

In which cases does it make sense to classify a trace into one population, and what are the alternatives?

Julika Schmitz

University of Freiburg

Jana Naue

University Medical Center Freiburg

Lukas Hoffmann

University Medical Center Freiburg

Zoe Lange

Ruhr University Bochum

Suchanek Ferdinand

University Medical Center Freiburg

Sabine Lutz-Bonengel

University Medical Center Freiburg

Holger Dette

Ruhr University Bochum

Peter Pfaffelhuber

`peter.pfaffelhuber@stochastik.uni-freiburg.de`

University of Freiburg

Carola Sophia Heinzl

University of Freiburg

Research Article

Keywords: Biogeographical ancestry, Ancestry informative markers, Admixture, Out-of-reference detection, Forensic genetics

Posted Date: October 19th, 2026

DOI: <https://doi.org/10.21203/rs.3.rs-11172480/v1>

License: © ⓘ This work is licensed under a Creative Commons Attribution 4.0 International License.

[Read Full License](#)

Additional Declarations: No competing interests reported.

In which cases does it make sense to classify a trace into one population, and what are the alternatives?

Julika Schmitz², Jana Naue ^{[iD](#)}², Lukas Hoffmann², Zoe Lange ^{[iD](#)}³,
Ferdinand Suchanek², Sabine Lutz-Bonengel², Holger Dette ^{[iD](#)}³,
Peter Pfaffelhuber ^{[iD](#)}^{1*}, Carola Sophia Heinzl ^{[iD](#)}¹

^{1*}Department of Mathematical Stochastics, University of Freiburg,
79104 Freiburg im Breisgau, Germany.

²Institute of Forensic Medicine, Medical Center – University of
Freiburg, Faculty of Medicine, University of Freiburg, Freiburg im
Breisgau, Germany.

³Faculty of Mathematics, Ruhr University Bochum, Bochum, Germany.

*Corresponding author(s). E-mail(s):

peter.pfaffelhuber@stochastik.uni-freiburg.de;

Abstract

During biogeographical ancestry (BGA) analysis, an unknown individual for whom ancestry informative markers (AIMs) were analyzed is often assigned to one of a few reference populations. For some individuals, such an assignment is not meaningful: the two parents of the individual may come from different populations, or the individual may come from a population not included in the reference database. In this study, we examined classification methods in combination with 104 AIMs from the VISAGE Enhanced Tool to provide one of the following three possible outcomes: (i) in-reference (IR), with an assignment to one of the reference populations; (ii) admixed (AD), with parents from two in-reference populations; and (iii) out-of-reference (OR), where the individual's ancestry is not sufficiently represented by the reference populations. We therefore compared seven classifiers in combination with class-specific acceptance thresholds for the OR decision. The best-performing method according to our selection criterion was a naive Bayes classifier over pairs of populations, which we call the recent naive Bayes (RNB) classifier. Populations held out of the reference panel were correctly rejected in 84.9% of cases. We further tested the method on 800 individuals from seven additional groups not included in the reference panel. Only 29.4% of individuals from four non-admixed populations were rejected,

resembling joint ancestry with reference populations. In addition, we investigated the prediction accuracy in a newly collected dataset of 28 individuals with self-reported ancestry by analyzing AIMS using massively parallel sequencing.

Keywords: Biogeographical ancestry, Ancestry informative markers, Admixture, Out-of-reference detection, Forensic genetics

1 Introduction

In recent years, forensic DNA analysis has evolved beyond the traditional identification of individuals using short tandem repeat (STR) profiles. In addition to individual identification, it increasingly includes methods for predicting externally visible characteristics, chronological age, and biogeographical ancestry of an unknown individual who left a trace at a crime scene or an unknown deceased person [26, 27, 40, 54]. Such extended analyses become particularly important when the determined individual STR profile does not match a (suspect) reference sample or any entries in (inter-)national DNA databases [32, 51]. In these cases, additional phenotypic or population genetic information can support investigations by narrowing down the pool of potential suspects or prioritizing investigative measures [5, 27, 51, 54]. In Germany, for example, the prediction of eye, hair, and skin color as well as age is permitted under Section 81e of the Code of Criminal Procedure (regulating forensic DNA analysis), while the determination of biogeographical ancestry (BGA) continues to be the subject of social and legal debate. [14, 69]. Methodologically, these prediction approaches rely on different genetic and epigenetic markers. Chronological age is typically predicted from age-correlated DNA methylation patterns at specific CpG sites [15, 34, 64], whereas externally visible characteristics and biogeographical ancestry are mainly predicted using single nucleotide polymorphisms (SNPs) [27, 57]. For biogeographical ancestry prediction, ancestry informative markers (AIMs), typically biallelic SNPs showing substantial differences in allele frequencies between populations, are compared with reference datasets to infer an individual’s most likely population of origin [25, 29, 33, 35, 37, 40, 46, 47, 55, 61, 62]. BGA can be traced back to continental populations using forensic AIMs panels, including the populations of sub-Saharan Africa (AFR), Western Europe (EUR), Middle East (MEA), South Asia (SAS), East Asia (EAS), Native Americans (AMR), and Oceania (OCE) [10, 11, 28, 36, 44, 48, 58, 59, 66, 68]. A finer differentiation of geographically close populations usually requires the analysis of a larger number of AIMs. In forensic practice, however, only small amounts of DNA are often available, so smaller marker panels with particularly high population divergence are preferred in order to ensure sufficient sensitivity [41, 49]. Various classification and clustering methods are used for the statistical evaluation of such markers, including the naive Bayes classifier SNIPPER [42], the machine learning method TabPFN [19], and the admixture model (AM) as implemented in STRUCTURE [43]. These methods use reference datasets with population-specific allele frequencies to calculate the probability of a DNA profile belonging to different groups of origin [52]. The classification of individuals without

mixed ancestry at the continental level is generally reliable [12, 28, 38]. However, more complex situations such as admixed ancestry or ancestry from populations that are not sufficiently represented in reference datasets cannot be handled correctly. Genetic admixture, in which different proportions of the genome originate from two or more populations, therefore poses a particular challenge. Hence, the question arises under which circumstances classifying an individual into a single population is justifiable. Two situations make it inadvisable. The individual may be admixed. In this setting, we have two possibilities: we can classify the parents of the individual into two different classes or, if this is not possible, classify the individual as "admixed". Another reason why classifying an individual into a single population might not be meaningful is that the individual may originate from a population that is not represented in the reference data. A usable method must be capable of distinguishing both scenarios from those in which a classification is appropriate. Furthermore, this must be achievable without prior knowledge of the specific class category.

Admixed ancestry has been approached from several directions. STRUCTURE and ADMIXTURE answer it indirectly [1, 43]; Pfaffelhuber et al. [39] refine this to the possibility to two admixed parents. Posing admixture as a hypothesis test is more recent. Smyser et al. [56] test whether L ancestral populations contribute to a trace against $K := L + 1$ in the alternative. Dette et al. [9] test that the individual descends from exactly one reference population against the admixture alternative, using the constraint based bootstrap approach.

Tvedebrink et al. [62] addressed the second situation. They test H_0 : the individual originates from population k , against H_1 : the individual is his or her own population, i.e. the reference dataset is not sufficient (referred to as "TVE" in this study). Tvedebrink and Eriksen [60] extended the test to a pair of populations, so that a first-order admixed individual becomes the hypothesis (referred to as "TVP" in this study); that extension answers both questions at once, which is why TVP but not TVE can name two parental populations in the comparison below. Their approach and the approach developed in this study are related but differ because we aim to construct a classifier (assigning the individual to a discrete class) rather than a statistical test.

Outside forensics, a classifier that may refuse to decide is long established. Chow [7] derived the optimal error-versus-reject trade-off for known distributions; classification with a reject option has since received a statistical learning treatment [3, 8, 22], and is called selective classification [17]. What we call out-of-reference is, in that literature, an open-set or out-of-distribution problem [18, 21, 53], and a common strategy is to combine a scalar score of fit with a threshold calibrated on held-out data [13, 31, 67].

Here, we combine a classifier with an *acceptance rule*: the classifier assigns a class to every individual, and the acceptance rule decides whether that class is reported or rejected. The acceptance rule has two parts: a score, which ranks how well an individual fits the reference panel, and a threshold, which turns that ranking into a decision. Because an individual assigned to a small population and one assigned to a large one are not comparable on the same scale, the threshold is fitted per predicted class (population), on data held out for that purpose.

As an alternative to classifying an individual into one population, we propose to classify its two parents into potentially different populations. This is related to the

recent admixture model (RAM), where the aim is to estimate the share of the genome originating from each population for each parent [39]; we therefore call the classifier the recent classifier. Because its classes are pairs of populations, a non-admixed individual is the pair (k, k) and admixture is a hypothesis of the model rather than a class that has to be learned from simulated training data. In total, seven classifiers for IR/AD/OR classification were considered in this study: naive Bayes (NB) [38, 42], recent naive Bayes (RNB), which is introduced here, the admixture model (AM) [1, 43], the recent admixture model (RAM) [39], TabPFN, a tabular foundation model [23, 24], which Heinzl et al. [19] applied to biogeographical ancestry, and the goodness-of-fit statistics classifiers (TVE, TVP) [60, 62]. Furthermore, an acceptance/rejection rule for classification is proposed.

This manuscript is organized as follows. Section 2.2 to section 2.4 describe the reference data, the evaluation design, the seven classifiers we compared (see above), the acceptance rule, and the quantities we report. Section 3.1 onwards compares the seven classifiers using reference data with one population absent in training, and examines the resulting error patterns. Section S1 describes the full pipeline, including the alternatives that were run and rejected, and Section S2 gives the full probabilistic models. Figure 1 visualizes the proposed methods and fig. 4 gives an overview of the evaluation of the approach. Finally, the developed approach was applied to 800 independent samples from [50], and novel buccal samples from 28 individuals with self-reported ancestry.

Every number and every figure below can be recomputed from the reference implementation at our [GitHub repository](#). It contains the full code we ran on all samples. In addition, a standalone tool in that repository scores new samples under the recommended configuration and writes one report per sample.

2 Material and Methods

2.1 Workflow

An overview of the pipeline is provided in Figure 1. The inputs are the genotypes of the 104 AIMs of the VISAGE Enhanced tool (note that missing values are handled by the final model); see section 2.2. For the BGA analysis, we apply a classifier (see section 2.3), and return one of three things for every sample:

In-reference (IR) The genetic data from the individual is most consistent with ancestry from one reference population.

Admixed (AD) The genetic data from the individual is most consistent with parents from two different reference populations;

Out-of-reference (OR) The population of the individual is not represented in the reference data.

No classifier directly returns “out of reference” as a class. In contrast, each classifier assigns an individual to one of its classes and provides a score for that assignment. If the score falls below a defined threshold, no final classification is reported because the assignment is rejected due to too high uncertainty. We call this the acceptance rule.

