$ 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 allele frequencies. In addition, although maximizing (2) could also

be handled using a Newton method as in [3], 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 in silico recently admixed individuals from a reference database. 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, $(G_m)_{m=1,\dots,M}$ are independent with $G_m = X_m + Y_m$, where $X_m = 1$ with probability $\tilde{G}_m/2$, $X_m = 0$ with probability $1 - \tilde{G}_m/2$ and $Y_m = 1$ with probability $\bar{G}_m/2$, $Y_m = 0$ with probability $1 - \bar{G}_m/2$. When iterating this procedure, we can also model second-generation admixed individuals etc. in silico.

Consider the subset of the 1000 genomes dataset consisting of Africans (AFR), East-Asians (EAS), Europeans (EUR) and South-East-Asians (SAS). All cases for second generation admixed individuals fall into one of seven categories. Writing up the ancestries of the four grandparents *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 [18] 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 cases with admixture ratio 25:25:25:25, e.g. AFR/EAS×EUR/SAS;

For each of these 55 cases, we simulated 500 individuals in silico by picking four grandparents at random from the populations, creating mother and father from the grandparents, 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 use deviations from $q^{\text{TRUE}} = (q_k^{\text{TRUE}})_{k=1,\ldots,K}$, where $q_k^{\text{TRUE}} = 1$ for a non-admixed individual from population k, $q_k^{\text{TRUE}} = q_{k'}^{\text{TRUE}} = 0.5$ for an admixed individual with parents from populations k and k', etc. We then use the estimation error for the admixture model

$$\sum_k |q_k - q_k^{\text{TRUE}}| \text{ and } \sum_k |\tfrac{1}{2}(q_k^M + q_k^P) - q_k^{\text{TRUE}}| \tag{6}$$

for the recent-admixture model, respectively. 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 $\frac{1}{2}(q^M + q^P)$ is contained in the estimates for this model.

2.5 Likelihood ratios for recent admixture

We want to see if data $G = (G_m)_{m=1,\dots,M}$ 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 $q^M = q^P = q$, this amounts to a likelihood ratio test of $H_0 : q^M = q^P$ 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). As usual in likelihood ratio tests, 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. The downside here is that we do not know the distribution of $\Delta\ell$ under H_0 and therefore cannot translate the observed value for $\Delta\ell$ to a p-value. Therefore, we only report the $\Delta\ell$-value.

In order to get more insight into $\Delta\ell$, recall that we assume that AIMs segregate independently. As a consequence, both $\ell(\hat{q}|G)$ from (2) and $\ell(\hat{q}^M, \hat{q}^P|G)$ from (4) are sums over all M loci, such that we can report the contribution of every AIM to $\Delta\ell$, as well as contributions of subsets of AIMs (e.g. heterozygote and homozygote sites).

2.6 Data from the 1000 genomes project

In order to detect recent admixture in publicly available 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` [19]. 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 [20] and Kidd AIMset [21], 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 [6]. 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.

The analysis of this dataset relies on allele frequencies used to estimate IA and PIA. Here, we use the samples of AFR, EAS, EUR and SAS. We did not use AMR since they are known to be admixed.

2.7 AIMs analysis of two collected individuals with recent admixture

Within a larger study about biogeographical inference, buccal swabs from two individuals with one European parent from Germany or Italy and one from either the Philippines or Venezuela were collected using a DNA-free swab (Sarstedt, Nümbrecht, Germany). Approval for collection and DNA analysis was obtained from the ethical committee of the University of Freiburg (414/18). DNA was extracted using the QIAamp Mini Kit (Qiagen, Hilden, Germany) and AIMs sequenced using the ForenSeq DNA Signature Prep Kit (Mix B) with the MiSeq FGx® Reagent Micro Kit on a MiSeq FGx (all Verogen, San Diego, CA, USA). Sample preparation and sequencing was performed according to the Manufacturer's recommendations. SNPs were analyzed and exported for inclusion in the model using the ForenSeq Universal Analysis Software (Verogen).

