Biogeographical Ancestry

Workshop

Torben Tvedebrink, PhD
Department of Mathematical Sciences – Aalborg University, Denmark

Section of Forensic Genetics – Department of Forensic Medicine
Faculty of Health and Medical Sciences – University of Copenhagen, Denmark

ISFG2022 – International Society for Forensic Genetics
Washington, DC – 30th August 2022

Version: 30/08/2022 08:05

Welcome

Today's topic

Welcome to the half-day workshop on **Biogeographical Ancestry** at ISFG2022.

Disclaimer: Workshop is not about selecting AIMs

Website abstract (*excerpt*):

In recent years, biogeographical ancestry markers have gained attention in the forensic genetics community [...]

There exists several methods for modelling AISNPs [...] However, a common assumption of these approaches is that a true population exists in the reference material [...]

The workshop will discuss the strengths and weaknesses of the listed methodologies, introduce the participants to the ideas behind genogeographer.org, and show how to use this tool in relation to casework.

Agenda, times and breaks

Theory and hands-on

The official ISFG2022 time slots divides the workshop into two main time slots divided by a common coffee break for all workshops.

Time	Session
09:00 - 10:30	Lecture
10:30 - 11:00	Coffee break
11:00 - 13:00	Hands-on and discussions

We will have some additional small breaks during the lecture and hands-on session.

Lecture topics

Please interrupt with questions (raise hand) during the presentation

- What are AIMs
- What is the objective of biogeographical ancestry analysis?
- Solution/Approaches to analysis
  - Likelihood based methods (e.g. likelihood ratios)
  - Principal Components Analysis (PCA, e.g. EIGENSTRAT)
  - Clustering (e.g. STRUCTURE, ADMIXTURE, etc.)
  - Classification (e.g. machine learning approaches)
  - Outlier detection (e.g. GenoGeographer)

Ancestry Informative Markers

Disclaimer: Workshop is not about selecting AIMs

An **ancestry informative marker (AIM)** is a marker that can inform us about the **ancestral origin** of an individual.

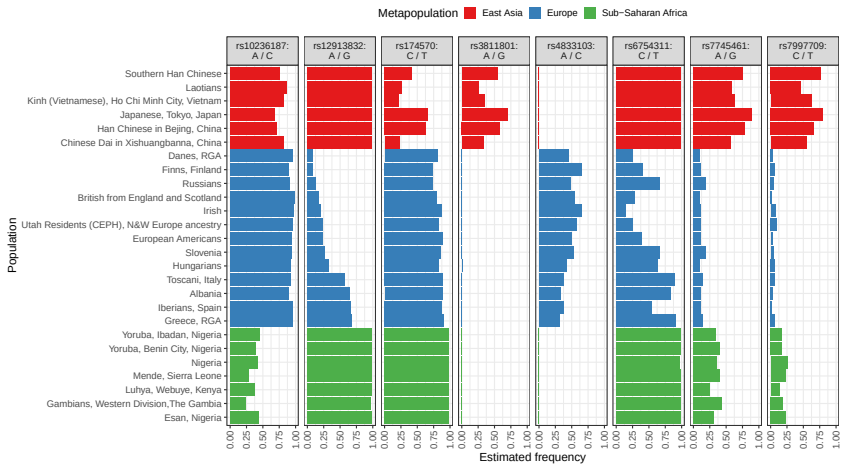
Thursday morning's keynote speaker Professor Noah **Rosenberg** previously published two excellent overviews of various statistical methods for **selecting AIMs** (Rosenberg et al., 2003; Rosenberg, 2005).

Hence, this workshop is **not about** identifying new markers, but to make **proper inference** of the results of a pre-selected **set of markers**. **Chris Phillips**' scientific price lecture tomorrow morning will have more on that subject.

The panel of AIMs used throughout this workshop, has been selected in order to discriminate between individuals over **large continental distances**. However, the methodologies are not limited to this type of panels.

AIMs set

Thermo Fisher Precision ID Ancestry Panel (excerpt)



AIMs set

Thermo Fisher Precision ID Ancestry Panel (excerpt)

Specifically, the AISNP set considered here is the Applied Biosystems™ **Precision ID Ancestry Panel**, which includes $\boldsymbol{L} =$ **165 autosomal markers (AISNPs)**.