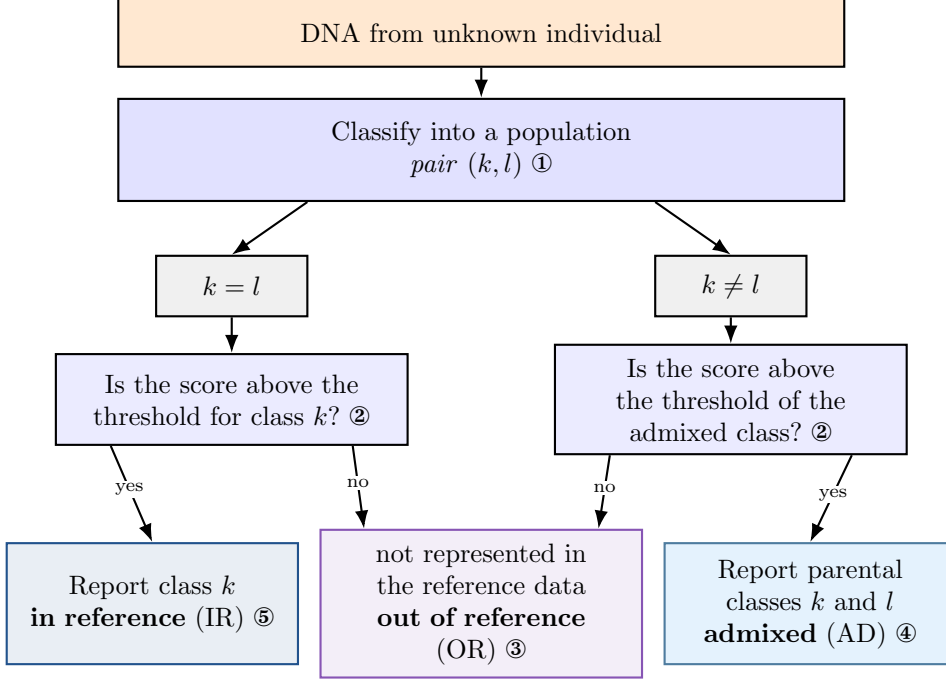


Fig. 1 Overview of the classification pipeline and the quantities used for evaluation. For every pair-capable classifier (recent naïve Bayes (RNB), recent admixture model (RAM), admixture model (AM), and goodness-of-fit statistics classifier [60] (TVP)), we make a prediction of the population pair (k, l) ①. A score is then computed and compared with a threshold (which depends on the predicted class, with one shared threshold for Admixed) and the classification is retained only if the score is sufficiently high ②. If the score is below the threshold, the acceptance rule rejects the classification, resulting in out-of-reference ③. If the score is high enough and $k \neq l$, the sample is reported as admixed ④. If the score is high enough and $k = l$, the sample is reported as in reference ⑤.

If we have K populations in the reference data, there is a total of $K + \binom{K}{2} + 1$ possible outcomes.

2.2 Used reference data

In this study, we used Supplementary Table S1A of Ruiz-Ramírez et al. [50] as the reference dataset, which lists genotypes of the 104 AIMs of the *VISAGE Enhanced Tool for Appearance and Ancestry* for 4341 individuals in 14 groups (group names used as provided in Ruiz-Ramírez et al. [50]). These markers also formed the basis for the analysis of 28 individuals in this study, allowing direct comparison (cf. section 2.7 and section 3.4).

We divided the 14 groups into two sets: Seven of them are continental populations and make up our reference panel, comprising $K = 7$ classes and $N = 3541$ individuals with the following composition in descending frequency: 935 (26%) East Asian (EAS), 827 (23%) European (EUR), 730 (21%) South Asian (SAS), 623 (18%) African (AFR), 240 (7%) Middle East (MEA), 109 (3%) Native American (AMR), and 77 (2%) Oceanian (OCE). Note that the sample distribution is highly imbalanced, both

in class size and in genetic heterogeneity within classes, which can affect methods that weight individuals and populations differently.

The remaining seven groups, comprising a further 800 individuals, served as additional downstream test data, so they did not enter training and calibration of the classifiers; see section 3.3: CENTRAL ASIAN (64), EAST AFRICAN (47), NORTH AFRICAN (29) and ROMA (3); MIDDLE EAST samples ('Sanger population data') from [2] (136); and two cohorts carrying an admixed label, ADMIXED AMERICAN (364) and ADMIXED AFRICAN (157).

It has to be considered that two of the three possible BGA situations (AD and OR) cannot be observed directly in the reference data.

Origin of the out-of-reference individuals

By definition, an out-of-reference individual is absent from the reference data. To construct this setting, we hold out one population at a time (we call this a *fold*) and treat its individuals as out-of-reference, with one modification: EUR and MEA are held out together. The two are at $F_{ST} = 0.047$ (Section S1.1), the smallest separation in the panel, and holding out one of them alone most often leads to classification of these (out-of-reference) individuals into the other population. Section S1.2 describes the alternative evaluation designs we ran as controls. This setting results in six different folds, depending on which population or populations are held out.

Origin of the admixed individuals

There are no admixed individuals of known parentage in the reference panel, so that class is constructed in silico. In each run, admixed individuals are generated from the corresponding in-reference individuals. An offspring in the AD class is an F_1 child of two in-reference individuals, one (independent) allele per marker from each parent, 15 per pair of parental populations, seeded identically for all methods. An out-of-reference individual is never used as a parent, so both parents of every admixed individual lie inside the reference. Note that the construction is restricted to first-generation admixture; no backcrosses or deeper admixture are introduced.

2.3 Classifiers compared for BGA prediction

Seven classifiers were compared. For each individual, each classifier returns a predicted class and at least one score; see (3) and Section S2.

NB: naive Bayes over populations [38, 42]. Its classes are the seven populations plus one pooled AD class containing the simulated "training offspring".

RNB: recent naive Bayes introduced here: naive Bayes over unordered population *pairs* with repetition, $\binom{K+1}{2} = 28$ for $K = 7$, where a pair (k, k) is the non-admixed case. Because admixed individuals are represented directly within the model, no simulated admixed training offspring are required. Figure 2 illustrates the underlying model.

AM: the admixture model [1, 43]: continuous proportions $q \in \Delta^{K-1}$ (the $K - 1$ -dimensional simplex) fitted per individual against the training data frequencies. It has

no discrete non-admixed/admixed criterion, so we fit the criterion during calibration: an individual counts as admixed when its largest proportion falls below a cut-off.

RAM: the recent admixture model [39] (fig. 3): proportions fitted separately for the two parental halves. The individual is non-admixed when the two halves agree and admixed otherwise, so unlike AM the criterion needs no cut-off.

TabPFN: a pre-trained in-context tabular neural network (transformer) for tabular data [23, 24], applied to biogeographical ancestry by Heinzl et al. [19], fitted on the training data and giving a posterior over the classes it was trained on. Note that no likelihood-based probabilistic interpretation is available because the model does not provide a likelihood.

TVE, *TVP*: goodness-of-fit statistics, per population [TVE, 62] and per population pair [TVP, 60]. The first has no notion of two parents and therefore never gives an admixed label.

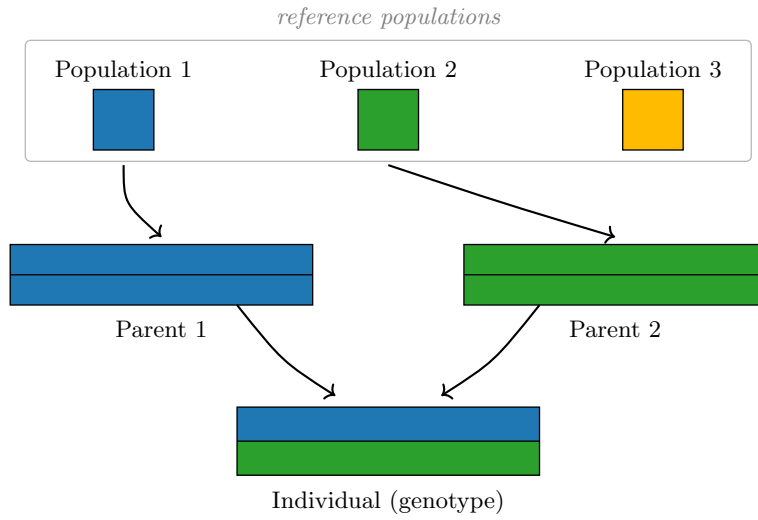


Fig. 2 Model for the RNB. The three ancestral populations at the top are colour keys, shown as upright boxes; parental and individual genomes are represented by horizontal bars. Both parents are diploid and therefore carry two bars each. Under the RNB a parent belongs to exactly one population, so its two bars share one colour (in this example, POPULATION 3 contributes nothing). The individual receives one allele per marker from each parent, and the markers are inherited independently. The RNB scores the *pair* of populations directly, its estimate is therefore a pair (k, l) , with the non-admixed case represented by (k, k) (both bars have the same colour).

Restrictions of the classifiers

The pipeline must determine three aspects from the genotyped AIMs: the population assignment, whether the individual is admixed, and, if so, the contributing populations. Of the seven classifiers above, four can answer all three: RNB, RAM, AM and

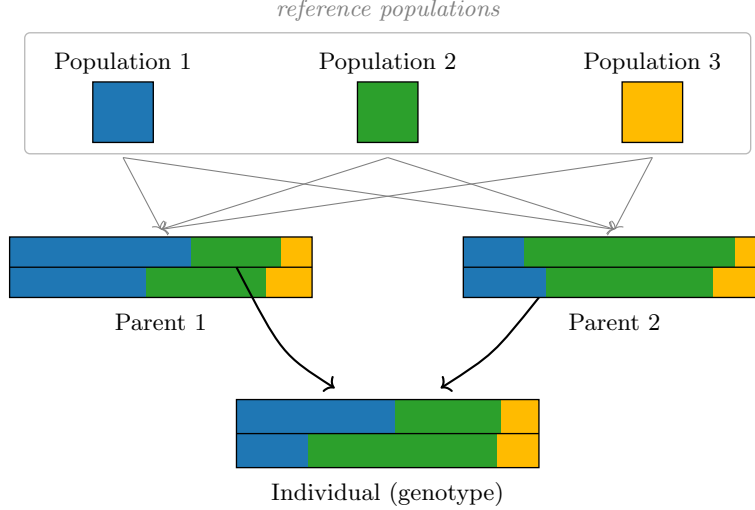


Fig. 3 Model for the RAM. The same schematic as in fig. 2 but with already admixed parents: the parents are again diploid, but now each of their two bars is divided between the three ancestral populations rather than carrying one colour. Because the markers are inherited independently, the genomic contribution an individual receives from a parent is a mosaic drawn from both of that parent’s bars, and its composition is therefore that of the parent as a whole — which is what the individual’s two bars show here. Those two bars are the estimates $q^{(1)}$ and $q^{(2)}$, and this is why they are parental *halves*: each summarizes one parent, not one of that parent’s two bars. An individual is classified as non-admixed when the largest components of both halves agree and as admixed otherwise. See [39] for more details.

TVP. These are exactly the four that name a pair rather than a single class, and therefore the only ones for which an estimated pair is available. NB and TabPFN have a pooled admixed class but no notion of which populations contribute, so they answer two of the three. TVE has no admixed class at all.