As a reference dataset for the analysis of the recent-admixture model (used for computing allele frequencies for continental populations), we use the Forensic *MPS AIMs Panel Reference Sets*, taken from `http://mathgene.usc.es/snipper/illumina_55.xlsx` which comes with the software SNIPPER [5]. This dataset contains data from the 1000 genomes project (504 out of 661 individuals from Africa (AFR) excluding the samples from African Caribbeans in Barbados and Americans of African Ancestry; 85 out of 347 Admixed Americans (AMR) only including Peruvians from Lima; 504 East Asians (EAS), 503 Europeans (EUR) and 489 South Asians (SAS)), as well as 13 Oceanian, Papua New Guinea, (OCE) samples from the Human Genome Diversity Panel. In the reference dataset, rs3811801 (contained in the ForenSeq DNA kit) is missing and therefore excluded from further analysis. This SNP has some discriminatory power for EAS (allele frequencies 1 (AFR); 1 (AMR); 0.49 (EAS); 0.99 (SAS)), as seen from the 1000 genomes data. Since data for rs1919550 and rs2024566 is missing for the Oceanic samples of the reference database, we also excluded these AIMs. Both only have low discriminatory power on a continental level. In total, this amounts to a total of 53 AIMs in the analysis, all of which are contained in the Kidd AIMset [21]. Allele frequencies are displayed in Figure S9.

3 Results

3.1 Classification fails frequently for recently admixed individuals in a three island model

We simulated genome-wide data from a sample, taken from a population genetic model with three islands, A, B and C, using [22]. Migration is such that only A, B and B, C are connected, but not A, C; see Figure 1(A). 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 [6] to look for 10 Ancestry Informative Markers (AIMs). When using a naive Bayes approach as in SNIPPER [5], this AIMset gives a vanishing misclassification error for the task of classifying the $3{\times}400$ simulated, non-admixed individuals. However, on recently (first generation; see Section 2.3) $A \times C$-admixed individuals, the classifier fails in many cases and gives

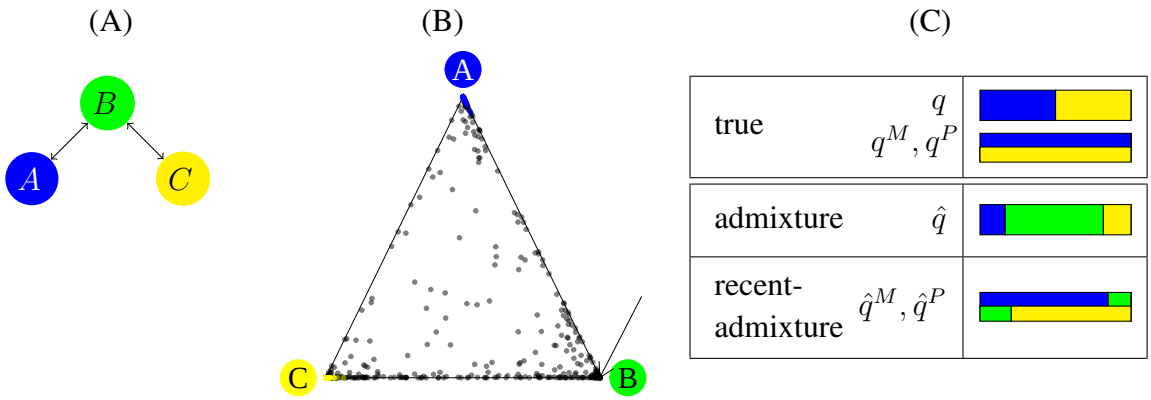

Figure 1: (A) Illustration of the population model for the simulations. (B) Classification results for non-admixed and $A \times C$ admixed individuals (individuals from A: blue; individuals from C: yellow; $A \times C$ admixed individuals: grey). The individual from (C) is indicated by the arrow and is assigned to B with probability almost 100%. (C) For the individual from (B), estimates of IA ($\hat{q}$) using the admixture model and of PIA ($\hat{q}^M, \hat{q}^P$) using the recent-admixture model.

population B as the best guess; see Figure 1(B).

3.2 The recent-admixture model improves accuracy of ancestry proportions

Subsequently, we used the admixture and recent-admixture model to estimate IA and PIA for both, non-admixed and first generation admixed individuals. 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. Figure 1(C) shows an example of an $A \times C$-recently-admixed individual, with misleading estimate for IA, but a more enlightening estimate for PIA. Here, IA predicts ancestry mostly in B when using admixture, since allele frequencies in B are between A and C. However, recent-admixture correctly predicts two parents of different ancestry, one mostly A, the other mostly C. The individual is taken from all $A \times C$-admixed individuals, which are displayed in Figure S1.