The 165 markers are a union (13 overlapping markers) of AISNPs selected by

- The Seldin-lab (Kosoy et al., 2009): **123 AISNPs**
- The Kenneth Kidd-lab (Kidd et al., 2014): **55 AISNPs**

Each marker is bi-allelic, e.g. `A/C`, and we denote `A` **allele 1**, and `C` allele 2, i.e. we use lexicographic ordering. Hence, an **individual** has **0**, **1 or 2** copies of **allele 1**.

Selected markers and panels

Some remarks

When a panel is **designed** it is done with **specific application** in mind (e.g. segment populations on a global scale or identify regional and local differences). Thus, if used out of context, the panel may perform different than anticipated.

For example, the Precision ID Ancestry Panel has a **low discriminatory power among African populations**.

Markers that are included in panels to discriminate on local distances will typically be constant across most other populations (i.e. being non-informative) and basically just **add some random noise**.

For example, within GenoGeographer's Sub-Saharan African reference sample, Precision ID Ancestry Panel has **11 markers without any variability** and **18 markers where five or less** of the alternative alleles have been observed (out of 668 individuals or 1336 alleles).

Objective of forensic AIMs analysis

Investigative leads

In **population genetics** AIMs are interesting in their own right in the study of **population structures**, anthropology and **human evolution**.

However, from a forensic point of view we are not necessarily interested in same research questions as the population geneticists are.

Within a group of people with shared ancestry, AIMs are non-informative and can not be used for person identification purposes.

However, AIMs can be used to provide **investigative leads** in cases where little else is known (e.g. no hits in DNA databases or no/uncertain witness statements)

Objective of forensic AIMs analysis

Questions

Hence, it may be interesting to ask

- which population/ethnic group/subpopulation/geographical area/region is the observed AIM profile most frequent?
- is the relative frequency of the profile in region A compared to region B?
- is likely that the profile originates from the country/region where is was found?
- the unidentified body more likely to originate from a foreign country?
  And, if so – which country/region/continent?

Objective of forensic AIMs analysis

Questions

All forensic scientists know that such questions preferably should be evaluated in terms of **likelihood ratios**, where the evidence is evaluated under two competing and mutually exclusive hypothesis.

$$LR = \frac{P(\text{AIM profile} \mid \text{Profile originates from region A})}{P(\text{AIM profile} \mid \text{Profile originates from region B})}$$

Well-defined hypothesis

Likelihood ratios

The use of **likelihood ratios** is **advised** by several commissions under the International Society of Forensic Genetics.

The hypotheses considered are typically exhaustive implying that their union constitutes **all relevant hypotheses**.

In such circumstances, the use of likelihood ratios is unproblematic (and often straight forward). When it comes to the ancestry of an individual this may, however, not be the case.

In the case of **ancestry**, the hypotheses will typically be *generated* by the **populations**, from which we have samples.

Potential issue with sample-driven hypothesis

Limited population samples

Because the estimation of the profile frequency relies a population sample we are limited to evaluate likelihood ratios for populations, for which we have sufficient number of sampled individuals.

There exists many publicly available repositories (e.g. gnomAD) containing whole genome sequences for a broad selection of human population across the globe. However, several populations (defined by language, culture, ethnicity, geography, etc.) remains untyped (or at insufficient numbers), which limits the number of likelihood ratios we can formulate.

When limited to the sampled populations, the human evolutionary history will also influence our interpretation of the results.

Human evolution

Related populations

Human evolution

Related populations

Because of the historical evolution, **human populations are related** through the migration patterns and other events.

Consequently, South American populations are closer to Asian population than those of Europe and Africa. Hence, an AIMs profile sampled from South America will have a higher likelihood in Asia than in Africa.

If compared though a LR, one would conclude that the profile would be orders of magnitude more likely to observe in Asia than in Africa. This could potentially result in criminal investigators would search for perpetrators with Asian origins.

However, if we had worldwide coverage of populations this issue would reduced.