2.4 Description of the procedure to evaluate the classifiers

Splitting of reference data

Each classifier described in section 2.3 was evaluated over 20 repeated iterations. Each iteration contains in-reference individuals as well as one or more out-of-reference populations (see section 2.2). For the in-reference individuals, we have a 70/15/15 train/calibration/test split. Admixed individuals are simulated and out-of-reference individuals are defined as described above. In the calibration part, the offspring are pooled into a single AD class. All methods use identical splits, individuals and simulated offspring. The split is summarized in fig. 4. The composition of a split is the same in every iteration. Pooled over the six folds, one iteration’s test part contains on average 700 EAS in-reference individuals, 615 EUR, 546 SAS, 477 AFR, 177 MEA, 83 AMR and 60 OCE, together with 1275 admixed individuals. On the out-of-reference side a held-out population contributes 468, 414, 365, 312, 120, 55 and 39 individuals respectively. Within a single fold, MEA, AMR and OCE contribute fewer than

Evaluation design and reported values

every classifier: NB, RNB, AM, RAM, TabPFN, TVE, TVP
 $B = 20 \text{ iterations} \times 6 \text{ folds}$

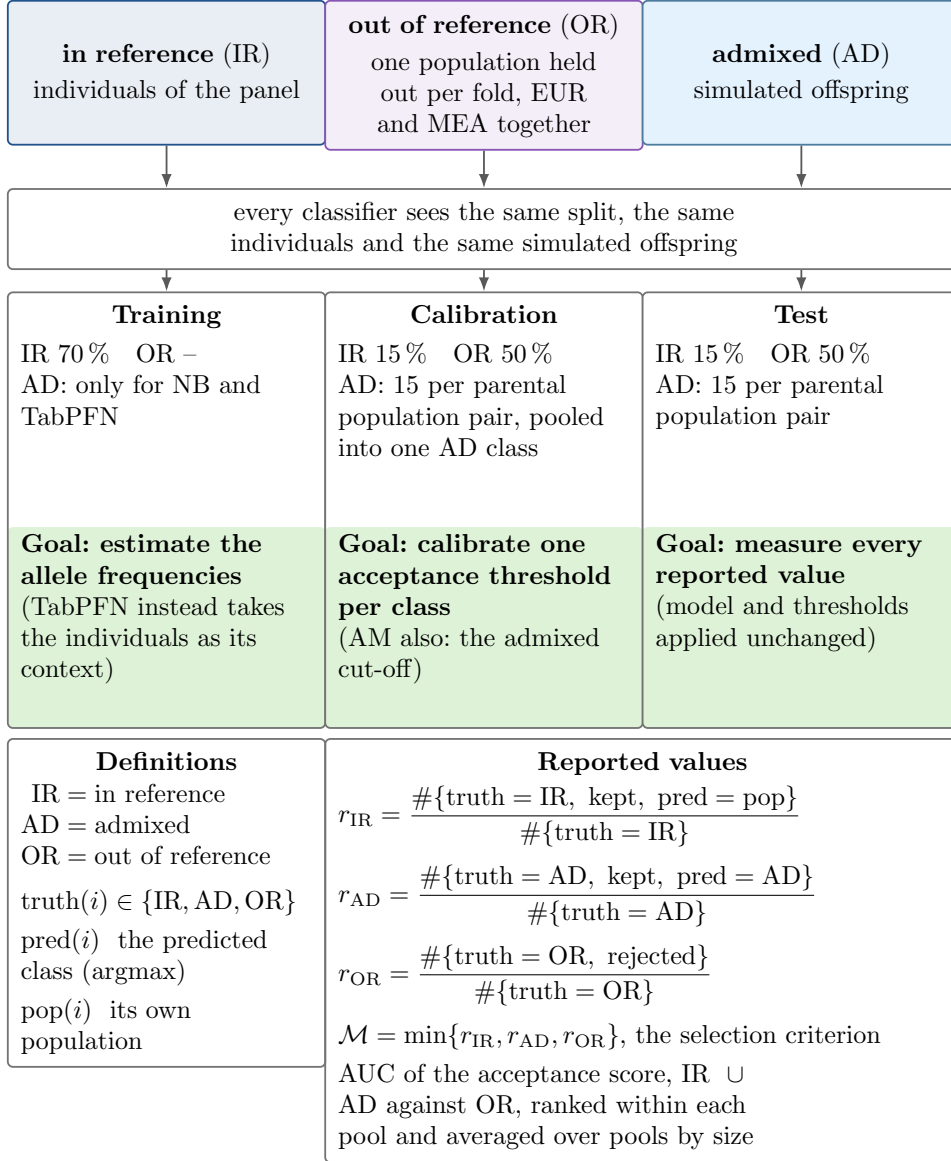


Fig. 4 The evaluation design and the values reported from it. Every iteration has three groups of individuals — the panel itself (IR), the population held out in that fold (OR), and simulated offspring (AD) — and splits them into three parts, with separate subsets used to estimate allele frequencies, calibrate class-specific acceptance thresholds, and evaluate model performance. All classifiers are evaluated using the same split, individuals and simulated offspring.

50 test individuals for evaluation each, which should be considered when interpreting population-specific results.

Training of classifiers

For six of the seven classifiers, allele frequencies are estimated from the in-reference individuals. TabPFN does not estimate allele frequencies at this step: it is an in-context learner, so the training individuals are handed to it as context and the pre-trained network uses them directly at prediction time. Two methods, NB and TabPFN (see section 2.3), additionally receive simulated admixed offspring here, because they can only predict a class they have seen; the other five obtain admixture structurally and train on the in-reference individuals alone.

Calibration

On the calibration set, every individual receives its acceptance score; see (3). Within each iteration, calibration predictions and scores from all six folds are pooled. For each classifier and acceptance score, one threshold $T_{c,b}$ is fitted within each predicted-class pool c in iteration b , by maximizing the mean of the class-specific retention and rejection rates defined in section 2.5. All admixed parental pairs share one Admixed pool and one acceptance threshold. The fitted thresholds are applied unchanged to the test predictions of the same iteration. For AM, a separate cut-off determining whether an individual is labeled Admixed is fitted on the calibration data within each fold, because the reference panel differs between folds.

Test of model performance

The thresholds estimated during calibration are applied unchanged to the test set. Every value reported below is computed on the test set.

Reported values

Two quantities ($\mathcal{M}$ and $\mathcal{A}$) are reported for every comparison. Write $\text{truth}(i) \in \{\text{IR}, \text{AD}, \text{OR}\}$ for the group an individual belongs to (in-reference, admixed, out-of-reference), $\text{pred}(i)$ for the class with the highest score (also called the $\arg \max$) and $\text{pop}(i)$ for an individual's own population. Then

$$\begin{aligned} r_{\text{IR}} &= \frac{\#\{i : \text{truth}(i) = \text{IR}, \text{kept}, \text{pred}(i) = \text{pop}(i)\}}{\#\{i : \text{truth}(i) = \text{IR}\}}, \\ r_{\text{AD}} &= \frac{\#\{i : \text{truth}(i) = \text{AD}, \text{kept}, \text{pred}(i) = \text{AD}\}}{\#\{i : \text{truth}(i) = \text{AD}\}}, \\ r_{\text{OR}} &= \frac{\#\{i : \text{truth}(i) = \text{OR}, \text{rejected}\}}{\#\{i : \text{truth}(i) = \text{OR}\}}, \end{aligned} \tag{1}$$

$$\mathcal{M} = \min\{r_{\text{IR}}, r_{\text{AD}}, r_{\text{OR}}\}, \quad \mathcal{A} = \frac{1}{3}(r_{\text{IR}} + r_{\text{AD}} + r_{\text{OR}}). \tag{2}$$

$\mathcal{M}$ is the selection criterion: a method is chosen by maximizing it over the classifier, acceptance score and threshold rule. The minimum requires the method to be adequate

in every case. The accuracy $\mathcal{A}$ is the mean of the same three rates and is reported beside it. Note that r_{IR} requires the correct population label; assignment to a neighboring population is therefore counted as an error, as is rejection.

An additional model evaluation parameter is the area under the curve (AUC) of the acceptance score, computed with IR $\cup$ AD as positives and OR as negatives. The AUC measures the score’s ranking ability without requiring a threshold. We report it per pool, ranked inside each pool and averaged over pools by size. See Section S1.6 for additional scores not reported here.

2.5 The acceptance rule: a score and a threshold

In standard (also called closed-set) classification, an individual is assigned to the class with the highest score. Here, motivated by open-set recognition and classification with a reject option, we extend this decision rule by introducing OR as an additional outcome [20, 53]: the highest-scoring reference class is reported only if its score is at least the corresponding acceptance threshold; otherwise, the individual is assigned to OR.

Assume that $q_k(x)$ is the (a posteriori) probability of class k for individual x , and, if available, $P(x|\theta)$ the probability of obtaining individual x , which has $M(x)$ successfully typed markers, under parameters θ (which depend on the classifier). This yields the two scores

$$s_{\text{post}}(x) = \max_k q_k(x), \quad s_{\text{lik}}(x) = \frac{1}{M(x)} \max_{\theta} \log P(x | \theta). \quad (3)$$

The available score depends on the classifier: TabPFN is discriminative and has only s_{post} , NB, RNB, AM and RAM are full likelihood models and have both. TVE and TVP have neither: their statistic already measures fit in absolute terms. However, because a small value means a good fit here, we use its negative so that higher values indicate a better fit in all cases. Because $\sum_k q_k(x) = 1$, s_{post} cannot express that no panel class fits, whereas s_{lik} can.

Recall that a class is the set of individuals carrying that label as a truth. In contrast, we write pool c for the individuals a method assigned to class c by its argmax. Each pool receives its own threshold T_c , fitted on the calibration set and applied unchanged to the test set.

A pool should keep individuals that belong to that label. Write $\text{correct}(i)$ for the class individual i should be given: its own population if it is in reference and Admixed if it is admixed. An out-of-reference individual’s own population is not among the classes, so $\text{correct}(i) \neq c$ holds for it in every c : it has no correct in-panel label and therefore always counts on the rejection side. Because a high score indicates a good fit, define

$$\begin{aligned} \text{keep}_c(t) &= \frac{\#\{i \in \text{pool } c : \text{correct}(i) = c, s_i \geq t\}}{\#\{i \in \text{pool } c : \text{correct}(i) = c\}}, \\ \text{rej}_c(t) &= \frac{\#\{i \in \text{pool } c : \text{correct}(i) \neq c, s_i < t\}}{\#\{i \in \text{pool } c : \text{correct}(i) \neq c\}}, \end{aligned}$$

and the threshold T_c maximizes $\frac{1}{2}(\text{keep}_c + \text{rej}_c)$. Note that pool c is fitted only using data from pool c . Section S1.4 gives two weaker criteria and additional alternatives not used in the reported analysis.

2.6 Using RNB on novel samples

While the training/calibration/test split had its purpose in recommending one method, we apply the RNB to novel samples below, and we have to adapt the pipeline for this. For deployment, RNB allele frequencies are estimated from all individuals in the seven reference populations. The fixed acceptance threshold for each predicted class c is obtained as the median of its iteration-specific calibration thresholds from the evaluation described in section 2.4, i.e. $T_c^{\text{deploy}} = \text{median}_{b=1, \dots, 20} T_{c,b}$. Here, $T_{c,b}$ is fitted to the pooled calibration predictions from the six folds of iteration b , as described above. The resulting eight thresholds, one for each reference population and one for Admixed, are used unchanged for the additional 800 individuals and the 28 newly analysed samples (sections 3.3 and 3.4). A prediction is retained when $s_{\text{lik}}(x) \geq T_{\hat{c}(x)}^{\text{deploy}}$, where $\hat{c}(x)$ is the predicted class.