To get a picture of all non-admixed and recently admixed samples, we computed errors for estimating IA as given in (6) for the admixture and recent-admixture model. As described in the MM section, 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 ($A, B, A \times B, A \times C, B \times C$: $p < 0.001$, $C : p = 0.14$). This shows that on average the recent-admixture model performs better than the admixture model, even on non-admixed samples.

3.3 The recent-admixture model improves estimation accuracy of ancestry proportions for 1000 genomes project data

For comparing the accuracy of the admixture and recent-admixture model, we extended our analysis of the errors for estimating IA to the 1000 genomes dataset. We excluded all Admixed Americans (AMRs) since they are known to have an admixed background [2, 6] and do not form a well-defined own group. As the true IA, we use the continental origins as described in the dataset (AFR, EAS, EUR and SAS). This means e.g. that we set $q_{\mathrm{EUR}}^{\mathrm{TRUE}} = 1$ for a European sample in the dataset.

We ran three kinds of analyses. First, on the non-admixed samples. 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. grandparents were taken from the non-admixed samples (denoted AFR/EAS$\times$EUR/SAS etc). In the first case, Figure S3 shows that the resulting errors for the admixture and recent-admixture model are almost identical, when using the EUROFORGEN AIMset. Overall, recent-admixture has a smaller error in 1364 out of 2157 cases, i.e. the 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, when using the EUROFORGEN AIMset. 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) – recall from Section 2.3 – and shows that again, recent-admixture is more accurate, when using the EUROFORGEN AIMset. 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 also found in the SI. We stress that the recent-admixture model not only gives significantly better estimates for IA, but also provides more information than the admixture model, since the genetic decomposition of both parents is estimated.

3.4 Power of the Likelihood-ratio test for recent admixture

When fixing the minimal $\Delta\ell$ for deciding if a sample is recently admixed in the likelihood-ratio 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 $\Delta\ell$, we obtain the Receiver-Operation-Characteristic (ROC) curve [23]. As we see in Figure 3, for the EUROFORGEN AIMset, the power of the test differs with the kind of admixture. For first generation admixed (case (B)), one non-admixed parent (case (E)) and all grandparents 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. The picture is nearly identical for the Kidd AIMset; see Figure S8.