But we don't

The situation

Figuratively in terms of fruit

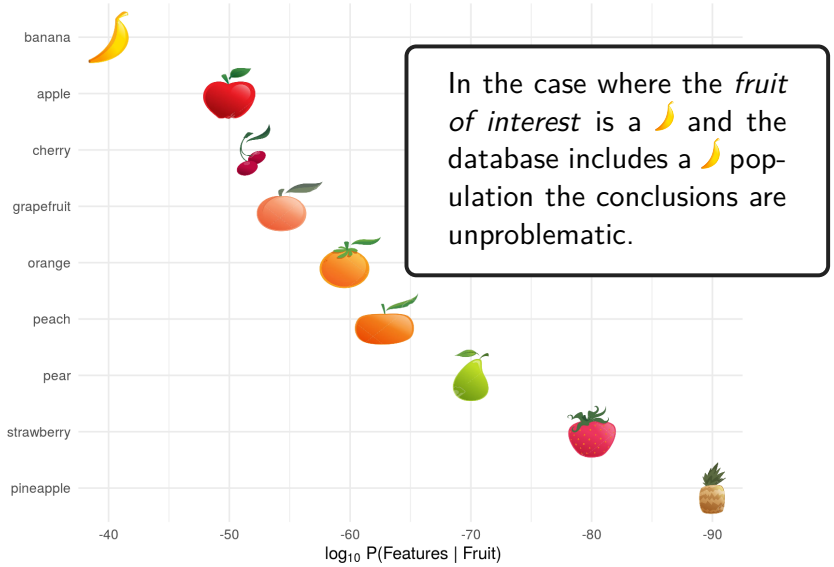
Approaches and solutions

Some popular choices

- Likelihood based methods (e.g. likelihood ratios as implemented in FROG-kb and Snipper)
- Principal Components Analysis (PCA, e.g.EIGENSTRAT)
- Clustering (e.g. STRUCTURE, ADMIXTURE, etc.)
- Classification (e.g. machine learning approaches)
- Outlier detection (e.g. GenoGeographer)

The non-exhaustive problem

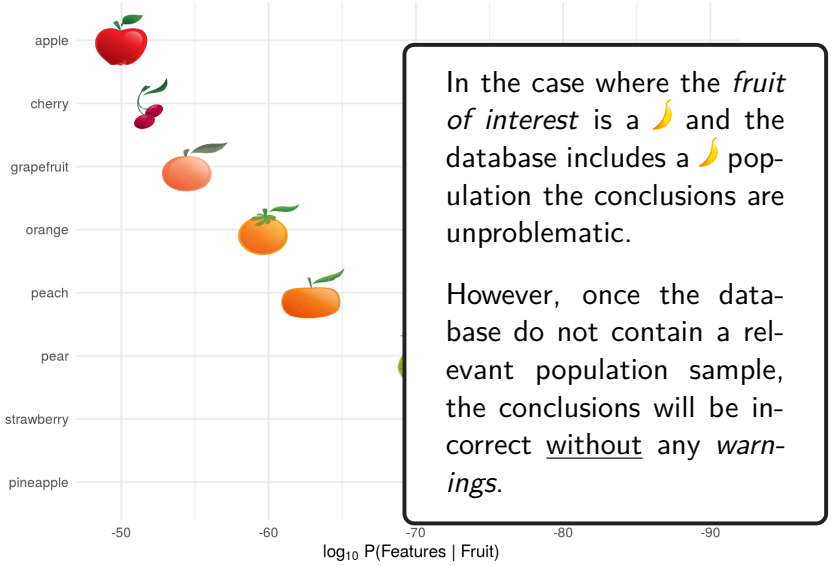
Likelihood (ratio) approach



In the case where the *fruit of interest* is a 🍌 and the database includes a 🍌 population the conclusions are unproblematic.

The non-exhaustive problem

Likeli



In the case where the *fruit of interest* is a 🍌 and the database includes a 🍌 population the conclusions are unproblematic.

However, once the database do not contain a relevant population sample, the conclusions will be incorrect without any *warnings*.

Principal Components Analysis

PCA

Principal components analysis (PCA) has a long history of application in the study of **population structure**. The usage of PCA in analysis of population structure was pioneered by Menozzi et al. (1978), as an efficient way to capture the underlying structure for visualisation purposes.

PCA is able to capture **continuous admixture** between populations, implying that admixed populations typically falls on the *lines* between its *parental* populations.

McVean (2009) derived a **close link** between the populations' genealogical **coalescent times** and the **primary axes** of a PCA.