2.7 Analysis of newly collected samples

Sample and library preparation

Buccal swab samples from 28 individuals were obtained with written informed consent using DNA-free swabs (Sarstedt, Nümbrecht, Germany). Approval was obtained from the Ethics Committee of the University of Freiburg (nr: 22-1427-S1). Biogeographical ancestry was inferred from information provided by the volunteers for the first (parents) and second (grandparents) generation; see Table S2. DNA was extracted using the QIAamp DNA Mini Kit (Qiagen, Hilden, Germany) according to the manufacturer’s protocol and quantified with the Power Quant System (Promega, Madison, WI, USA) on a QuantStudio 5 System with QuantStudio Design and Analysis Software v1.5.2 (Thermo Fisher Scientific (TFS), Waltham, MA, USA). An implemented assay for analysis of BGA markers selected from the AmpliSeq Enhanced VISAGE panel was used for sample analysis. The modified assay used here consists of 106 autosomal and 83 Y-chromosomal markers with PCR primers designed for a multiplex assay (primer mix A and B) by Illumina for library preparation using the NEBNext technology (New England Biolabs (NEB), Ipswich, MA, USA) and was run on a MiSeq FGx (Verogen (now a Qiagen company), San Diego, CA, USA). However, not all markers were used for BGA prediction within this study (cf. below). PCR reaction mix includes 1 ng of DNA, 5 μL Platinum Direct PCR Universal Master Mix (TFS), 5 μL Primer mix A or B, and DNA-free water to a volume of 20 μL . PCR was run under the following conditions: initial denaturation at 95 °C for 10 min, 35 cycles: 30 sec 95 °C (0.2 °C/s); 2 min 62 °C (0.2 °C/s); 2 min 68 °C, final elongation: 20 min 72 °C.

PCR products were purified with AMPure beads and library were prepared using the NEBNext UltraExpress DNA Library Prep Kit (E3325S, NEB) according to the manufacturer’s protocol. NEBNext Multiplex Oligos for Illumina (E6440S, NEB)

were used for sample indexing. After the final clean-up, sample quality and quantity were determined on the TapeStation 4150 using the High Sensitivity D1000 DNA ScreenTape (both Agilent, Santa Clara, CA, USA).

Sequencing and data analysis

Samples were pooled equimolarly to prepare a 10 nM library. The final library, including 20 pmol phiX, was sequenced using 2×150 bp cycles (300 bp cartridge and micro or standard flow cell) on a MiSeq FGX (Verogen, now Qiagen). Paired-end fastq files were trimmed (quality score: 20) using Trim Galore v0.4.3/Cutadapt v2.0 and aligned to the GRCh38 (primary assembly) using BWA-MEM (release 0.7.17). SNPs were called and filtered using BCFtools 1.22 (mpileup, call, query) with a quality threshold of 20. For final SNP calling a coverage threshold of 30 was set and variants were visually verified using IGV.

Population assignment

To compare the classifier output with the participant information, the reported ancestry of the parents was mapped to the seven populations used in this study: AFR, AMR, EAS, EUR, MEA, OCE and SAS (consistent with the populations of the reference data). For example, Cambodian and Japanese reference individuals were assigned to EAS. Individuals of Philippine ancestry were mapped to EAS because Aeta, Agta, Luzon, Vizayan and most Igorot individuals in the reference table are assigned to EAS, although two Igorot entries are assigned to OCE in the reference data. In most cases, only country information was used for population inference; however, in some circumstances, additional available information (i.e. ethnogeographic information) was given priority over broad country labels. For example, Kurdish Iraq and Assyrian reference individuals are assigned to MEA, and unadmixed Peruvian and other Indigenous American populations are classified to AMR. For one individual (ID27), a tri-population target class (EUR/AMR/AFR) was assigned as its reported ancestry comprised three relevant components due to admixed parents.

Where the reference data contained heterogeneous assignments or no exact country match was available, the closest of the seven model populations was selected and the resulting ambiguity was retained for interpretation. For example, Iranian reference individuals occur under both MEA and EUR, whereas populations such as Ecuador, Chile, Costa Rica and Cuba are not represented directly. Ancestry reported as Turkey (Kurdistan region) is similarly ambiguous: generic Turkey and Turkish reference individuals are assigned to EUR, whereas Kurdish Iraq reference individuals are assigned to MEA. For samples specifically reporting ancestry from the Kurdistan region, the more specific Kurdish reference was prioritized and the respective ancestry was therefore mapped to MEA. Accordingly, these targets represent standardized approximations of the self-reported ancestry within the seven-class framework rather than independent genetic ground truth.

3 Results

3.1 Comparison of classifiers

Ranking of classifiers by $\mathcal{M}$ and AUC

Table 1 summarizes the seven classifiers under the evaluation design described in section 2.4: twenty iterations, six folds each, one threshold per population, fitted on the criterion of section 2.5. All values are computed on the test split only.

Based on the criterion $\mathcal{M}$, the RNB classifier achieved the highest value with $\mathcal{M} = 0.752$. RNB exceeded RAM by 0.012 in $\mathcal{M}$ and by 0.079 in AUC. The same ordering was observed across five seeds, with an average difference of 0.035; see Table S1 in Section S1.8. TVE achieved the highest AUC. However, although TVE showed the strongest separation by AUC (0.893), it scores $\mathcal{M} = 0$, because it never gives an admixed label.

Table 1 Each method at its $\mathcal{M}$ -best acceptance score (see (2)), on the evaluation design of section 2.4: twenty iterations, six folds each, one threshold per population.

Method	Score	r_{IR}	r_{AD}	r_{OR}	$\mathcal{M}$	$\mathcal{A}$	AUC
RNB	s_{lik}	0.768	0.752	0.849	0.752	0.790	0.867
RAM	s_{post}	0.744	0.766	0.740	0.740	0.750	0.788
NB	s_{lik}	0.835	0.582	0.690	0.582	0.702	0.797
AM	s_{lik}	0.783	0.832	0.559	0.559	0.724	0.744
TabPFN	s_{post}	0.791	0.553	0.697	0.553	0.680	0.760
TVP	either	0.332	0.748	0.718	0.332	0.599	0.855
TVE	either	0.721	—	0.929	0	0.550	0.893

Error rates

As a refinement of table 1, Figure 5 shows confusion matrices for all seven methods and the resulting error patterns. The remaining methods show distinct error patterns: TVP assigns 0.513 of the truly in-reference individuals to AD; NB classifies various admixed and out-of-reference individuals in the wrong classes; and AM classifies 0.370 of the out-of-reference individuals to AD, consistent with a limitation of an acceptance rule based on absolute fit. TVE has an empty AD column, because it cannot output the class, hence $\mathcal{M}$ equals zero. TabPFN rejects 0.412 of the admixed individuals outright. This is notable because an increased level of heterozygosity is characteristic for F1 admixed individuals [16], and this signal is available to TabPFN during its in-context learning (defined as training here). RNB and RAM are the only two methods whose correct classification groups (IR/IR, AD/AD and OR/OR) all exceed 0.7. As mentioned before, RNB achieves a higher $\mathcal{M}$ than RAM, with all correctly labelled groups at ≥ 0.75 and a maximum of 0.21 for the true AD classified as OR incorrect group. RAM’s main weakness is the out-of-reference group, which are classified as in-reference in 0.243 of the individuals, in contrast to 0.080 for RNB.

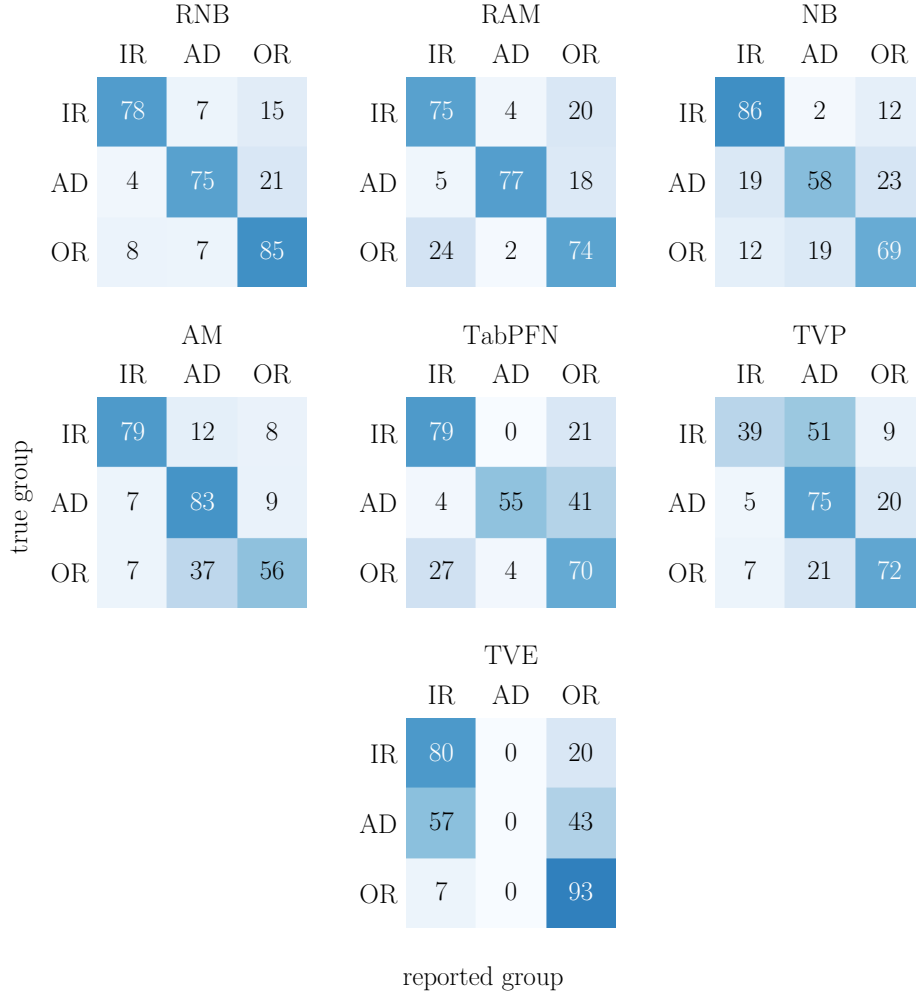


Fig. 5 Confusion matrix showing the reported group (IR, AD, OR) in relation to the true group. Rows: true groups, columns: reported groups. Representation in percent of the row, so each row sums to 100%. Here, the IR/IR cell counts individuals that were kept and not called AD, no matter which population was inferred (in contrast to r_{IR} in table 1), where the correct population must be inferred.

NB and TabPFN were excluded from subsequent analyses as their training on the admixed class led to poor performance, and both methods are unable to identify a specific pair of ancestral populations. TVE was also excluded from subsequent analyses because it cannot classify admixture.

3.2 The acceptance rule and the admixed class

Classification with a reject option is well established outside forensic genetics [7, 22]; within forensic ancestry inference it is rare, the exception being Tvedebrink and Erikssen [60], Tvedebrink et al. [62]. Section 4 discusses our acceptance rule in relation to these approaches. Here, we examine the effect of the acceptance rule on RNB, the best classifier based on $\mathcal{M}$ (results for the other classifiers are in Section S1.9).

The left panel of fig. 6 provides a refined view of the RNB results shown in fig. 5: for in-reference populations, diagonal entries above the line show correctly classified individuals, whereas for out-of-reference populations, the final column below the line shows rejected individuals. AFR shows a high accuracy for both IR (93%) and OR (if AFR is left out, 100%) individuals, whereas some other populations show lower accuracy; e.g., in the MEA class 57% of in-reference individuals were kept, whereas 88% of the out-of-reference individuals were rejected. Note that in practice only the reported class is known, i.e. the column an individual is assigned to. As defined above, each column corresponds to a pool of individuals. Because columns are not normalized to sum to 100%, their composition can be examined separately. Here, the two weakest pools are MEA and OCE: of the individuals the acceptance rule keeps in the MEA pool only 58.1% are truly MEA, and in the OCE pool 62.5%; section 4 reports the corresponding values for all eight pools.

