

Supplement to: 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 Heinzel ^{[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

This is the supplementary material (Online Resource 1) to the article named in the title. It states the pipeline in full, including the alternatives that were run and rejected, and gives the likelihoods of the compared methods. References to figures, tables and sections without an S refer to the main article.

S1 More details on the pipeline

This appendix provides additional details on the design decisions underlying section 2.4 and section 2.5 of the article, including alternative approaches that were evaluated but not used for the reported analysis. The aim is to document the analysis sufficiently for the reported results to be reproduced.

S1.1 Divergence between the reference populations

Pairwise F_{ST} values over the 104 markers were calculated using the Hudson estimator [4], in its multiallelic form and as a ratio of averages over markers, as recommended by Bhatia et al. [3], with the $n/(n-1)$ correction for within-population heterozygosity:

	AFR	AMR	EAS	EUR	MEA	OCE	SAS
AFR	—	0.456	0.406	0.426	0.307	0.341	0.304
AMR	0.456	—	0.182	0.372	0.313	0.243	0.226
EAS	0.406	0.182	—	0.349	0.278	0.135	0.171
EUR	0.426	0.372	0.349	—	0.047	0.327	0.135
MEA	0.307	0.313	0.278	0.047	—	0.252	0.079
OCE	0.341	0.243	0.135	0.327	0.252	—	0.144
SAS	0.304	0.226	0.171	0.135	0.079	0.144	—

The pairwise F_{ST} values range from 0.047 to 0.456, with a median of 0.278. These values are high in absolute terms, as expected because the 104 markers were selected for ancestry informativeness. Two pairwise comparisons are particularly relevant for the subsequent analysis. EUR and MEA show the smallest differentiation ($F_{ST} = 0.047$) by a clear margin, followed by MEA and SAS ($F_{ST} = 0.079$). The low differentiation between EUR and MEA motivates the paired hold-out fold described in section 2.4 of the article. AFR and AMR show the largest differentiation ($F_{ST} = 0.456$).

S1.2 Three designs

We evaluated three designs for constructing the out-of-reference class:

D1, leave-k-out

Each fold removes k populations from the panel and treats their individuals as out-of-reference, so the reference panel in that fold contains $7 - k$ populations. An iteration runs all $\binom{7}{k}$ folds and pools their predictions. At $k = 1$ ($k = 2$), the same individual is treated as in-reference in six (fifteen) of the seven (twenty-one) folds and as out-of-reference in one (six), while its genotype remains unchanged.

D2, the reported design

D2 follows D1 at $k = 1$ with one modification: the fold that would hold out EUR and the fold that would hold out MEA are replaced by a single fold holding out both, resulting in six folds. The rationale is described in section 2.4 of the article. All other aspects of D1 remain unchanged.

D3, synthetic out-of-reference

The panel is not reduced: all seven populations remain available, and the out-of-reference individuals are generated synthetically. Both synthetic generators use the same basic step: for ancestor k and marker m with reference frequencies p_{km} , a new

population is drawn from the Balding–Nichols model,

$$\tilde{p}_{km} \sim \text{Dirichlet}\left(p_{km} \frac{1-F_{ST}}{F_{ST}}\right),$$

which reduces to the usual Beta distribution for two alleles and extends to multiallelic markers. Individuals are then generated under Hardy–Weinberg equilibrium from $\tilde{p}_{km}$. We consider two variants:

Mixed. For each individual, we select m ancestral populations, draw weights $q_i \sim \text{Dirichlet}(b, \dots, b)$ for these populations, and form the mixed panel frequencies $p_{im}^{\text{mix}} = \sum_k q_{ik} p_{km}$. A Balding–Nichols step is then applied before generating one individual. Thus, out-of-reference individuals are generated from different simulated populations. With $m = 7$, mixtures lie at the centre of the simplex and are the easiest to detect, whereas with $m = 2$ the new population is derived from only two ancestral populations and is substantially more difficult to distinguish.

Drifted. One Balding–Nichols frequency vector is generated for each ancestral population, and all out-of-reference individuals assigned to that population are simulated from this vector. This variant isolates the effect of F_{ST} on the divergence between the simulated population and its ancestor.

Reasons to report D2

The designs address different questions and should not be interpreted as interchangeable parameter settings. D2 evaluates performance on observed individuals from populations excluded from the reference panel, whereas D3 evaluates performance under a synthetic model whose results depend on the chosen simulation parameters. Across the mixture settings considered in D3, the accuracy ranges from 0.51 to 0.89, and the ranking of the methods can change. A result such as $\mathcal{M} = 0.75$ for observed held-out populations therefore has a different interpretation from a value obtained under a synthetic generator. We therefore use D2 as the primary evaluation design. D3 is retained as a complementary analysis because it is the only design that keeps all seven populations in the reference panel, as in deployment, and the only design in which the degree of population novelty is a controlled variable.

S1.3 What the other designs show

The reported design holds out populations one at a time, with EUR and MEA held out together (section 2.4 of the article); section S1.2 above gives the other constructions in full. All seven methods were evaluated under each of the four. This subsection summarizes why the remaining evaluation settings were not selected as the primary reported design.

The synthetic design

It leaves the panel complete and draws its out-of-reference individuals from an in silico model instead, with one allele-frequency vector per simulated population drawn from a Balding–Nichols distribution [2]; the construction is D3 above. We do not use this as the primary design because the degree and structure of novelty are imposed by the simulation rather than determined by observed populations.

Leave-two-out

Holding out two populations at once, rather than one, reduces the seven-population reference panel to five populations, removing nearly a third of the panel and therefore providing a substantially smaller reference than a single-population hold-out.

Leave-one-out without the pairing

As described above, EUR and MEA ($F_{ST} = 0.047$, section S1.1) are held out only together. Because these populations are closely related, holding out either one alone confounds the remaining population with the rejection option. Holding them out jointly reflects the effective resolution of the panel as six separable classes rather than seven, instead of evaluating a distinction that is not supported by the data.