Principal Components Analysis

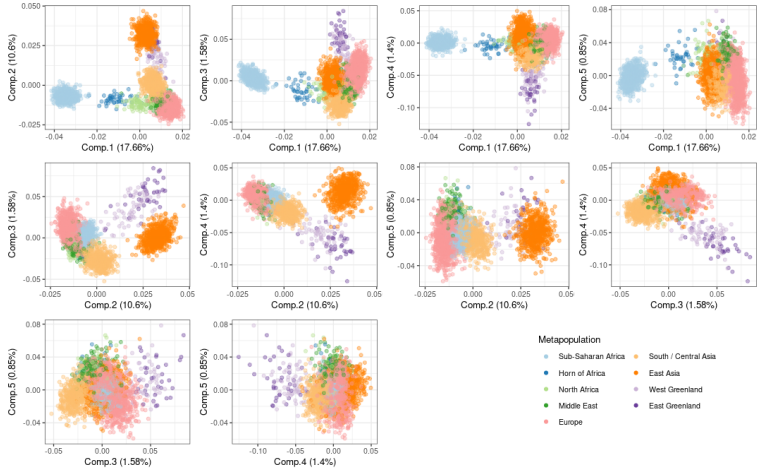
PCA

More recent PCA based methods for inference include SMARTPCA (Patterson et al., 2006) and **EIGENSTRAT** (Price et al., 2006). These methodologies provide further insight as to how **many PCs** are required to **capture detectable population structures** by the use of hypothesis testing on the magnitude of the PCA's eigenvalues.

However, several papers discuss some of the known pitfalls when using PCA. The most important issue to be aware of, is PCA's sensitivity to the sampling of individuals and populations. **Unbalanced sampling** of some populations forces the PCs to account for the variation caused by the **majority groups** (Novembre et al., 2008; McVean, 2009; Wangkumhang et al., 2018; Miller et al., 2020).

GenoGeographer reference database

PCA– Five significant PCs



Forensic application

PCA

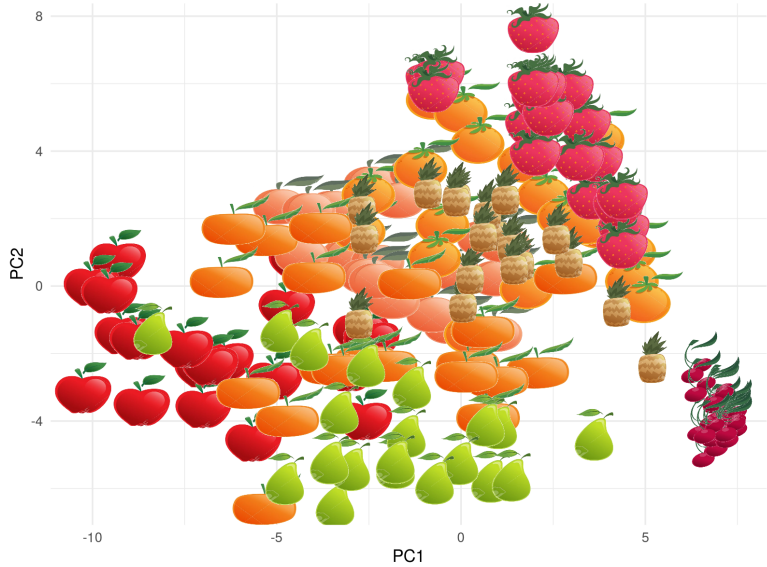
The PCA gives the projections of the AIM profiles to explain as much of the variability as possible with the first few axes. Alternatively, the best linear approximation of the profiles in a lower dimensional space.

However, from a **forensic point of view** it is rather difficult to assign a **quantitative value** to the projection of the profile of interest. The powerful feature of PCA is compression of the genetic data to provide a visual summary of the profile and reference database.

Furthermore, for a PC projection to work as intended a full profile is needed. In casework this is not always possible to genotype all markers resulting in **partial profiles**.