The admixed individuals classified by RNB are examined in more detail (right part of fig. 6). This expands the Admixed (IR) row from the left panel: every individual of that row appears again, those the acceptance rule rejects in the rejected column and the rest at the parent pair they were given. Results are shown after applying the acceptance rule because 20.6% of these individuals are rejected and their parental assignments are therefore not reported by the pipeline. Of the two true parents, both are classified correctly in 64% of all cases and exactly one in 15%; both parents are incorrect in only 0.6% of cases. MEA shows the lowest parental recovery rate: 44% of its parents are recovered, 38% are rejected, and 13% are classified as EUR, which is the pair Section S1.1 measures at $F_{ST} = 0.047$ and section 2.4 therefore holds out together. Individuals from SAS are also often rejected (28%) although the parents are well classified. The same analysis for the other three methods (RAM, AM, TVP) is in Section S1.9.

These results depend on the use of $\mathcal{M}$ as the selection criterion. In addition, results for EUR and MEA depend on combining them for the out-of-reference class; see section 2.4. Additional evaluation designs were considered, and all seven methods were evaluated under each design. More details can be found in Section S1.3.

3.3 Individuals from groups the panel does not contain

The reference table of Ruiz-Ramírez et al. [50] contains a further 800 individuals in seven groups that our reference panel does not contain (section 2.2). Since none of them entered the training/calibration/test data, they can be evaluated directly using the fitted RNB configuration.

Four of the seven groups represent real populations absent from the panel, and only 29.4% of their individuals are rejected; compare with the rejection of 84.9% for

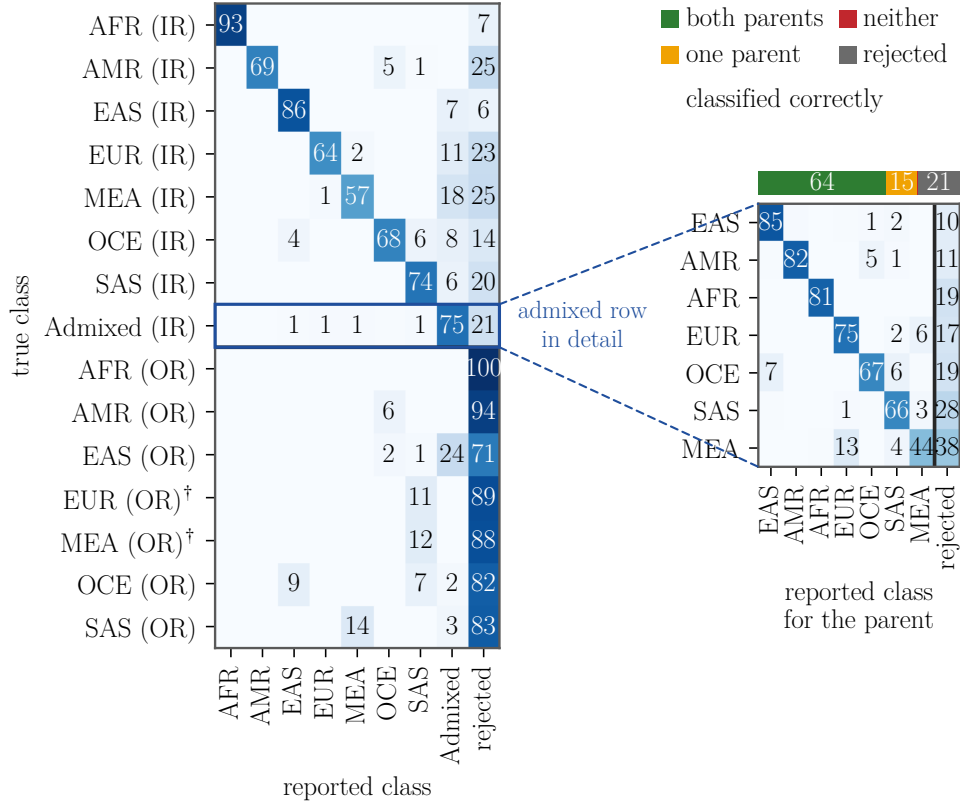


Fig. 6 Performance of the RNB classifier after application of the acceptance rule. *Left*: Rows indicate the true class and columns the reported class; cells below 0.5% are omitted and each row sums to 100% (up to rounding errors). The upper rows show in-reference (IR) individuals from populations included in the panel. The lower rows show individuals from populations treated as out-of-reference (OR) in the corresponding fold; these samples should ideally be assigned to the rejected column. (Recall that EUR and MEA are held out together in a single fold (section 2.4) and are therefore marked with †.) *Right*: Detailed results for the true Admixed (IR) individuals. The horizontal bar summarizes the outcome per individual: both parental populations recovered correctly, one parental population recovered, neither recovered, or the individual rejected by the acceptance rule. The matrix below evaluates the same samples at the level of individual parental components. Each admixed individual therefore contributes two entries, one for each true parental population. Rejected individuals contribute both parental components to the rejected column. Both panels pool the six folds and twenty iterations, with the threshold in each fold and iteration fitted on the corresponding calibration set. For a given population, the IR row is based on the five folds in which that population is part of the reference panel, whereas the OR row is based on the fold in which it is held out.

the populations held out of the panel in the reported design (table 1). This difference is consistent with genetic distance to the reference panel: three of those four groups (north african, central asian, roma) lie close to a panel population (table 2), whereas the fourth (east african) has both the largest F_{ST} and the highest rejection rate.

However, if an individual is not rejected, the reported class often provides a reasonable description relative to the reference panel: 27 of the 29 north african individuals

Table 2 The additional 800 individuals of the reference table, scored by the deployed configuration using the RNB classifier. F_{ST} of the corresponding group is computed against each of the seven panel populations with the estimator of Section S1.1, and the table reports the closest of the seven reference populations. The last column gives the classes that the individuals that passed the acceptance rule received most often. Groups are ordered by F_{ST} within each block.

Group	n	F_{ST}	to	rejected	classes given to the rest
<i>populations absent from the panel</i>					
NORTH AFRICAN	29	0.018	MEA	0.07	MEA 27
CENTRAL ASIAN	64	0.041	MEA	0.02	Admixed 50, EUR 11
ROMA	3	0.046	MEA	0.00	Admixed 3
EAST AFRICAN	47	0.087	MEA	0.83	MEA 7
<i>a panel region from another source</i>					
(SANGER) MIDDLE EAST	136	0.007	MEA	0.25	MEA 90, Admixed 9
<i>cohorts labelled admixed</i>					
ADMIXED AFRICAN	157	0.014	AFR	0.55	AFR 68, Admixed 2
ADMIXED AMERICAN	364	0.047	MEA	0.85	Admixed 27, MEA 19

are given MEA, and 50 of the 64 central asian ones are called admixed, most often as a mixture of EUR with MEA or with SAS. Section S1.11 gives the remaining groups, among them the two cohorts carrying an admixed label, for which a correct answer is not defined.

3.4 Application to newly analyzed samples

Analysis of AIMS

All 28 individuals were analyzed using the modified panel comprising 106 autosomal markers. The marker rs12594144 (part of the 104 VISAGE Enhanced tool) was removed from further analysis because of insufficient coverage in most samples. Furthermore, rs9306481 and rs924381 were not considered for BGA prediction because they were not part of the SNP set used for model development. Thus, a maximum of 103 autosomal markers was available for BGA prediction. Amplification of rs12405776 slightly co-amplified an additional region (chr10:38330354–38330121). However, alignment against the complete reference genome allowed the amplicons to be assigned either to the rs12405776 locus or to the co-amplified region, enabling reliable variant calling of rs12405776. Due to the set quality threshold for final SNP calling, marker numbers ranged 78 to 103 (cf. Table S2).

Classification using the RNB classifier and class-specific acceptance thresholds

All 28 individuals were classified with the recommended RNB classifier using the likelihood score s_{lik} and the class-specific acceptance thresholds described in Section 2.5.

That means a prediction was reported only when its likelihood score passed the threshold calibrated for the corresponding class. Samples that did not pass the relevant threshold were reported as *not classified*.

For evaluation, the reported population assignment was compared with the target derived from the participant information. (fig. 7, details in Table S2). A single-population output k was treated as the pair (k, k) . An *exact* match was defined if both populations were predicted correctly, whereas a *partial* match was assigned when at least one target component was classified correctly. No wrong classifications in respect to both parents were observed. Rejected samples were considered separately as *not classified* remaining an inconclusive result in a forensic case. ID27 was represented by three target populations but was rejected by the acceptance rule and therefore did not require an exact or partial pairwise classification.

Applying the RNB to the 28 individuals, 16 passed the acceptance rule and received a population assignment, whereas 12 were reported as *not classified* (fig. 7). Among the 16 reported classifications, 12 exactly matched the target and four were partial matches (one parent correctly assigned). Rejection and therefore a final non-classification occurred especially in cases of individuals of the AMR class (7 out of 8). This can be partly attributed to the fact, that at the parental level, also partly admixed ancestry is present. Admixed parents can lead to a non-evenly distribution of two populations within the child, e.g. in one of the rejected cases (ID 21), one parent is from Germany (EUR), whereas the other parent itself is an admixed parent from Italy (EUR) and Chile (AMR). Without using the acceptance threshold, only European was classified by the model, missing the American part of the second parent. In the case of ID27 a 3-population target AFR/AMR/EUR was present (parents: Portugal/Brazil (Indigenous) and Germany/Africa). As this is not represented by the model, a rejection occurred as expected ($s_{\text{lik}} = -1.122$, $\text{margin} = -0.2227$). (cf. Table S2). The rejection of a classification is intended to avoid wrong classifications. Of all 12 rejected individuals, 9 individuals would have received a partial wrong classification (EUR was correctly classified in all cases), whereas 3 cases below the acceptance threshold would have had a correct assignment (ID2 EUR, ID3 EAS/EUR, ID16 AMR/EUR). In the case of the four partial classifications, in two cases the second component was missed (ID1, ID20) and in the other two cases, the individual was wrongly classified as being admixed without having a component (based on the self-reported ancestry) (ID10, ID23), respectively. ID1, with reported Australian and Scottish ancestry, was assigned EUR rather than EUR/OCE. Notably, the Australian reference data are themselves heterogeneous, with two Australian-labeled individuals assigned to OCE and one assigned to EUR. ID20, with German and Philippine ancestry (EUR/EAS) was classified as EAS only, whereas ID17 and ID19, with comparable German and Philippine or Cambodian ancestry, were both classified as EAS/EUR. ID10, for which both parental sides were reported as Turkey (Kurdistan region) (MEA only), was classified as EUR/MEA. Given that generic Turkish reference individuals are assigned to EUR whereas Kurdish Iraq is assigned to MEA, this result cannot be interpreted as a straightforward disagreement. In addition, EUR and MEA are the two closest reference populations in the panel, with $F_{ST} = 0.047$. Finally, ID23, with German ancestry on one parental side and an Indigenous-American grandparent on