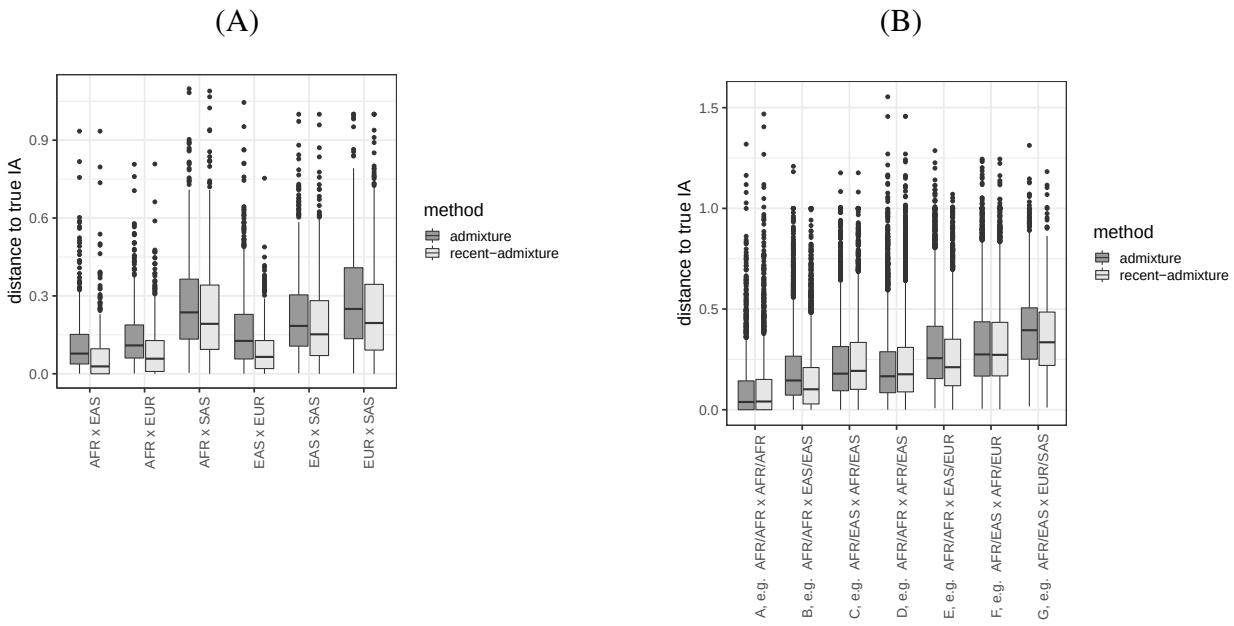

Figure 2: For all first generation admixed samples (A) and second generation admixed samples (B), we computed IA from the admixture ($\hat{q}$) and recent-admixture ($\frac{1}{2}(\hat{q}^M + \hat{q}^P)$) model using the EUROFORGEN AIMset. The distance to the true IA is computed as in (6). The cases in (B) are as described in Section 2.3.

3.5 The LR test identifies recent admixture in the 1000 genomes dataset

From the 1000 genomes dataset, we highlight individuals which give highly significant results for the test of recent admixture for the EUROFORGEN and Kidd AIMsets. As trainingset, for estimating allele frequencies, we use the individuals from `http://mathgene.usc.es/snipper/illumina_55.xlsx` which are part of the 1000 genomes dataset; see Section 2.7. In Figures 4, S10 and S11, we give the result of the most extreme individual (in the sense of the largest $\Delta\ell$ observed in the whole sample), a male from the African American (ASW) population. We note that it is known that the ASW population is admixed [2], but until now, it has not been tested if admixture is recent. We see that heterozygous sites are in fact the drivers of the large difference in log-likelihood between the recent-admixture and the admixture model, $\Delta\ell$. The likelihood-ratio test indicates that it is $e^{6.747} \approx 8.5 \cdot 10^2$ times more likely that the individual is recently admixed than non-recently, when using the EUROFORGEN AIMset and $e^{10.010} \approx 2.2 \cdot 10^4$ times more likely for the Kidd AIMset. Similar results appear in the Admixed American population, where we find an individual which appears to be recently admixed from Europe and Admixed American (individual NA19720); see Figures S12 and S13. We note, however, that the result for recent admixture may in some cases depend on the AIMset used. E.g., for another Admixed American from Mexico in the sample, NA19719, Figures S12 and S13 show that the evidence for recent admixture using the EUROFORGEN AIMset is much greater that for the Kidd AIMset.

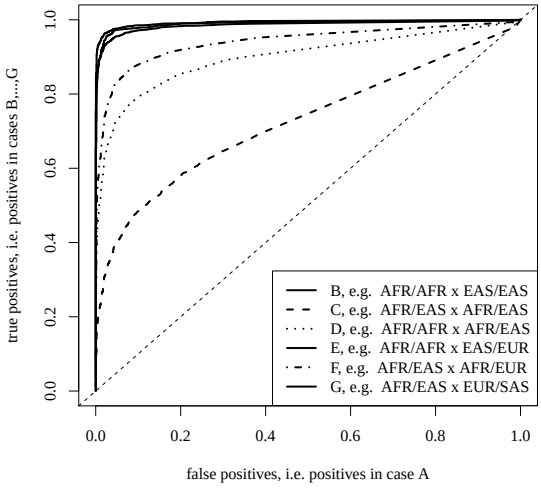

Figure 3: Using the EUROFORGEN AIMset, the ROC curve displays false positives (i.e. non-admixed individuals with high $\Delta\ell$), againts true positives (all cases of admixture in second generation).

(A) (B)

EUR, AFR, AMR, Other

(C)

	number	contribution to $\Delta\ell$
heterozygote sites	51	15.177
homozygote sites	70	-8.430
sum	121	6.747

Figure 4: Analysis of NA20278 from the 1000 genomes dataset using the EUROFORGEN AIMset. (A) Estimated IA. (B) Estimated PIA. (C) Contributions to the difference in log-likelihood $\Delta\ell$ from the likelihood-ratio test. More details on these contributions, broken down to single AIMs, is found in Figure S10.