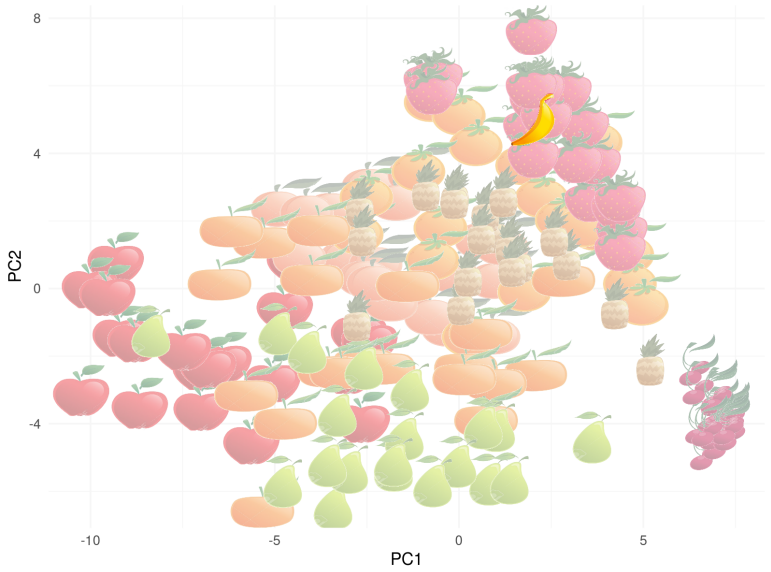
Fruity PCA

An example – The reference database



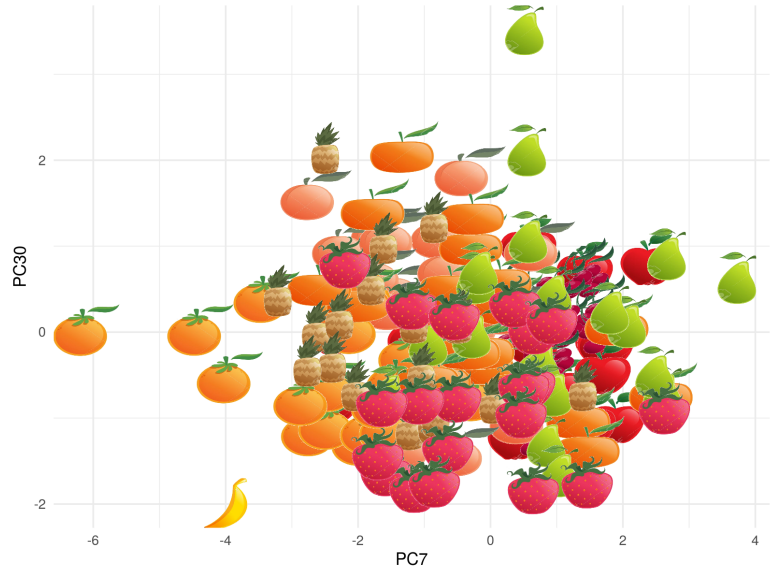
Fruity PCA

An example – The fruit of interest 🍌



Fruity PCA

An example – On some lower projection 🍓s and 🍌 don't match



STRUCTURE

Model-based clustering

The seminal **STRUCTURE** paper by Pritchard et al. (2000), introduced a population genetics methodology based on a statistical finite mixture model. Using a Bayesian approach, the posterior probabilities for **cluster memberships** and cluster-specific allele frequencies are estimated using a MCMC-algorithm (MCMC: Markov Chain Monte Carlo).

STRUCTURE has been used in numerous studies and has according to Google Scholar more than 30,000 citations, which indicates its **enormous influence** in the field of **population genetics** (Novembre, 2016)

STRUCTURE

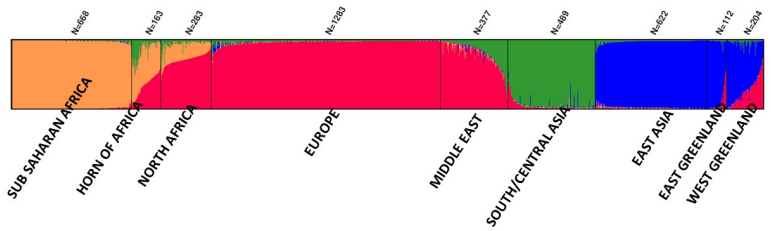
Model-based clustering

In essence, the initial STRUCTURE model is a simple **Hardy-Weinberg model** with the relaxation that subpopulations have different allele frequencies and individuals may inherit alleles from several subpopulations, i.e. being **admixed** (Novembre, 2016)

Following the initial publication, **several modifications** both on the population genetic model (allowing for e.g. linked markers and null alleles, Falush et al., 2003, 2007) and **computational aspects** (faster and more efficient algorithmic schemes, Tang et al., 2005; Alexander et al., 2009; Raj et al., 2014) has been suggested. See Novembre (2014, 2016) for further references and remarks.