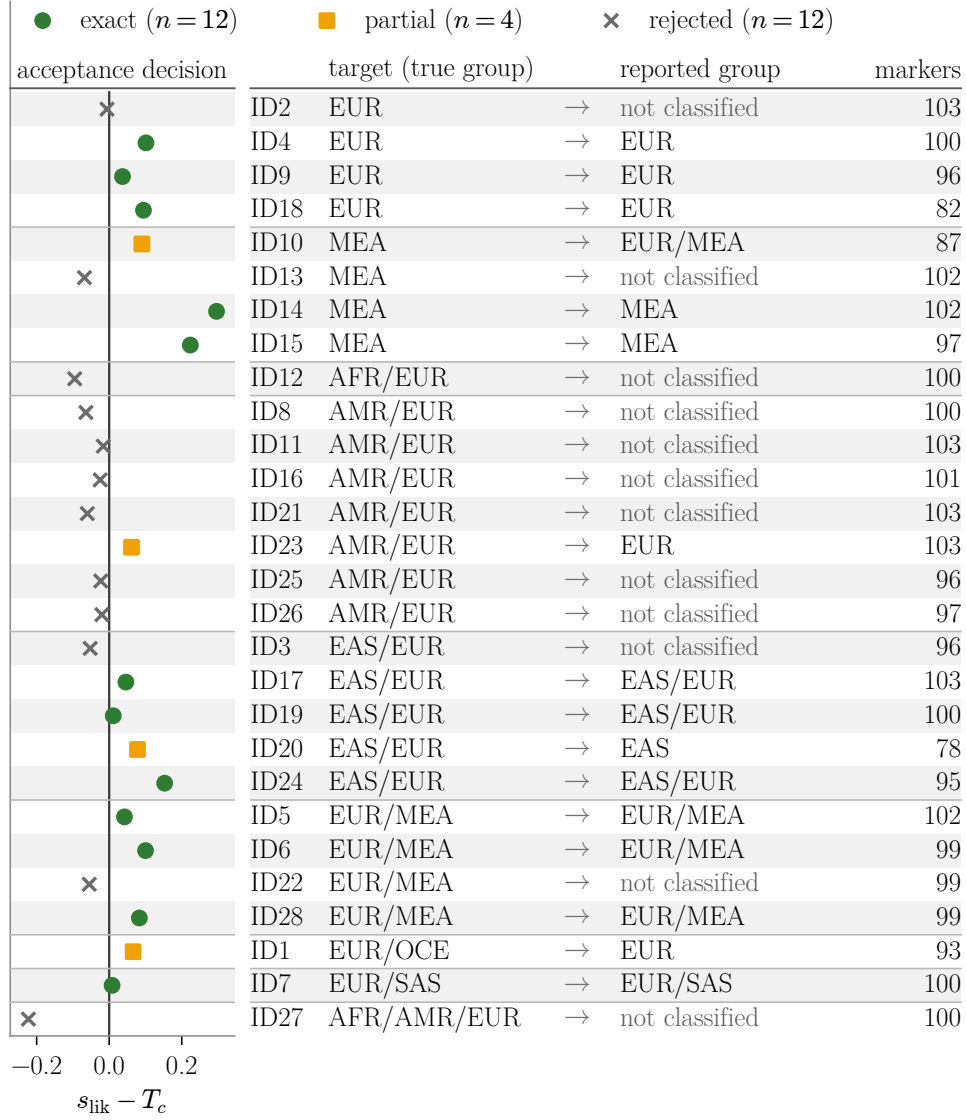


Fig. 7 Classification results for the 28 empirical samples using the RNB classifier and the likelihood-based acceptance rule. The likelihood score relative to the class-specific acceptance threshold, $s_{\text{lik}} - T_c$, is shown for each individual. Values at or above zero pass the acceptance rule, whereas values below zero result in *not classified*. Symbols distinguish exact, partial, and rejected classifications according to the ancestry targets derived from the participant information. For each individual, the target population assignment, reported classifier output, and number of markers available for classification out of the 103 usable markers are shown; the same data are listed in Table S2. Rows are grouped by target, with single-population targets shown first, followed by two-population targets and the three-population target of ID27. Target populations were derived from the self-reported parental ancestry according to the population-group assignments used in the reference data. For two-population targets, the pair is unordered; a pair (k, k) is written as k .

the other, was mapped to AMR/EUR but classified as EUR only. Here, the AMR component was reported at the grandparent rather than parental level, whereas the RNB model represents each parent by one population.

As mentioned, the markers used for classification of the individuals varied. However, no simple relationship was observed between the number of missing SNPs and reduced accuracy (a partially correct result or rejection). Samples with comparatively large numbers of missing markers could still pass the acceptance rule: ID18 was classified exactly as EUR with 21 of the 103 markers missing, and ID20 was classified as EAS with 25 missing. Conversely, several samples with no missing markers, including ID2, ID11 and ID21, were rejected. Only successfully typed markers contribute to s_{lik} , which is normalized by the number $M(x)$ of available markers. The results therefore do not support marker dropout alone as an explanation for the rejected classifications.

4 Discussion

A classifier with an admixed and a reject option

Under the reported design, we recommend the recent naive Bayes classifier with the likelihood score s_{lik} ; the method supports both admixed classifications and rejection. RNB achieved the highest selection criterion, $\mathcal{M} = 0.752$, defined as the minimum of the accuracies for the in-reference, admixed and out-of-reference groups, and an AUC of 0.867. Several additional properties support this choice. For RNB, s_{lik} performed best across all four designs and both evaluation quantities. Its errors are the most evenly spread of all methods, with no off-diagonal cell above 0.21 (fig. 5). It names admixture structurally, as a hypothesis over pairs of populations the two parents come from, so it needs no simulated training offspring. Additionally, s_{lik} measures fit in absolute terms rather than as a posterior normalised over the panel, it can express that none of the panel populations provides an adequate fit. Finally, its estimate is closed form: it evaluates the likelihood of every possible pair and takes the largest, so the prediction is deterministic for a fixed input. However, this recommendation depends on the evaluation design, as the recent admixture model (RAM) performs better in some alternative designs. RAM under s_{post} achieves a higher criterion value in two of the three designs we do not report, leave-one-out without the EUR/MEA pairing and leave-two-out (Section S1.2, discussed in Section S1.3); in the first of the two it leads on the AUC as well. The recommendation is therefore conditional on the choice of reported design (Section S1.2), and would have to be revisited if that choice changed.

Limitations

The approach has several limitations:

1. The admixed class in our calibration/test data consists only of simulated F1 hybrids. No admixed individual in the method comparison is real; this excludes other forms of admixture, including backcrosses and second-generation admixture.
2. The out-of-reference class depends on the genetic distance. Not surprisingly, a nearby missing population is not detectable as out-of-reference. Populations held out of the panel are rejected in 0.849 of cases, but these were distant from the

remaining reference populations. (For this reason, we only left out EUR and MEA in combination.) The case Tvedebrink et al. [62] raises, an individual from a population that the reference does not contain, is therefore a matter of degree.

3. In calibration, the thresholds are fitted under different panel sizes from those used for test data. While thresholds are fitted on a panel of 5 (if EUR and MEA are left out together) or 6 populations during calibration, they are used on 7 populations in the deployed system. The direction of that bias is known: as the number of panel populations increases, the best fit can only improve, so a threshold fitted at six under-rejects at seven.
4. Reference data for AMR (109 samples) and OCE (77 samples) are limited. Such small sample sizes may not represent the full heterogeneity of these populations and should be interpreted accordingly.

An additional limitation concerns the interpretation of predictions in casework. When analyzing a casework sample, the true ancestry is not known, but the reported class is. So, we should be interested in the question: provided we have a sample which is reported to be e.g. MEA, what is the probability that it really comes from MEA? (This is usually called the positive predictive value, PPV.) For the recommended configuration the question based on the pools is given in table 3.

Table 3 The PPV is the fraction of individuals which are true class c among those reported to be class c . (This is a reading down a *column* of fig. 6.) Take MEA, pooled over the six folds, an iteration keeps about 174 individuals in the MEA pool: about 100 are truly MEA, about 50 are South Asian individuals from the one fold in which SAS is held out of the panel — MEA and SAS are the second closest pair in the panel at $F_{ST} = 0.079$ (Section S1.1) — about 12 are admixed and about 10 are European, which gives the 0.581.

Pool	PPV	Pool	PPV	Pool	PPV	Pool	PPV
AFR	0.999	AMR	0.959	SAS	0.816	OCE	0.625
EAS	0.977	EUR	0.964	Admixed	0.752	MEA	0.581

The same question can be put to the rejection: Given a sample is reported OR, how likely is it really OR (negative predictive value, NPV)? Pooled over the six folds and twenty iterations, an iteration rejects about 2168 individuals, of which about 1506 are genuinely out of reference, about 400 are in reference and about 262 are admixed. The negative predictive value is therefore 0.695, against an out-of-reference share of 0.311 in the test part.

Both, the positive and negative predictive value, depend on how common each group is, and in our study this is set by design and not by casework. In other words, for any real population, PPV and NPV will differ from the numbers just reported.

In our analysis, we used separate training, calibration and test sets within each split to avoid reusing the same observations for multiple stages. However, note that the selection criterion $\mathcal{M}$ is used twice: (i) to choose among several (39, to be exact)

configurations formed by seven classifiers, two scores, and three threshold rules; (ii) as the performance we report for the chosen configuration. Consequently, the observed performance of RNB may overestimate what performance the selected configuration would achieve on fresh data [6, 63, see also 30 for the general issue of non-independent selection and subsequent analysis]. Yet another design decision was made after seeing the results: The joint EUR and MEA holdout was not fixed in advance. Rather, it followed from the comparison itself, where holding out either population alone led to its individuals being assigned to the other (section 2.4), irrespective of the classifier. This observation is corroborated by $F_{ST} = 0.047$ for that pair against 0.079 for the next closest (Section S1.1). For these reasons, the recommendation of RNB should be interpreted as conditional on the reported evaluation design. We investigated the robustness of the resulting recommendation in two complementary ways. First, we repeated the complete comparison across five seeds with different random splits and RNB ranked first under $\mathcal{M}$ in every run. This demonstrates stability with respect to the random partition, but not with respect to the panel, since all five runs use the same 3541 individuals at the same 104 markers. Second, the selected RNB configuration was applied, without using either additional dataset for model selection or threshold calibration, to 800 held-out individuals (section 3.3) and 28 newly analyzed samples (section 3.4). These analyses provide additional out-of-sample evidence about the behavior of the deployed classifier, although neither of them yields an $\mathcal{M}$ -value. An evaluation separated from configuration selection could be obtained through nested resampling, with all configuration choices made within each outer development set, or through a further untouched evaluation dataset. Nested evaluation is constrained by the small number of reference populations and the limited sample sizes within some groups, with a single fold containing as few as 39 out-of-reference individuals. Additional independent evaluation would require suitable new data. In principle, following the simultaneous-inference approach to post-selection validity [4], confidence intervals covering the true $\mathcal{M}$ -values of all candidate configurations simultaneously would remain valid for the selected configuration. Although, applying this principle here would require a construction that accounts for our nonlinear criterion and dependencies in the evaluation design.

All analyses in this study are based on Supplementary Table S1A of Ruiz-Ramírez et al. [50]. Two properties of this dataset limit interpretation. The first is its composition. The seven labels are continental aggregates of very unequal size (from 935 EAS individuals down to 77 OCE). The labels are also heterogeneous within groups. Australian-labeled reference individuals are split between OCE and EUR, and individuals from Turkey are assigned EUR whereas individuals from the Kurdish region of Iraq are assigned MEA. These reference assignments affected the interpretation of the 28 samples, in ID1 and ID10 (section 3.4), where disagreement between the target and the reported group may reflect the reference assignment as well as classifier performance. The second property is the marker set. The 104 AIMs of the VISAGE Enhanced Tool were selected to differentiate continental groups, and the F_{ST} matrix of Section S1.1 illustrates this limitation: EUR and MEA at $F_{ST} = 0.047$ are not separable and have to be held out as one fold, and three of the four genuinely absent populations of section 3.3 lie closer to a panel population than EUR and MEA lie to

each other, which helps explain why they are often not rejected. Consequently, the 104 AIMs can only answer *this population cannot be resolved by the panel* (rather than certain absence from the panel). A larger or differently chosen marker set moves that boundary, and how far it can be moved is an active question [41, 45, 49].