The score

For the RNB, all four evaluated designs (D1 with $k = 1$ and $k = 2$, D2, and D3) favour the likelihood score s_{lik} over s_{post} for both reported quantities, with AUC differences of 0.12 to 0.28. In the reported design, it is 0.867 against 0.587 on the AUC and 0.752 against 0.533 on the $\mathcal{M}$ -criterion. For RAM, the second-ranked method, the two reported quantities favour different scores. Its AUC is higher for s_{lik} (0.834 versus 0.788), whereas its $\mathcal{M}$ -criterion value is higher for s_{post} (0.740 versus 0.482), a difference of 0.26. This disagreement between the two evaluation quantities provides no consistent basis for recommending one of the two scores for RAM.

S1.4 The threshold rule and the alternatives

Three criteria, nested

We considered three nested criteria for defining which individuals should be retained within a predicted pool: whether the predicted class is correct, whether the broader group is correct, or whether the individual belongs to the reference set at all. For an individual predicted as MEA:

Criterion	truth of an individual in the MEA pool			
	MEA	other panel pop.	admixed	out-of-reference
name (in use)	keep	reject	reject	reject
coarse group	keep	keep	neither	reject
reference membership	keep	keep	keep	reject

The distinction between these criteria is important. Under the coarse-group criterion, an in-reference individual retained under the wrong population is still counted as correct, whereas an admixed individual assigned to a population pool contributes to neither side of the threshold objective. Consequently, retaining or rejecting such an individual does not affect the fitted threshold.

Three weightings

Within each pool, the two rates defined in section 2.5 of the article are combined in one of three ways: weighting them according to the number of individuals on each side (`name_count`, equivalent to the overall proportion correctly handled by the threshold), weighting them equally (`name_bal`), or taking their minimum (`name_min`). All three criteria combine the same two rates but assign different importance to imbalances between them. In particular, `name_count` can achieve a high value at the expense of the smaller group, whereas `name_min` is determined by the weaker of the two rates.

`name_bal` is used throughout the reported analysis. Although `name_count` yields higher $\mathcal{M}$ -criterion values in some settings, by up to 0.15—more than most differences between methods—this occurs at the expense of individual predicted classes: it results in 94 of 2016 class-iteration combinations with no retained individuals, compared with none for `name_bal`, and results in no accepted predictions for the Admixed class for one method. Thus, `name_count` increases the overall criterion by leaving some classes without accepted predictions, which is precisely the behaviour the criterion was intended to avoid. `name_min` is stricter still with respect to this imbalance and reduces performance on other quantities. Consequently, criterion values obtained with different threshold-weighting rules are not directly comparable.

S1.5 How sharply the threshold is determined

A fitted threshold is informative only if the objective it maximises has a sufficiently well-defined maximum. Figure S1 shows the two components of this objective as the threshold is varied for the reported configuration.

S1.6 Quantities that are not reported

Four additional quantities were computed but are not reported as primary metrics. Three are discussed here because they are natural alternatives to the quantities used in the main analysis.

Overall accuracy, with each individual weighted equally, depends on the class composition of the test set. A method that accepts nearly all individuals can therefore obtain a high value when the test set is dominated by individuals that should be retained. The accuracy $\mathcal{A}$ of section 2.4 of the article is the same quantity with the three groups weighted equally.

Balanced accuracy for retained versus rejected individuals merges in-reference and admixed individuals into a single positive group and therefore does not evaluate whether the retained class label is correct. For example, an admixed individual retained under a population label and an in-reference individual retained under Admixed both count as true positives.

Accuracy over all nine outcomes – the seven populations, Admixed, and rejection – assigns equal weight to each outcome. Consequently, Admixed and rejection each contribute only one ninth of the total. Under the reported design, this weighting changes the ranking substantially and places a method with $\mathcal{M} = 0.124$ in the middle of the comparison. More generally, a per-class mean across seven population classes,

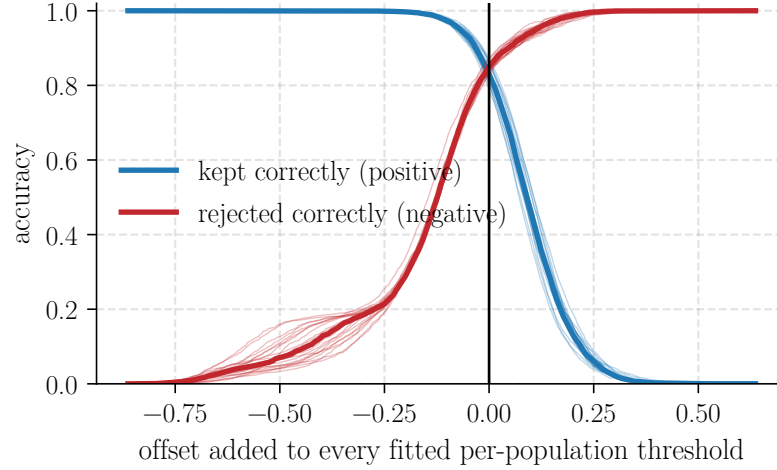


Fig. S1 In each iteration, a class-specific threshold is determined. Each thin line corresponds to one of the twenty iterations, and the heavy line shows their median. The blue curve shows the proportion of individuals that should be retained and are retained, whereas the red curve shows the proportion that should be rejected and are rejected. On the x axis, the fitted threshold of each individual's predicted class has been subtracted from its score. Zero (vertical line) therefore denotes the fitted operating point, and the axis represents a common offset from that point.

one admixed class, and rejection gives relatively little weight to the two outcomes that are central to the present pipeline.