GenoGeographer reference database

STRUCTURE– Nine metapopulations ($K = 4$)



STRUCTURE critique

Model assessment and choice of K

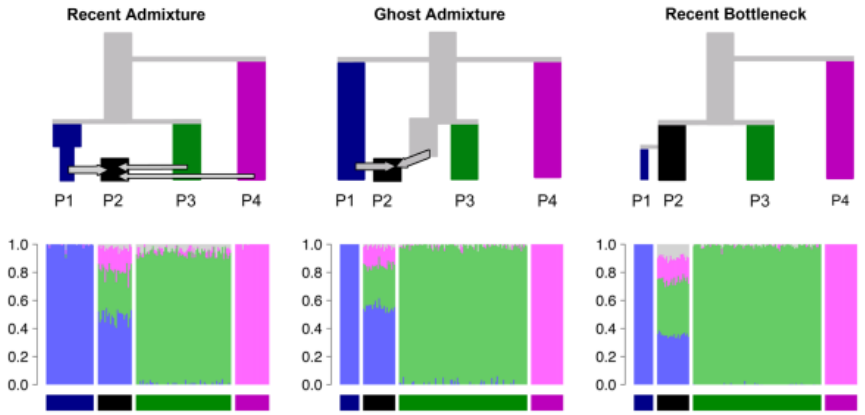
Because of STRUCTURE's complexity, it is hard to **conduct assessment of the model fit**. A recently published instructive tutorial (Lawson et al., 2018), provides a critical view on how to assess the outcome of STRUCTURE analysis using the tools badMIXTURE GLOBETROTTER, fineSTRUCTURE and CHROMOPAINTER (Lawson, 2018; Hellenthal et al., 2014; Lawson et al., 2012).

Furthermore, choosing the appropriate value for K, the number of population clusters, is **not a well defined problem**, i.e. only heuristic methods exists for guiding the specification of K (Pritchard et al., 2000; Novembre, 2014, 2016; Lawson et al., 2018).

Different choices of K may result in rather **different results and interpretation of the population stratification and history**.

STRUCTURE critique

badMIXTURE (Lawson et al., 2018) – Indistinguishable admixture plots



Forensic application

STRUCTURE

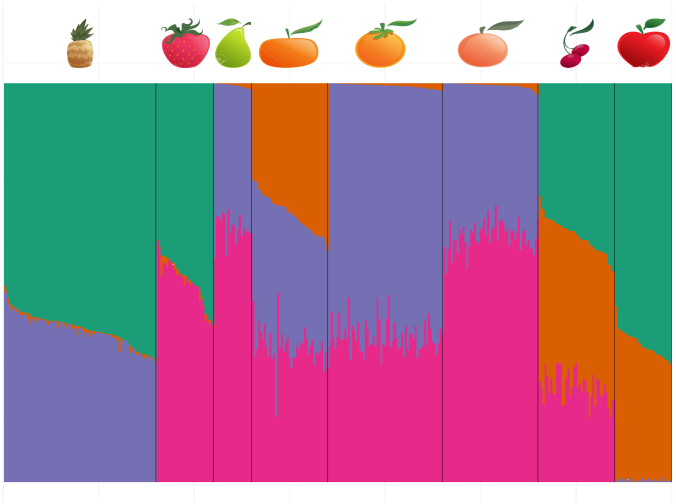
There exist several guides and tutorials for how to prepare the population data and choosing parameter settings for STRUCTURE (e.g. Porras-Hurtado et al., 2013; Santos et al., 2016, with some emphasis on forensic applications).

However, identifying which cluster the profile of interest resembles the most is difficult, which was also concluded by Porras-Hurtado et al. (2013, pp. 7):

> "One disadvantage of STRUCTURE highlighted by this study is the difficulty of analyzing single genotype profiles"