Resutik et al. [49] is the study closest to ours: they put GENOGEOGRAPHER to 2820 individuals with a 1900-SNP panel over 21 metapopulations and – as we do – evaluate it on first-generation offspring simulated one allele per parent and marker. Enabling its admixture module, which tests all 210 pairs of reference groups, lowers the share of unassigned individuals from 3.79 % to 0.71 % and of discordant ones from 7.38 % to 1.06 %; but 63 % of the individuals are then assigned to between two and 21 of the 231 candidates at once and counted as concordant if the true one is among them. Remarkably, none of their 300 simulated offspring is left unassigned at all. We note that a classifier is the more practical instrument than multiple tests: a test returns every hypothesis it fails to reject, so its answer is a set that grows with the number of hypotheses tested and with any correction for multiple testing — which Resutik et al. [49] note would be desirable here. A classifier returns one class by construction, so a report either gives one population, a pair of populations, or a rejection, which is the form casework needs.

The newly analyzed samples

Overall, the newly analyzed samples demonstrate both the applicability and the limitations of representing real-world ancestry using seven broad reference populations. When the acceptance rule was applied, the inferred population assignment was generally compatible with the standardized target derived from the participant information, and no accepted classification was completely discordant. At the same time, 12 of 28 samples were not classified, including individuals whose operational target populations were represented in the reference panel. However, the use of the acceptance rule avoided a partially incorrect classification in nine of the 12 samples.

ID27 illustrates the limitations of the two-population strategy particularly clearly. Its paternal ancestry was reported as Portugal/Brazil (Indigenous), whereas its maternal ancestry was reported as Germany/Africa, jointly comprising EUR, AMR and AFR components (cf. Results application). As mentioned, the sample was rejected by the acceptance rule. ID27 had the lowest likelihood score among the 28 empirical samples ($s_{\text{lik}} = -1.122$), consistent with the pronounced complexity of its reported ancestry and the difficulty of representing it within the model’s two-population framework. This case illustrates an inherent limitation of the RNB model for ancestry profiles that cannot be adequately represented by at most two parental populations.

A *not classified* result should therefore not be interpreted as direct evidence that an individual belongs to a population absent from the reference data. Rather, it indicates that the observed genotype did not provide sufficient support for reporting the class selected by the RNB classifier under the calibrated acceptance rule. Complex ancestry, geographically intermediate populations, ambiguities in the reference assignments and the necessary simplification of heterogeneous self-reported ancestry may all contribute to such outcomes.

The reject option

A classifier that may refuse to decide is long established outside forensics. Accordingly, the contribution of this paper is not the reject option itself but its adaptation to the BGA setting: a per-class rather than a global threshold, a likelihood score whose scale is comparable across pools of very different size, and a selection criterion, the minimum of the three rates, that prevents trading one outcome for another. Conformal prediction [65] could provide a route to a stronger statement, of the form that the true population lies in a reported set with probability at least $1 - \alpha$. We did not pursue this approach here. The forensic approaches most closely related to our work are Tvedebrink and Eriksen [60], Tvedebrink et al. [62], implemented in GenoGeographer [61]. These studies address the same setting considered here, an individual from a population the reference does not contain, and use a goodness-of-fit test of the profile against each claimed population, with the individual treated as its own population when nothing fits.

Our approach follows the same general principle: the quantity used for acceptance should measure fit in absolute terms rather than relative to the panel, motivating the use of s_{lik} as the score. Our approach differs in two respects. We construct a classifier rather than a statistical test, and we fit the cut-off separately for each predicted class on a calibration set rather than using a global test level. These differences are reflected in table 1, where TVE and TVP are included in the comparison: TVE is the strongest of all methods at rejecting out-of-reference individuals, at $r_{\text{OR}} = 0.929$, and among the weakest at naming the correct population, at $r_{\text{IR}} = 0.721$, while lacking an admixed class, while TVP names a population correctly for only 0.332 of the in-reference individuals. We find that the design using $\mathcal{M}$ as a selection criterion is more balanced, and identifies RNB as the recommended method.

Supplementary information. The appendices are supplied as Online Resource 1 (supplement_sn.pdf), whose sections are numbered S1, S2 and so on and are referred to by those numbers throughout.

Declarations

ORCID. Jana Naue 0000-0001-9471-9013, Zoe Lange 0009-0001-1867-5271, Holger Dette 0000-0001-7048-474X, Peter Pfaffelhuber 0000-0002-6421-5460, Carola Sophia Heinzel 0009-0000-9042-949X. ORCID ERGAENZEN fuer Julika Schmitz, Lukas Hoffmann, Ferdinand Suchanek und Sabine Lutz-Bonengel.

Data Availability. Due to data protection, underlying SNP data of the 28 individuals are not shown. They can be obtained on reasonable request.

The code to reproduce our results and to apply our methods to new data is available at <https://github.com/pfaffelh/BGA-pairs-and-reject>. In particular, we provide a stand-alone Python script which produces a report on a novel genotype of an individual.

Acknowledgments. The authors thank Timo Sanger for support with labwork.

Funding. This work was funded by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) – Project-ID 499552394 – SFB Small Data.

Use of Generative AI. Large language models were used within the study and for manuscript preparation. Anthropic’s Claude was used as a coding and writing assistant (wrote and ran parts of the evaluation and figure code, checked reported numbers against the pipeline output, and revised passages of the text). Furthermore, Google’s (DeepMind) Gemma 4 and OpenAI’s GPT-5.6 Sol were used for improvement of phrasing. Every analysis, every number and every claim was specified, checked and approved by the authors, who take full responsibility for the content.

Author Contributions. Conceptualization: Carola Heinzl, Sabine Lutz-Bonengel, Peter Pfaffelhuber

Funding Acquisition: Sabine Lutz-Bonengel, Peter Pfaffelhuber, Holger Dette

Investigation: Lukas Hoffmann, Jana Naue, Ferdinand Suchanek

Formal analysis: Carola Heinzl, Peter Pfaffelhuber, Julika Schmitz

Software: Carola Heinzl, Peter Pfaffelhuber, Julika Schmitz

Validation: Carola Heinzl, Jana Naue, Peter Pfaffelhuber, Julika Schmitz

Writing -Original Draft: Carola Heinzl, Sabine Lutz-Bonengel, Jana Naue, Peter Pfaffelhuber, Julika Schmitz

Writing - Review & Editing: all authors

Supervision: Holger Dette, Carola Heinzl, Sabine Lutz-Bonengel, Jana Naue, Peter Pfaffelhuber

Declaration of Conflicts of Interest. The authors declare that there are no conflicts of interest.

Ethics approval. All procedures performed in studies involving human tissue were in concordance with the ethical standards of the institutional and/ or national research committee and with the 1964 Helsinki declaration and its later amendments or comparable ethical standards. Participants gave informed consent and approval was granted by the Ethics Committee of the University of Freiburg (22-1427-S1). This article does not contain any studies with animals performed by any of the authors.

Clinical trial number. not applicable

References

- [1] Alexander DH, Novembre J, Lange K (2009) Fast model-based estimation of ancestry in unrelated individuals. *Genome research* 19(9):1655–1664. <https://doi.org/10.1101/gr.094052.109>
- [2] Almarri MA, Haber M, Lootah RA, et al (2021) The genomic history of the middle east. *Cell* 184(18):4612–4625
- [3] Bartlett PL, Wegkamp MH (2008) Classification with a reject option using a hinge loss. *Journal of Machine Learning Research* 9:1823–1840

- [4] Berk R, Brown L, Buja A, et al (2013) Valid post-selection inference. *The Annals of Statistics* 41(2):802–837. <https://doi.org/10.1214/12-AOS1077>
- [5] Butler K, Peck M, Hart J, et al (2011) Molecular “eyewitness”: forensic prediction of phenotype and ancestry. *Forensic Science International: Genetics Supplement Series* 3(1):e498. <https://doi.org/10.1016/j.fsigss.2011.09.109>
- [6] Cawley GC, Talbot NLC (2010) On over-fitting in model selection and subsequent selection bias in performance evaluation. *Journal of Machine Learning Research* 11:2079–2107
- [7] Chow CK (1970) On optimum recognition error and reject tradeoff. *IEEE Transactions on Information Theory* 16(1):41–46. <https://doi.org/10.1109/TIT.1970.1054406>
- [8] Cortes C, DeSalvo G, Mohri M (2016) Learning with rejection. In: *Algorithmic Learning Theory (ALT 2016)*, Lecture Notes in Computer Science, vol 9925. Springer, pp 67–82, https://doi.org/10.1007/978-3-319-46379-7_5
- [9] Dette H, Heinzl CS, Lange Z, et al (2026) Testing for single-population ancestry in the admixture model. *arXiv preprint arXiv:260601990*
- [10] Diepenbroek M, Bayer B, Schwender K, et al (2020) Evaluation of the ion ampliseq phenotrivium panel: Mps-based assay for ancestry and phenotype predictions challenged by casework samples. *Genes* 11(12):1398. <https://doi.org/10.3390/genes11121398>
- [11] Eduardoff M, Gross TE, Santos C, et al (2016) Inter-laboratory evaluation of the eurofor gen global ancestry-informative snp panel by massively parallel sequencing using the ion pgm. *Forensic Science International: Genetics* 23:178–189. <https://doi.org/10.1016/j.fsigen.2016.04.008>
- [12] Escher LM, Naslavsky MS, Scliar MO, et al (2022) Challenges in selecting admixture models and marker sets to infer genetic ancestry in a brazilian admixed population. *Scientific Reports* 12(1):21240. <https://doi.org/10.1038/s41598-022-25521-7>
- [13] Esposito C, Landrum GA, Schneider N, et al (2021) Ghost: adjusting the decision threshold to handle imbalanced data in machine learning. *Journal of Chemical Information and Modeling* 61(6):2623–2640
- [14] Fleckhaus J, Anslinger K, Lutz-Bonengel S, et al (2025) Aktueller Stand der “erweiterten DNA-Analyse”. *Rechtsmedizin* 35:405–413. <https://doi.org/10.1007/s00194-025-00768-2>
- [15] Freire-Aradas A, Phillips C, Lareu MV (2017) Forensic individual age estimation with dna: From initial approaches to methylation tests. *Forensic Science Review*

- [16] Garnier-Géré P, Chikhi L (2013) Population subdivision, Hardy–Weinberg equilibrium and the Wahlund effect. In: Encyclopedia of Life Sciences (eLS). Wiley, <https://doi.org/10.1002/9780470015902.a0005446.pub3>, covers both directions: the heterozygote deficit of a pooled structured sample, and the heterozygote excess of first-generation offspring of two differentiated populations
- [17] Geifman Y, El-Yaniv R (2017) Selective classification for deep neural networks. Advances in neural information processing systems 30
- [18] Geng C, Huang SJ, Chen S (2021) Recent advances in open set recognition: A survey. IEEE Transactions on Pattern Analysis and Machine Intelligence 43(10):3614–3631. <https://doi.org/10.1109/TPAMI.2020.2981604>
- [19] Heinzel CS, Purucker L, Hutter F, et al (2025) Advancing biogeographical ancestry predictions through machine learning. Forensic Science International: Genetics p 103290. <https://doi.org/10.1016/j.fsigen.2025.103290>
- [20] Hendrickx K, Perini L, Van der Plas D, et al (2024) Machine learning with a reject option: A survey. Machine Learning 113(5):3073–3110. <https://doi.org/10.1007/s10994-024-06534-x>
- [21] Hendrycks D, Gimpel K (2017) A baseline for detecting misclassified and out-of-distribution examples in neural networks. In: International Conference on Learning Representations
- [22] Herbei R, Wegkamp MH (2006) Classification with reject option. The Canadian Journal of Statistics 34(4):709–721. <https://doi.org/10.1002/cjs.5550340410>
- [23] Hollmann N, Müller S, Eggenesperger K, et al (2023) TabPFN: A transformer that solves small tabular classification problems in a second. In: Proceedings of the 11th International Conference on Learning Representations (ICLR), [arXiv:2207.01848](https://arxiv.org/abs/2207.01848)
- [24] Hollmann N, Müller S, Purucker L, et al (2025) Accurate predictions on small data with a tabular foundation model. Nature 637(8045):319–326. <https://doi.org/10.1038/s41586-024-08328-6>
- [25] Kayser M (2015) Forensic dna phenotyping: Predicting human appearance from crime scene material for investigative purposes. Forensic Science International: Genetics 18:33–48. <https://doi.org/10.1016/j.fsigen.2015.02.003>
- [26] Kayser M, de Knijff P (2011) Improving human forensics through advances in genetics, genomics and molecular biology. Nature Reviews Genetics 12(3):179–192. <https://doi.org/10.1038/nrg2952>