Taking the minimum over populations is also problematic because, in practice, the lowest values largely reflect the trade-off between MEA and EUR, while three populations fall below the required $n \geq 50$ sample-size threshold and would be excluded entirely. The resulting minimum would therefore be taken over only six of the nine possible outcomes. Population-specific behaviour is instead examined separately in fig. 6 of the article and in the analysis below.

S1.7 Two checks the criterion cannot make

The global criterion does not capture either of the following checks for the same reason: small predicted pools contribute in proportion to their size. OCE, for example, contains 77 of the 3541 reference individuals, so poor performance for this population has only a small effect on the global criterion while potentially affecting the entire OCE pool.

A first additional check is whether any predicted class becomes unreachable after thresholding. The three global rates do not distinguish the weighting rules on this property. Across the eight output classes, however, the ordering is reversed: the weighting rules with the highest criterion values are those that leave classes with no accepted predictions. The separable, pool-specific weighting rules cannot fail in this way because each pool is fitted using its own observations.

A second check concerns the composition of the accepted part of each pool. Restricted to individuals retained by the threshold, prec_c is the proportion whose correct class is c . This quantity describes the reliability of an emitted class label and is

therefore directly relevant to casework interpretation. It is the precision counterpart to r_{IR} , which instead describes what happens to an individual of known type. The values for the recommended configuration are tabulated in section 4 of the article.

S1.8 How much the numbers move between runs

All values in section 3 of the article are based on one run comprising twenty iterations with seed 42. The seed determines both the twenty train/calibration/test splits and, for the two admixture models, the initialization of the fitting procedure. To assess sensitivity to these random choices, the full comparison was repeated for five seeds separated by increments of twenty, such that no two runs use the same split.

Table S1 Variation of the reported quantities across runs. The full comparison was repeated for five seeds, from 42 to 122 in steps of 20, such that no two runs use the same split. For each method, the first line gives the mean across the five runs and the second line gives the corresponding range. Every quantity is defined in (1) and (2) of the article; TVE has no admixed class, so its r_{AD} is empty rather than zero and its criterion is 0 by definition.

	r_{IR}	r_{AD}	r_{OR}	$\mathcal{M}$	AUC
RNB	0.770 (0.764–0.779)	0.762 (0.752–0.776)	0.843 (0.836–0.849)	0.760 (0.752–0.766)	0.877 (0.867–0.888)
RAM	0.752 (0.737–0.764)	0.765 (0.758–0.772)	0.727 (0.709–0.743)	0.726 (0.709–0.740)	0.786 (0.782–0.789)
NB	0.839 (0.835–0.842)	0.562 (0.483–0.608)	0.697 (0.678–0.723)	0.562 (0.483–0.608)	0.796 (0.790–0.799)
AM	0.786 (0.781–0.791)	0.837 (0.827–0.853)	0.554 (0.543–0.562)	0.554 (0.543–0.562)	0.743 (0.734–0.751)
TabPFN	0.787 (0.784–0.791)	0.545 (0.528–0.553)	0.690 (0.677–0.698)	0.545 (0.528–0.553)	0.761 (0.755–0.766)
TVP	0.330 (0.327–0.333)	0.748 (0.724–0.765)	0.718 (0.702–0.740)	0.330 (0.327–0.333)	0.856 (0.854–0.858)
TVE	0.732 (0.721–0.738)	—	0.928 (0.925–0.933)	0.000	0.895 (0.893–0.898)

RNB has the highest criterion value in all five runs, with margins of 0.012, 0.053, 0.048, 0.040, and 0.022 over the next method; its AUC margin ranges from 0.079 to 0.099. In the run reported in section 3 of the article, RNB has its lowest criterion value as well as its smallest criterion margin over RAM and its smallest AUC margin among the five runs. The reported run is therefore the least favourable of the five for the relative performance of RNB, so the main-text comparison understates rather than overstates its advantage.

Variation across seeds differs between methods. The two goodness-of-fit statistics and the two naive Bayes classifiers vary by at most two to three percentage points for

most quantities, because their predictions are based on closed-form expressions and the seed primarily changes the allocation of individuals to the data splits. The main exception is the admixed rate of NB, which ranges from 0.483 to 0.608 because the admixed class is learned from simulated offspring and therefore depends more strongly on the particular draw. The recent admixture model shows its largest variation for the out-of-reference rate (0.709 to 0.743). As discussed in section 4 of the article, its optimisation can converge to different optima depending on the initialisation.

S1.9 The acceptance rule and the parents, method by method

Fig. 6 of the article shows, for the recommended method, the outcome of the acceptance rule and, for predictions labelled Admixed, the inferred parental populations. The figures below show the corresponding results for the other six methods, using the acceptance score reported for each method in table 1 and following the order in section 2.3 of the article: first the three methods retained for detailed comparison alongside the recent classifier, followed by the three remaining methods. Figures S2 to S4 show both components; figs. S5 to S7 show only the left component because these three methods do not provide a parental-population pair. Rows, columns, the line separating in-reference from out-of-reference rows, and the dagger follow the conventions of fig. 6 of the article. The parental rows are shown in the same order, and every panel pools the same six folds and twenty iterations, allowing direct comparison across methods and with the main-text figure.

S1.10 The 28 newly analysed samples, one row each

Table S2 lists the 28 samples of section 3.4 of the article individually. The *stated ancestry* records the ancestry reported by the donor for each parent, including the reported country or region. The *target* maps this information to the classes represented in the reference panel and therefore provides the operational reference against which the prediction is compared. *Markers* gives the number of the panel’s 104 markers available for the respective sample. One marker, **rs12594144**, failed in almost every sample, and two additional markers of the assay are not part of the panel, so the attainable maximum is 103. The acceptance score s_{lik} is compared with the threshold of the predicted class. The margin is defined as their difference, so a sample is retained when the margin is non-negative.