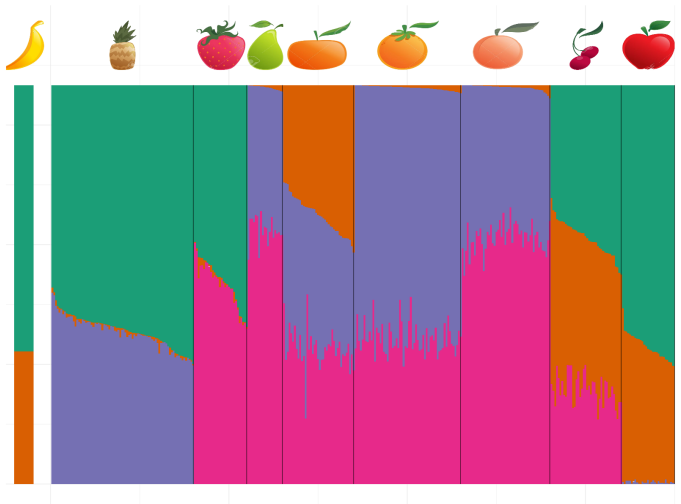
Fruity STRUCTURE

An example – The reference database ($K = 4$)



Fruity STRUCTURE

An example – The fruit of interest 🍌. How to assign?



Classification

Likelihood based

The likelihood (ratio) approach computes the profile frequency based on the assumed population genetic model.

Assuming Hardy-Weinberg equilibrium at least marker, which are assumed to be mutually independent, this is simply the product of the estimated allele frequencies in each population:

$$\begin{aligned} P(\boldsymbol{G} = \boldsymbol{g}) &= P(G_1 = g_1, G_2 = g_2, \ldots, G_L = g_L) \\ &= P(G_1 = g_1)P(G_2 = g_2)\cdots P(G_L = g_L) \\ &= \prod_{l=1}^{L} P(G_l = g_l) \\ &= \prod_{l=1}^{L} \binom{2}{x_{0l}} \hat{p}_l^{x_{0l}}(1-\hat{p}_l)^{2-x_{0l}}, \end{aligned}$$

where x_{0l} counts the number of allele 1 and $\hat{p}_l$ is the estimated frequency of allele 1 at locus l.

Classification

Machine learning

However, the classification does not need to be driven by the population genetic likelihood function.

Several other algorithms have and can been used in the context of AIMs prediction. To name a few:

- Multinomial regression (e.g. McNevin et al., 2013; Cheung et al., 2017, 2018),
- Discriminant Analysis (e.g. Jombart et al., 2010),
- Classification Trees or Random Forest, and
- Neural Networks (e.g. Qu et al., 2019)

Classification

Common issues

The above mentioned algorithms will typically all solve the problem of assigning a new observation $\boldsymbol{x}_0$ to the most probable population used in the training dataset.

However, as previously discussed, one such population may not exist since the profile of interest, $\boldsymbol{x}_0$, may originate from a population not represented in the reference database. Thus, $\boldsymbol{x}_0$ will be assigned to the **least improbable population**.

In addition, partial profiles are often also an issue for machine learning methods (as it where for PCA). In some of the input features are missing, the machine learning algorithm will typically not work out of the box.

GenoGeographer

Background and Motivation

The initial interest in developing a new tool for Biogeographical Ancestry prediction, was to answer the question:

> When is a profile frequency estimate too rare (too small) to be trusted?

Following that, would there be a way to detect when the *true* population wasn't available in the reference database?

> Another concern was, how many samples were needed per reference population to give reliable allele frequency estimates?

Likelihoods

Uncertainty

The second question could be answered using existing methodologies.

The uncertainty in allele frequencies is inversely proportional to the number of sampled alleles. This implies that fewer sample sizes increases the uncertainty in the allele frequency estimates.

Chakraborty et al. (1993) derived expressions for DNA profile frequency estimates, which we (Tvedebrink et al., 2018) adopted to assess the variability of the frequency estimates.

Simulation results (data not shown) suggested that at least $n = 75$ individuals were needed to obtain nominal coverage of the computed 95%-confidence intervals for a range of varying allele frequencies.

Hypothesis

The population of origin

To address the first question, we proposed to use a statistical likelihood ratio test framework, in order to overcome the focus on relative frequencies of AIMs profiles.

This corresponds to an **absolute measure of concordance** between an AIMs profile, $\boldsymbol{x}_0$, and those of a population, j.

It is informative to state the hypothesis that we are inquiring:

Hypothesis:

H_0: The AIMs profile **originates** from population j
H_1: The AIMs profile **does not originate** from population j.