- [27] Kayser M, Branicki W, Parson W, et al (2023) Recent advances in forensic dna phenotyping of appearance, ancestry and age. *Forensic Science International: Genetics* 65:102870. <https://doi.org/10.1016/j.fsigen.2023.102870>
- [28] Kidd KK, Speed WC, Pakstis AJ, et al (2014) Progress toward an efficient panel of snps for ancestry inference. *Forensic Science International: Genetics* 10:23–32. <https://doi.org/10.1016/j.fsigen.2014.01.002>
- [29] Kosoy R, Nassir R, Tian C, et al (2009) Ancestry informative marker sets for determining continental origin and admixture proportions in common populations in america. *Human Mutation* 30(1):69–78. <https://doi.org/10.1002/humu.20822>
- [30] Kriegeskorte N, Simmons WK, Bellgowan PSF, et al (2009) Circular analysis in systems neuroscience: the dangers of double dipping. *Nature Neuroscience* 12(5):535–540. <https://doi.org/10.1038/nn.2303>
- [31] Masana M, Ruiz I, Serrat J, et al (2018) Metric learning for novelty and anomaly detection. *arXiv preprint arXiv:180805492*
- [32] McNevin D (2020) Forensic inference of biogeographical ancestry from genotype: The genetic ancestry lab. *Wiley Interdisciplinary Reviews: Forensic Science* 2(2):e1356. <https://doi.org/10.1002/wfs2.1356>
- [33] Mogensen HS, Tvedebrink T, Børsting C, et al (2020) Ancestry prediction efficiency of the software genogeographer using a z-score method and the ancestry informative markers in the precision id ancestry panel. *Forensic Science International: Genetics* 44:102154. <https://doi.org/10.1016/j.fsigen.2019.102154>
- [34] Naue J (2023) Getting the chronological age out of dna: using insights of age-dependent dna methylation for forensic dna applications. *Genes & Genomics* 45(10):1239–1261. <https://doi.org/10.1007/s13258-023-01392-8>
- [35] Pereira R, Phillips C, Pinto N, et al (2012) Straightforward inference of ancestry and admixture proportions through ancestry-informative insertion deletion multiplexing. *PLoS One* 7(1):e29684. <https://doi.org/10.1371/journal.pone.0029684>
- [36] Pereira V, Freire-Aradas A, Ballard D, et al (2019) Development and validation of the eurofor gen name (north african and middle eastern) ancestry panel. *Forensic Science International: Genetics* 42:260–267. <https://doi.org/10.1016/j.fsigen.2019.06.010>
- [37] Pereira V, Santangelo R, Børsting C, et al (2020) Evaluation of the precision of ancestry inferences in south american admixed populations. *Frontiers in Genetics* 11:966. <https://doi.org/10.3389/fgene.2020.00966>
- [38] Pfaffelhuber P, Grundner-Culemann F, Lipphardt V, et al (2020) How to choose sets of ancestry informative markers: A supervised feature selection approach.

- Forensic Science International: Genetics 46:102259. <https://doi.org/10.1016/j.fsigen.2020.102259>
- [39] Pfaffelhuber P, Sester-Huss E, Baumdicker F, et al (2022) Inference of recent admixture using genotype data. Forensic Science International: Genetics 56:102593. <https://doi.org/10.1016/j.fsigen.2021.102593>
 - [40] Phillips C (2015) Forensic genetic analysis of bio-geographical ancestry. Forensic Science International: Genetics 18:49–65. <https://doi.org/10.1016/j.fsigen.2015.05.012>
 - [41] Phillips C, de la Puente M (2021) The analysis of ancestry with small-scale forensic panels of genetic markers. Emerging Topics in Life Sciences 5(3):443–453. <https://doi.org/10.1042/ETLS20200327>
 - [42] Phillips C, Salas A, Sanchez J, et al (2007) Inferring ancestral origin using a single multiplex assay of ancestry-informative marker snps. Forensic Science International: Genetics 1(3-4):273–280. <https://doi.org/10.1016/j.fsigen.2007.06.008>
 - [43] Pritchard JK, Stephens M, Donnelly P (2000) Inference of population structure using multilocus genotype data. Genetics 155(2):945–959. <https://doi.org/10.1093/genetics/155.2.945>
 - [44] de la Puente M, Ruiz-Ramírez J, Ambroa-Conde A, et al (2021) Development and evaluation of the ancestry informative marker panel of the visage basic tool. Genes 12(8):1284. <https://doi.org/10.3390/genes12081284>
 - [45] de la Puente M, Casanova-Adán L, González-Bao J, et al (2025) Evaluating the effect of marker panel sizes on estimation of bio-geographical co-ancestry proportions. Forensic Science International: Genetics 78:103275. <https://doi.org/10.1016/j.fsigen.2025.103275>
 - [46] Rajeevan H, Soundararajan U, Pakstis AJ, et al (2012) Introducing the forensic research/reference on genetics knowledge base, frog-kb. Investigative Genetics 3:18. <https://doi.org/10.1186/2041-2223-3-18>
 - [47] Rajeevan H, Soundararajan U, Pakstis AJ, et al (2020) Frogancestrycalc: a standalone batch likelihood computation tool for ancestry inference panels catalogued in frog-kb. Forensic Science International: Genetics 46:102237. <https://doi.org/10.1016/j.fsigen.2020.102237>
 - [48] Resutik P, Aeschbacher S, Krützen M, et al (2023) Comparative evaluation of the maplex, precision id ancestry panel, and visage basic tool for biogeographical ancestry inference. Forensic Science International: Genetics 64:102850. <https://doi.org/10.1016/j.fsigen.2023.102850>

- [49] Resutik P, Gysi M, Lemmin T, et al (2026) From unsupervised selection of aims to likelihood-based bga inference: Challenges revealed by openadmixture and genogeographer. *Forensic Science International: Genetics* 83:103433. <https://doi.org/10.1016/j.fsigen.2026.103433>
- [50] Ruiz-Ramírez J, de la Puente M, Xavier C, et al (2023) Development and evaluations of the ancestry informative markers of the VISAGE Enhanced Tool for Appearance and Ancestry. *Forensic Science International: Genetics* 64:102853. <https://doi.org/10.1016/j.fsigen.2023.102853>
- [51] Samuel G, Prainsack B (2019) Forensic dna phenotyping in europe: views “on the ground” from those who have a professional stake in the technology. *New Genetics and Society* 38(2):119–141. <https://doi.org/10.1080/14636778.2018.1549984>
- [52] Santos C, Phillips C, Gomez-Tato A, et al (2016) Inference of ancestry in forensic analysis ii: Analysis of genetic data. In: *Methods in Molecular Biology*, vol 1420. p 255–285, https://doi.org/10.1007/978-1-4939-3597-0_19
- [53] Scheirer WJ, Rocha AdR, Sapkota A, et al (2013) Toward open set recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 35(7):1757–1772. <https://doi.org/10.1109/TPAMI.2012.256>
- [54] Schneider PM, Prainsack B, Kayser M (2019) The use of forensic dna phenotyping in predicting appearance and biogeographic ancestry. *Deutsches Ärzteblatt International* 116(51-52):873–880. <https://doi.org/10.3238/arztebl.2019.0873>
- [55] Shriver MD, Smith MW, Jin L, et al (1997) Ethnic-affiliation estimation by use of population-specific dna markers. *American Journal of Human Genetics* 60(4):957–964
- [56] Smyser TJ, Pfaffelhuber P, Giglio RM, et al (2024) Probabilistic genetic identification of wild boar hybridization to support control of invasive wild pigs (*sus scrofa*). *Ecosphere* 15(2):e4774. <https://doi.org/10.1002/ecs2.4774>
- [57] Terrado-Ortuño N, May P (2024) Forensic dna phenotyping: a review on snp panels, genotyping techniques, and prediction models. *Forensic Sciences Research* 10(1):owae013. <https://doi.org/10.1093/fsr/owae013>
- [58] Tillmar A, Kling D (2025) Comparative study of statistical approaches and snp panels to infer distant relationships in forensic genetics. *Genes* 16(2):114. <https://doi.org/10.3390/genes16020114>
- [59] Truelsen D, Tvedebrink T, Mogensen HS, et al (2021) Assessment of the effectiveness of the eurofor name and precision id ancestry panel markers for ancestry investigations. *Scientific Reports* 11(1):18595. <https://doi.org/10.1038/s41598-021-97654-0>

- [60] Tvedebrink T, Eriksen PS (2019) Inference of admixed ancestry with ancestry informative markers. *Forensic Science International: Genetics* 42:147–153. <https://doi.org/10.1016/j.fsigen.2019.06.013>
- [61] Tvedebrink T, Eriksen PS, Mogensen HS, et al (2017) Genogeographer – a tool for genogeographic inference. *Forensic Science International: Genetics Supplement Series* 6:e463–e465. <https://doi.org/10.1016/j.fsigs.2017.09.196>
- [62] Tvedebrink T, Eriksen PS, Mogensen HS, et al (2018) Weight of the evidence of genetic investigations of ancestry informative markers. *Theoretical Population Biology* 120:1–10. <https://doi.org/10.1016/j.tpb.2017.12.004>
- [63] Varma S, Simon R (2006) Bias in error estimation when using cross-validation for model selection. *BMC Bioinformatics* 7:91. <https://doi.org/10.1186/1471-2105-7-91>
- [64] Vidaki A, Kayser M (2018) Recent progress, methods and perspectives in forensic epigenetics. *Forensic Science International: Genetics* 37:180–195. <https://doi.org/10.1016/j.fsigen.2018.08.008>
- [65] Vovk V, Gammerman A, Shafer G (2005) *Algorithmic Learning in a Random World*. Springer, New York, <https://doi.org/10.1007/b106715>
- [66] Wang S, Lewis CM, Jakobsson M, et al (2007) Genetic variation and population structure in native americans. *PLoS Genetics* 3(11):e185. <https://doi.org/10.1371/journal.pgen.0030185>
- [67] Wu M, Ye J (2009) A small sphere and large margin approach for novelty detection using training data with outliers. *IEEE transactions on pattern analysis and machine intelligence* 31(11):2088–2092
- [68] Xavier C, de la Puente M, Phillips C, et al (2020) Forensic evaluation of the asia pacific ancestry-informative maplex assay. *Forensic Science International: Genetics* 48:102344. <https://doi.org/10.1016/j.fsigen.2020.102344>
- [69] Zieger M (2022) Forensic dna phenotyping in europe: How far may it go? *Journal of Law and the Biosciences* 9(2):lsac024. <https://doi.org/10.1093/jlb/lsac024>

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- [supplementsn.pdf](#)