S1.11 The remaining groups of the reference table, group by group

Table 2 in section 3.3 of the article summarizes all seven groups; here, we describe the groups individually. Among the 47 east african individuals, which form the group with the largest distance from the reference panel and have a rejection rate of 0.83, 37 are described as a mixture of AFR and MEA in the underlying pair assignment. The ROMA group contains only three individuals, so it has little influence on the aggregate results and its F_{ST} estimate should be interpreted with caution. The 136 (SANGER) MIDDLE EAST individuals represent the same broad region as the panel’s MEA population but originate from a different data source [1]. Their distance to the panel’s MEA population

these samples, the predicted class and parental pair indicate the most likely model output before rejection and are not reported as final classification results.

ID	stated ancestry and target			output of the pipeline		
	paternal	maternal	target	markers	predicted class	most likely parental pair
ID1	Australia	Scotland (GB)	Oceanian, European	93 kept	European	European, European, Euro- pean
ID2	Russia/Latvia	Germany	European	103 rejected	European	European, Euro- pean
ID3	Germany	Philippines	European, East Asian	96 rejected	Admixed	East Asian, Euro- pean
ID4	Russia/Kazakhstan	Germany	European	100 kept	European	European, Euro- pean
ID5	Iran	Germany	Middle East, Euro- pean	102 kept	Admixed	European, Middle East
ID6	Syria/Iraq	Germany	Middle East, Euro- pean	99 kept	Admixed	European, Middle East
ID7	Germany	Guyana/India	European, South Asian	100 kept	Admixed	European, South Asian
ID8	Italy/Colombia/CostaR	Indigenous Peru/Spain	European, American	100 rejected	Admixed	European, South Asian
ID9	Germany	Canada (franco- phone)	European	96 kept	European	European, Euro- pean
ID10	Turkey (Kurdistan region)	Turkey (Kurdistan tan region)	Middle East	87 kept	Admixed	European, Middle East
ID11	Germany	Brazil	European, American	103 rejected	Admixed	European, South Asian
ID12	Brazil (AfroBrazil- ian/Indigenous/Caribbe an region)	Germany	African, European	100 rejected	Admixed	European, Middle East
ID13	Turkey (Kurdistan region)	Turkey (Kurdistan tan region)	Middle East	102 rejected	Admixed	European, Middle East
ID14	Turkey (Kurdistan region)	Turkey (Kurdistan tan region)	Middle East	102 kept	Middle East	Middle East, Mid- dle East
ID15	Turkey (Kurdistan region)/Assyrian	Turkey (Kurdistan tan region)	Middle East	97 kept	Middle East	Middle East, Mid- dle East
ID16	Germany	Ecuador (mes- tizo) (GGP; Lebanon)	European, American	101 rejected	Admixed	American, Euro- pean
ID17	Germany	Cambodia	European, East Asian	103 kept	Admixed	East Asian, Euro- pean
ID18	Germany	Germany/Italy	European	82 kept	European	European, Euro- pean
ID19	Germany	Philippines	European, East Asian	100 kept	Admixed	East Asian, Euro- pean
ID20	Germany	Philippines	European, East Asian	78 kept	East Asian	East Asian, East Asian
ID21	Germany	Chile/Italy	European, American	103 rejected	European	European, Euro- pean
ID22	Turkey (Arameans)	Turkey	European, Middle East	99 rejected	Admixed	European, Middle East
ID23	USA (GP: indigenous)	Germany	American, European	103 kept	European	European, Euro- pean
ID24	Germany	Japan	European, East Asian	95 kept	Admixed	East Asian, Euro- pean
ID25	Costa Rica	Germany	American, European	96 rejected	European	European, Euro- pean
ID26	Cuba	Germany	American, European	97 rejected	Admixed	European, Middle East
ID27	Portugal/Brazil (Indigenous)	Germany/Africa	American, European, African	100 rejected	Admixed	Middle East, South Asian
ID28	Germany	Lebanon	European, Middle East	99 kept	Admixed	European, Middle East