3.6 Individuals with recent admixture sequenced with the Verogen MiSeq FGx™ Forensic Genomics Solution

For the German/Philippine female, when using a classification tool based on a naive Bayes approach (e.g. SNIPPER), data from the 53 autosomal markers indicate a 61% chance to be European and 39% to be South-East-Asian (with reference samples from India, Pakistan etc). The ForenSeq Universal

Analysis Software did not provide a clear classification result into one cluster of the training dataset, but the sample falls together with the Admixed American samples of the 1000 Genome project. The closest centroid contains samples from the 1000 genome populations in which also mainly samples from Puerto Rico and Colombia fall into. The use of the admixture model leads to a mixed ancestry from Europe, East-Asian and Oceania; see Figure 5(A). Oceania does not fit with the self-reported data and might result from a wrong conclusion due to the mixed SNPs of European and East-Asian ancestry. Using the recent-admixture model, one parent with European and one parent with mainly an East-Asian ancestry are revealed which fits the self-declaration. This individual has $\Delta\ell = 3.513$, i.e. a likelihood ratio of $e^{\Delta\ell} \approx 33$, such that the recent-admixed model is favoured.

For the Italian/Venezuelan male, when using a naive Bayes classifier, his DNA is classified as European. The ForenSeq Universal Analysis Software provides a classification rather into the European cluster, but states the closest centroid in which also single reference samples from the 1000 genome project from European as well as Middle- and South-American ancestry fall into.

The admixture model estimates mainly European ancestry, and contribution of South-East-Asian, and a small fraction African ancestry. The recent-admixture model estimates an European ancestry (explained by the Italian father) and one parent with mostly South-East Asian origin. So, recent-admixture is correctly predicted with a likelihood ratio of $e^{0.820} \approx 2.71$ relative to non-recent admixture, but the East-Asian ancestry of one parent is in contrast to the Venezualian ancestry of the mother. However, note that the only reference population near Venezuela are Admixed Americans from Peru.

(C)

	number	contribution to $\Delta\ell$
heterozygote sites	19	4.639
homozygote sites	34	-1.127
sum	53	3.513

(F)

	number	contribution to $\Delta\ell$
heterozygote sites	18	1.766
homozygote sites	35	-0.946
sum	53	0.820

Figure 5: (A)–(C) Same as Figure 4, but for the first collected individual. (D)–(F) Same for the second collected individual. In Figure S16, contributions to $\Delta\ell$ are broken down to single AIMs.

3.7 Runtimes

The analysis of the admixture and recent-admixture model is fast. The main reason is that allele frequencies are only computed from a reference database (and not estimated on the fly, as in STRUC-

TURE and ADMIXTURE). As a consequence, runtimes scale linearly with the number of analysed traces. E.g. once allele frequencies for all AIMs from the reference dataset are given, one of 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

Our first objective was to assess how recent admixture affects the outcome of standard methods of BGA inference. Using simulated island populations as well as data generated with the Verogen MiSeq FGx™ Forensic Genomics Solution, we provide in depth evidence that using classification software such as SNIPPER (all-or-nothing classifiers) is not suitable in such cases, as has been suggested by [18].Prestes et al. (2016) [24] notes that mixed membership models as used in STRUCTURE [4] and ADMIXTURE [3] are in many cases capable of inferring mixed ancestry in such individuals and [18] introduced a genetic distance algorithm, which improves results from ADMIXTURE in specific cases. However, we show that also the gold-standard for BGA inference, STRUCTURE and ADMIXTURE, can be misleading in recently admixed individuals. These models fail because the assumption of equal chances for the two alleles at a locus to stem from any population is violated in recently admixed individuals.

Our second objective was improving the inference of BGA (or estimation of admixture composition) and to develop a test to identify recently admixed individuals by inference of PIA. We used the excess of heterozygote sites in recently admixed individuals – called the Wahlund principle [10] – in order to estimate parental admixture (see also [13]) and test for recent admixture. Note that we still assume that AIMs are spread out in the genome, leading to independent segregation. This is in contrast to approaches using genome-wide dense SNP data, where linkage has to be taken into account [12, 11]. Another assumption made in our analysis is that allele frequencies in all populations are provided in a reference database. This is in contrast to the approach made by STRUCTURE, where IA and allele frequencies are estimated at the same time. A consequence of the STRUCTURE approach is that the analyzed traces will shift the allele frequencies in the clusters that are posthoc assumed to represent the reference populations. We believe that in forensic genetics, the outcome of the analysis should not depend on e.g. the number of traces which are studied, and thus a method that relies on a reference data base should be preferred. A welcome side effect of leveraging a reference data base is that our analysis is faster than STRUCTURE and ADMIXTURE. Since our method does not need genome-wide data, but only a few AIMs, analysis can be further accelerated and applied to forensic casework.