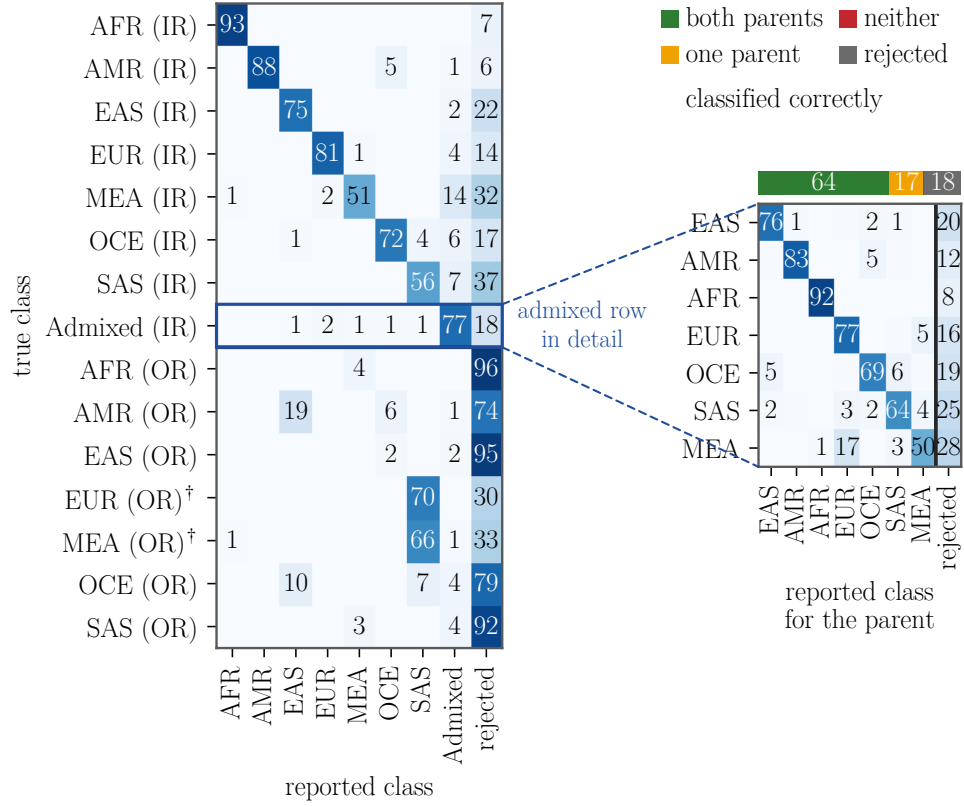


Fig. S2 The recent admixture model under s_{post} . Among the six comparison methods, its decision pattern is most similar to that of the recommended RNB classifier in fig. 6 of the article, consistent with RAM ranking second on the criterion. For out-of-reference individuals, RAM assigns a population label in 0.243 of cases, approximately three times the corresponding proportion for RNB, resulting in a lower rejection rate. For parental-population recovery, both parents are recovered in 0.645 of cases compared with 0.636 for RNB, while 0.182 are rejected compared with 0.206 for RNB. MEA is the weakest parental row for both methods.

is $F_{ST} = 0.007$, and 0.25 are rejected. This group therefore provides information on the behaviour of the MEA threshold for independently sourced individuals from a represented region rather than on rejection of an absent population. For a north african individual, MEA is the closest available reference population. Whether such an assignment should be counted as an error is therefore primarily a question of the reference panel's definition and resolution rather than of the classification method.

The two admixed cohorts are included for completeness, although a unique correct classification is not defined for them. The classifier assigns the Admixed class to 0.61 of these individuals before application of the acceptance rule, while the acceptance rule rejects 0.76. This pattern suggests that the cohorts contain ancestry components not represented by the reference panel.

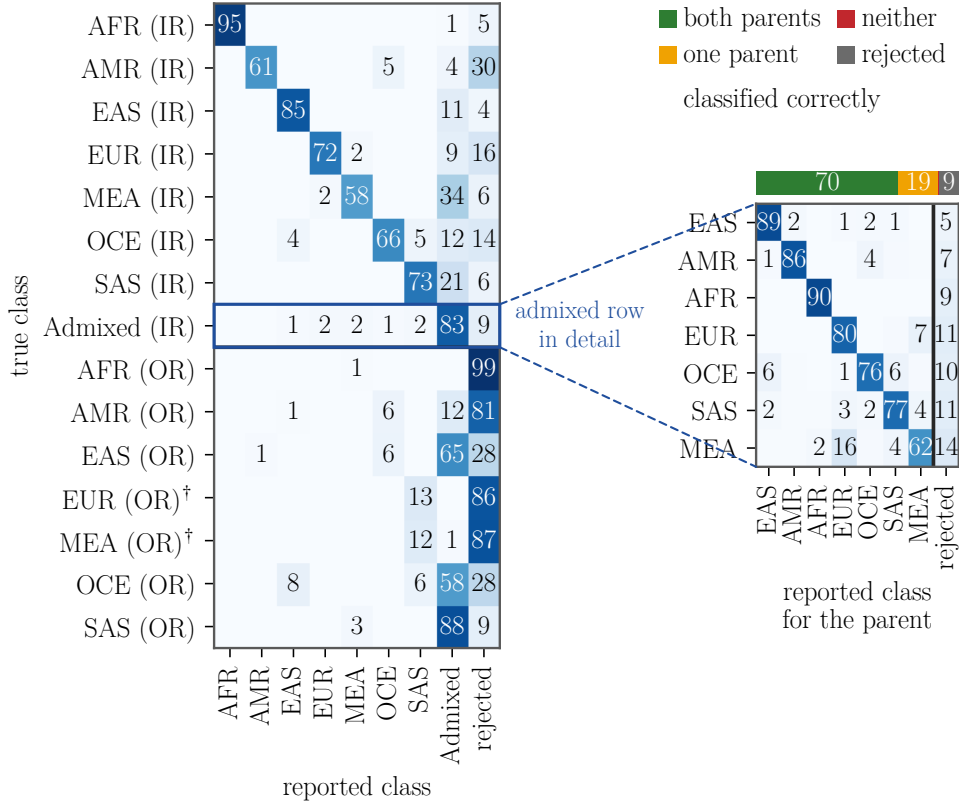


Fig. S3 The admixture model under s_{lik} . Among out-of-reference individuals, 0.370 are assigned to the Admixed class rather than rejected. This reflects the flexibility of the admixture model, which can represent a genotype as a combination of reference populations even when none provides a sufficient fit on its own. The model recovers both parental populations in 0.704 of cases, the highest proportion among the four methods that provide parental information, and rejects 0.095, the lowest proportion among them. Parental populations are defined as the two populations with the largest fitted proportions rather than as an explicitly modelled pair, which explains the similarity of the resulting matrix to those of the pair-based methods.

S2 Likelihoods and tests

S2.1 Likelihoods of the compared methods

Assume we have unphased genetic training/reference data $(x_{ikjm})_{i,k,j,m}$, where $x_{ikjm} \in \{0, 1, 2\}$ is the number of alleles j in individual i from population k at marker m . Because humans are diploid, $\sum_j x_{ikjm} = 2$ for all i, k, m , and $s_{kjm} := \sum_i x_{ikjm}$ is the number of appearances of allele j in population k at marker m , such that $n_k := \sum_j s_{kjm}$ does not depend on m and is the number of (haploid) training samples from population k .

In addition, let $(y_m)_m$ denote the test data, where $y_m = (y_{m1}, y_{m2})$ is the pair of alleles observed at marker m . We describe the methods used in this study through their

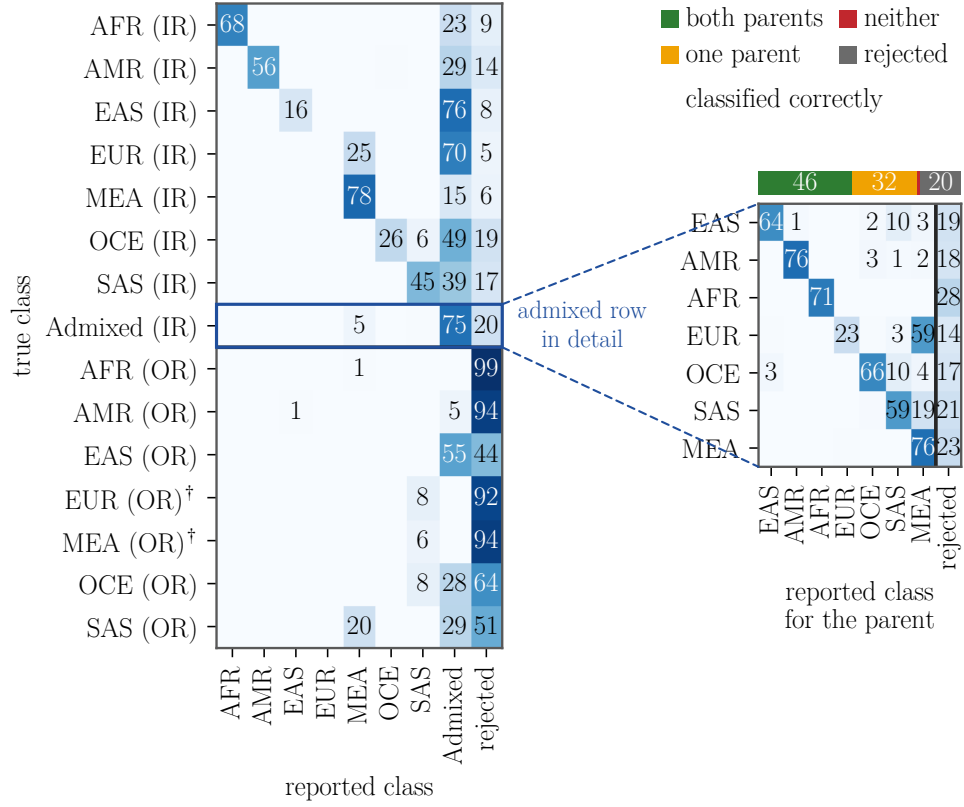


Fig. S4 The pairwise goodness-of-fit statistic under either score. The method frequently assigns in-reference individuals to the Admixed class: 0.513 of truly in-reference individuals receive this label. Its in-reference rate of 0.332 therefore mainly reflects the class assignment made before application of the acceptance rule rather than the placement of the threshold. It is also the weakest of the four methods that provide parental information: both parental populations are recovered in 0.461 of cases, with 0.200 rejected. The loss is concentrated in EUR, the same population that drives the over-assignment to Admixed (fig. 5 of the article).

log-likelihoods, up to an additive constant, which are then maximized. Naive Bayes and Recent Naive Bayes yield classifications into one population or a pair of populations, respectively. Admixture and Recent Admixture instead estimate admixture proportions, denoted by q , for the test sample.

Naive Bayes:

Assuming independence of the alleles in y ,

$$\ell(c | y) = \sum_m \left[\log((s_{c,y_{m1},m} + 1)(s_{c,y_{m2},m} + 1)) - 2 \log(n_c + 4) \right]$$

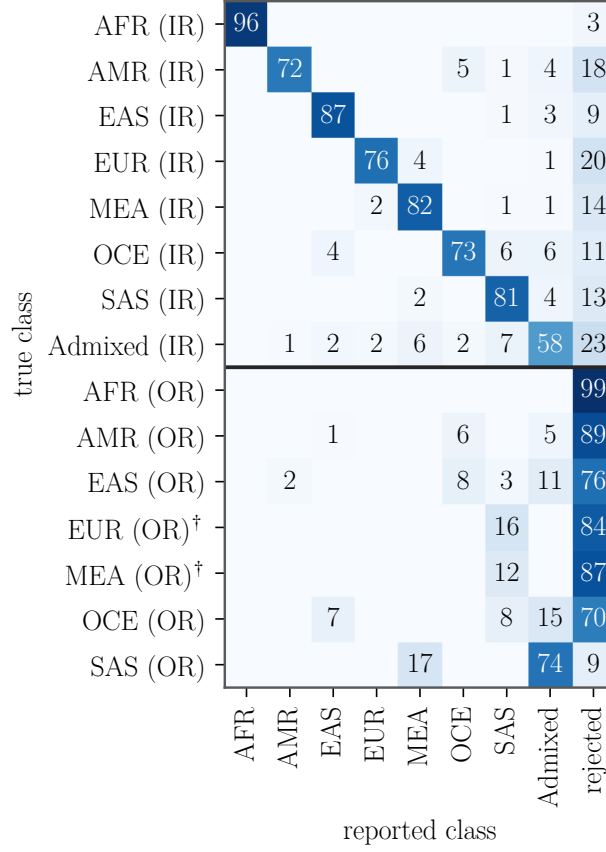


Fig. S5 Naive Bayes over populations under s_{lik} , after applying the class-specific acceptance thresholds. Errors occur in both directions: 0.192 of admixed individuals receive a population label, whereas 0.187 of out-of-reference individuals are labelled Admixed. This pattern results from the pooled Admixed class learnt from simulated offspring.