In inference of BGA, several hypothesis can be tested by a likelihood ratio (LR) test. For example, Tvedebrink and colleagues have recently developed statistical tests with the null hypothesis that the trace is a non-admixed sample from one of the populations in the reference database [15]. This test was extended in [25] to the null hypothesis that the trace is a recent admixture of the (non-admixed) parents from two populations in the reference database. In contrast, the LR test presented here tests if the recent-admixture model (with two parents with different IA) fits the data significantly better than

the admixture model (where both alleles at a locus have the same chance given by IA to originate in any of the continents), and thus can be considered a formal test of recent admixture. As expected for a method that relies on a genome wide increase in heterozygosity to detect recently admixed individuals, we see that heterozygote sites are responsible for the biggest contribution to the increase in likelihood in the recent-admixture model.

In our real-world samples, we detect signs of recent-admixture both in the 1000 genomes dataset and the two self-sequenced individuals with recent-admixture. Note that the 1000 genomes project specifically targeted individuals that are most likely to have all of their grandparents in only one location or population [26]. Nevertheless, we find striking evidence for recent admixture in some African individuals from the southwest of America. For the two self-sequenced samples , the analysis of recent-admixture was compared with the self-reporting of the individuals. The ancestry information is therefore based on the assumption of having a biologically correct family tree and correct information on birth place and ancestry. In case of the female individual, Philippine and German ancestry were declared for the last four generations. The other individual, with self-reported maternal Venezuelian ancestry of the last two generations, was ”insecure” about the generation before. This fits our results only if the maternal line has South-East ancestors. However, we also have to note that Venezuela is no part of the reference dataset. Another explanation for the diverging reported ancestry from the estimation result on parental ancestry could be that the only admixed American population (Peruvians) in the reference dataset is genetically further apart from Venezuela than the South-East Asian reference samples.

The recent-admixture model gives accurate results even on small AIMsets. Most importantly, we see in our simulations and real world examples, that the results from the recent-admixture model outperform the admixture model in virtually all cases. In addition, they provide additional insights of the recent ancestry of the trace which can guide casework.

Acknowledgements

We thank Denise Syndercombe Court for comments on an earlier version of the manuscript. The sequencing results are part of a larger study which is party funded by the Wissenschaftliche Gesellschaft Freiburg.

References

[1] 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.

[2] 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 Con-

sortium, 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.

[3] 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.

[4] F. Staubach. Germany: Note limitations of DNA legislation. *Nature*, 545(7652):30, 2017.

[5] 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.

[6] 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*, 46:102259, 2020.

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

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

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

[10] S. Wahlund. Zusammensetzung von Populationen und Korrelationserscheinungen vom Standpunkt der Vererbungslehre aus betrachtet. *Hereditas*, 11:65–106, 1928.

[11] J. Y. Zou, E. Halperin, E. Burchard, and S. Sankararaman. Inferring parental genomic ancestries using pooled semi-Markov processes. *Bioinformatics*, 31(12):i190–196, Jun 2015.

[12] J. Pei, Y. Zhang, R. Nielsen, and Y. Wu. Inferring the ancestry of parents and grandparents from genetic data. *PLoS Comput. Biol.*, 16:e1008065, 2020.

[13] D. J. Crouch and M. E. Weale. Inferring separate parental admixture components in unknown DNA samples using autosomal SNPs. *Eur. J. Hum. Genet.*, 20:1283–1289, 2012.

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

[15] T. Tvedebrink, P. S. Eriksen, H. S. Mogensen, and N. Morling. Weight of the evidence of genetic investigations of ancestry informative markers. *Theoretical Population Biology*, 120:1–10, 2018.

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

[17] 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.

[18] 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.

[19] 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.

[20] 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.

[21] 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.

[22] 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.

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

[24] 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.

[25] T. Tvedebrink and P. S. Eriksen. Inference of admixed ancestry with ancestry informative markers. *Forensic Science International. Genetics*, 42:147–153, 2019.

[26] 1000 Genomes Project Consortium. 1000 genomes project: Developing a research resource for studies of human genetic variation. consent to participate. `https://www.internationalgenome.org/sites/1000genomes.org/files/docs/Informed%20Consent%20Form%20Template.pdf`, download 26.8.2020.