Recent Naive Bayes:

For $k \neq l$, either y_{m1} originates from population k and y_{m2} from population l , or vice versa, giving

$$\ell(k, l | y) = \sum_m \left[\log((s_{ky_{m1}m} + 1)(s_{ly_{m2}m} + 1) + 1_{y_{m1} \neq y_{m2}}(s_{ly_{m1}m} + 1)(s_{ky_{m2}m} + 1)) \right. \\ \left. - \log((n_k + 4)(n_l + 4)) \right]$$

Admixture:

Here, we set $p_{kjm} := \frac{1}{N_k} s_{kjm}$ as the frequency of allele j in population k at marker m . Let $q = (q_1, \dots, q_K)$, with $q_k \geq 0$ and $\sum_k q_k = 1$, denote the proportions of the test individual's genome originating from populations $1, \dots, K$, i.e. the admixture

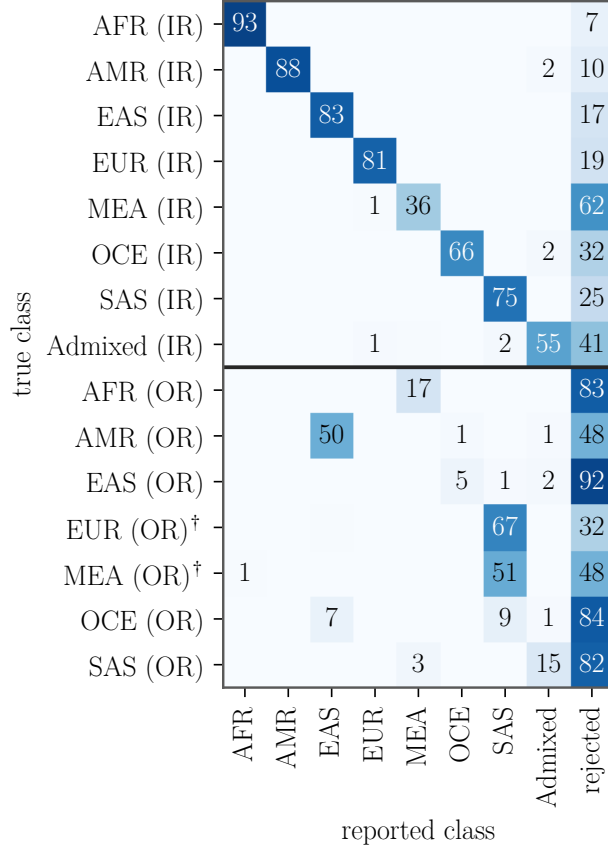


Fig. S6 The tabular transformer under s_{post} , after applying the class-specific acceptance thresholds. The largest loss occurs for admixed individuals, of whom 0.412 are rejected, the highest proportion among the methods shown here. Its posterior is normalised over the classes available during training, which can provide good relative ranking among those classes but cannot directly represent the possibility that none of them provides an adequate fit.

proportions. Then,

$$\ell(q|y) := \sum_m \log \langle q, p_{\cdot y_{m1}m} \rangle + \log \langle q, p_{\cdot y_{m2}m} \rangle,$$

where $\langle q, p \rangle := \sum_k q_k p_k$ is the scalar product and $p_{\cdot jm} := (p_{kjm})_{k=1,\dots,K}$.

Recent admixture:

Let q' and q'' denote the maternal and paternal admixture proportions of the individual, respectively. The corresponding log-likelihood is

$$\ell(q', q''|y) := \sum_m \log(\langle q', p_{\cdot y_{m1}m} \rangle \cdot \langle q'', p_{\cdot y_{m2}m} \rangle + 1_{y_{m1} \neq y_{m2}} \langle q'', p_{\cdot y_{m1}m} \rangle \cdot \langle q', p_{\cdot y_{m2}m} \rangle).$$

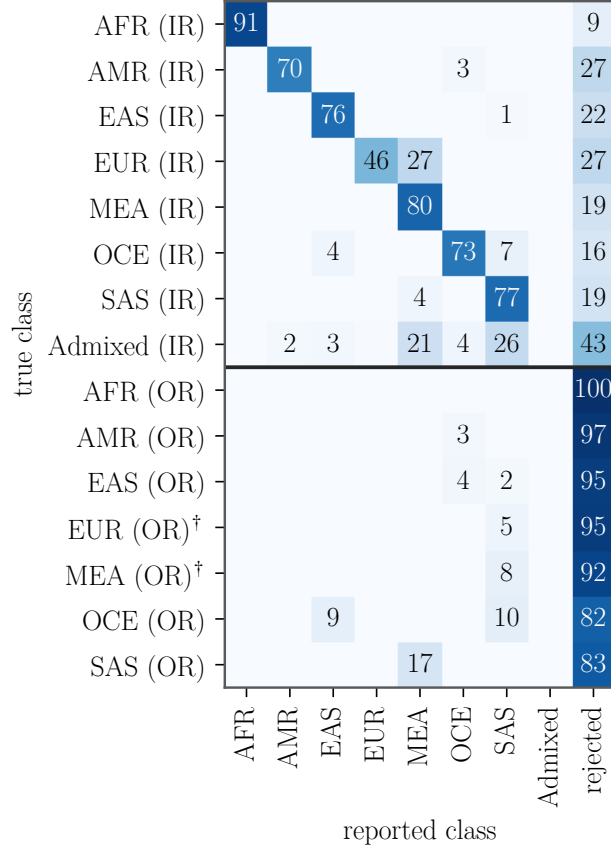


Fig. S7 The single-population goodness-of-fit statistic, after applying the class-specific acceptance thresholds. The Admixed column is empty because the statistic does not represent two parental populations. At the same time, this method has the highest rejection rate among the six comparison methods, which explains why its AUC is highest in every design while its criterion value is zero. The figure therefore illustrates why both quantities are reported side by side.

S2.2 Tests from Tvedebrink

Tvedebrink et al. [6] considered the following test problem for a given population k :

$$H_0 : y \text{ originates from population } k$$

against the alternative

$$H_1 : y \text{ does not originate from population } k.$$

For the biallelic markers considered by Tvedebrink et al. [6], only the counts s_{k0m} and s_{k1m} are required.

We define $s_{+km} := y_m + s_{k1m}$ as the total count of allele 1 at locus m in population k together with the test individual. A likelihood-ratio approach for biallelic markers leads to the following locus-specific statistic. Writing $\beta := \frac{s_{+km}}{2(n_k+1)}$ for the allele frequency in population and individual together and $\gamma := \frac{s_{+km} - y_m}{2n_k}$ for the one in the population alone,

$$Q(y_m, s_{+km}) = \frac{\beta^{s_{+km}} (1 - \beta)^{2(n_k+1) - s_{+km}}}{\left(\frac{y_m}{2}\right)^{y_m} \left(1 - \frac{y_m}{2}\right)^{2 - y_m} \gamma^{s_{+km} - y_m} (1 - \gamma)^{2n_k - s_{+km} + y_m}}.$$

Assuming independence across markers, the resulting test statistic is

$$T := \frac{\sum_{m=1}^M -\log(Q(y_m, s_{+km})) + \mathbb{E}_{y_m}(Q(y_m, s_{+km}))}{\sqrt{\sum_{m=1}^M \mathbb{V}_{y_m}(Q(y_m, s_{+km}))}}$$

which is assumed to follow a normal distribution. To calculate the expectation and variance in this expression, we use

$$\begin{aligned} p(y_m = 0 | S_{+km} = s_{+km}) &= \frac{(2(n_k + 1) - s_{+km}) (2n_k + 1 - s_{+km})}{2(n_k + 1) (2n_k + 1)}, \\ p(y_m = 1 | S_{+km} = s_{+km}) &= \frac{2(2(n_k + 1) - s_{+km}) s_{+km}}{2(n_k + 1) (2n_k + 1)}, \\ p(y_m = 2 | S_{+km} = s_{+km}) &= \frac{s_{+km} (s_{+km} - 1)}{2(n_k + 1) (2n_k + 1)}. \end{aligned}$$

Multiallelic markers

Tvedebrink et al. [6] restrict the analysis to biallelic markers, whereas the panel used here also contains multiallelic markers. We therefore treat each allele a of a marker as a separate locus, “ a against the remaining $2 - a$ copies”, and sum the numerator and variance in T over these allele-specific loci in the same way as the original statistic sums over markers. One allele per marker is redundant and is dropped as the reference, so a bi-allelic marker leaves a single allele and T reduces to the original statistic. Only fully typed loci enter, and the reference panel is built on the training fold, so an allele unseen there is treated as missing rather than paired against a count of zero.

TVP: the pairwise extension

Tvedebrink and Eriksen [5] extend the test to a pair of populations. Under

$$H_0 : y \text{ is a first-order admixture of populations } k \text{ and } l,$$

one allele of the individual comes from each parent, so with $\hat{p}_{kam} = (s_{kam} + 1)/(2n_{km} + 2)$ the Laplace-smoothed frequency of allele a at marker m in population k , and

$y_{am} \in \{0, 1, 2\}$ its count in the profile,

$$\begin{aligned}\mathbb{P}(y_{am} = 0 \mid k, l) &= (1 - \hat{p}_{kam})(1 - \hat{p}_{lam}), \\ \mathbb{P}(y_{am} = 1 \mid k, l) &= \hat{p}_{kam}(1 - \hat{p}_{lam}) + (1 - \hat{p}_{kam})\hat{p}_{lam}, \\ \mathbb{P}(y_{am} = 2 \mid k, l) &= \hat{p}_{kam}\hat{p}_{lam},\end{aligned}$$

which is the same convolution used by the RNB classifier. Against the alternative of a random draw at allele frequency $\frac{1}{2}$, that is $(\frac{1}{4}, \frac{1}{2}, \frac{1}{4})$ as in [5], the log-likelihood ratio per locus is $\log Q_{am} = \log \mathbb{P}(y_{am} \mid k, l) - \log \mathbb{P}(y_{am} \mid H_1)$, and the pair is scored by

$$z(k, l \mid y) := \frac{\sum_{a,m} [-\log Q_{am} - \mathbb{E}_{H_0}(-\log Q_{am})]}{\sqrt{\sum_{a,m} \mathbb{V}_{H_0}(-\log Q_{am})}},$$

the sums running over the typed loci of the individual. Mean and variance are taken exactly over $y_{am} \in \{0, 1, 2\}$ under H_0 rather than by the delta approximation of [5]; z is then approximately standard normal when (k, l) is the true pair, and larger values of z provide stronger evidence against the pair. The pair (k, k) is the single-population test above, so the $\binom{K+1}{2} = 28$ candidate pairs contain the $K = 7$ single populations as their diagonal. The individual is assigned the pair attaining the smallest z . This is reported as a single population when $k = l$ and as Admixed otherwise. The quantity $\min_{k,l} z(k, l \mid y)$ is used for the acceptance rule, with its sign inverted because smaller values indicate a better fit.

Compared with Tvedebrink and Eriksen [5], we make one simplification: the profile is not added back to the reference counts, i.e. we omit the EM step for allocating heterozygous alleles to the two parents. This affects the calibration of z but not the ranking of the population pairs.

References

- [1] Almarri MA, Haber M, Lootah RA, et al (2021) The genomic history of the middle east. *Cell* 184(18):4612–4625
- [2] Balding DJ, Nichols RA (1995) A method for quantifying differentiation between populations at multi-allelic loci and its implications for investigating identity and paternity. *Genetica* 96:3–12. <https://doi.org/10.1007/bf01441146>
- [3] Bhatia G, Patterson N, Sankararaman S, et al (2013) Estimating and interpreting F_{ST} : the impact of rare variants. *Genome Research* 23(9):1514–1521. <https://doi.org/10.1101/gr.154831.113>
- [4] Hudson RR, Slatkin M, Maddison WP (1992) Estimation of levels of gene flow from DNA sequence data. *Genetics* 132(2):583–589. <https://doi.org/10.1093/genetics/132.2.583>

- [5] 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>
- [6] 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>