Outlier detection

z-score approach

These hypotheses can be thought of as a way of detecting whether $\boldsymbol{x}_0$ is an **outlier** or not **relative to sample $\boldsymbol{x}_j$ from population $\boldsymbol{j}$**.

By arguments similar to those of **Fisher's exact test** for $r \times c$ tables, we can compute the exact distribution, from which we evaluate the expectation and variance of the likelihood ratio test (LRT) statistic.

This approach also removed the need for a lower limit on the number of reference samples in the LRT setting. Furthermore, partial profiles was not an issue since all calculations are done marker-wise.

Some mathematical details

Joint distribution

With the **independence assumption of the markers**, we just focus on a single marker.

Under the null hypothesis (cf. above), x_0 and x_j are from the same population j, where allele 1 has probability p_j.

Then we know that

$$\begin{aligned} x_0 &\sim \text{bin}(2, p_j), \\ x_j &\sim \text{bin}(2n_j, p_j) \\ x_+ &\sim \text{bin}(2(n_j{+}1), p_j), \end{aligned}$$

where $x_+ = x_0 + x_j$ is the number of allele 1 observed between *both* the reference population and $\boldsymbol{x}_0$ among their $2n_j + 2 = 2(n_j + 1)$ alleles.

Some mathematical details

Conditional distributions

Conditional on x_+ (the sum of x_j and x_0), it can be shown that x_0 follows a hyper-geometric distribution (Tvedebrink et al., 2018):

$$P(X_0 = x_0 \mid X_+ = x_+) = \frac{\binom{2}{x_0}\binom{2n_j}{x_+ - x_0}}{\binom{2(n_j{+}1)}{x_+}},$$

where x_+ is the sufficient statistic under the null hypothesis.

The conditioning trick implies that the uncertainty about the allele frequency is modelled directly in the distribution. Hence, the smaller the reference sample is the weaker conclusions can be made.

The test statistic

Likelihood ratio test

We form a **likelihood** ratio **of the data** under the hypotheses.

In the numerator, we assume a **common population**. Hence, $x_+ = x_0 + x_j$ is the sufficient statistic under the null hypothesis.

In the denominator, we assume **two different populations** (x_0 is not from population j as x_j is). Hence, we estimate the allele frequencies separately:

$$Q(x_0, x_+) = \frac{\left(\frac{x_+}{2(n_j{+}1)}\right)^{x_+}\left(1 - \frac{x_+}{2(n_j{+}1)}\right)^{2(n_j+1)-x_+}}{\left(\frac{x_0}{2}\right)^{x_0}\left(1 - \frac{x_0}{2}\right)^{2-x_0}\left(\frac{x_+ - x_0}{2n_j}\right)^{x_+-x_0}\left(1 - \frac{x_+ - x_0}{2n}\right)^{2n_j - x_+ + x_0}},$$

where $2n_j$ is the number of sampled alleles from population j.

The test statistic

Why this expression?

$$Q(x_0, x_+) = \frac{\overbrace{\left(\frac{x_+}{2(n_j+1)}\right)^{x_+}\left(1-\frac{x_+}{2(n_j+1)}\right)^{2(n_j+1)-x_+}}^{\text{We assume the same population}}}{\underbrace{\left(\frac{x_0}{2}\right)^{x_0}\left(1-\frac{x_0}{2}\right)^{2-x_0}}_{\text{Assumed } x_0 \text{ not from population } j}\ \underbrace{\left(\frac{x_+-x_0}{2n_j}\right)^{x_+-x_0}\left(1-\frac{x_+-x_0}{2n}\right)^{2n_j-x_++x_0}}_{x_j \text{ from population } j}},$$

which can be expressed as the ratio of two likelihoods

$$\approx \frac{\tilde{p}^{x_+}(1-\tilde{p})^{2(n_j+1)-x_+}}{{p_0}^{x_0}(1-p_0)^{2-x_0}\hat{p}_j^{x_j}(1-\hat{p}_j)^{2n_j-x_j}},$$

where p_0 is the frequency of allele 1 in x_0, $\hat{p}_j$ in population j and $\tilde{p}$ when they are joined together.

Marker-wise z-score

Standardising $-\log Q(x_0 \mid x_+)$

Since x_+ is the sufficient statistic under H_0, conditioning on x_+ brings us to **Fisher's exact test**.

The numerator in $Q(x_0 \mid x_+)$ is a **constant when conditioning**.

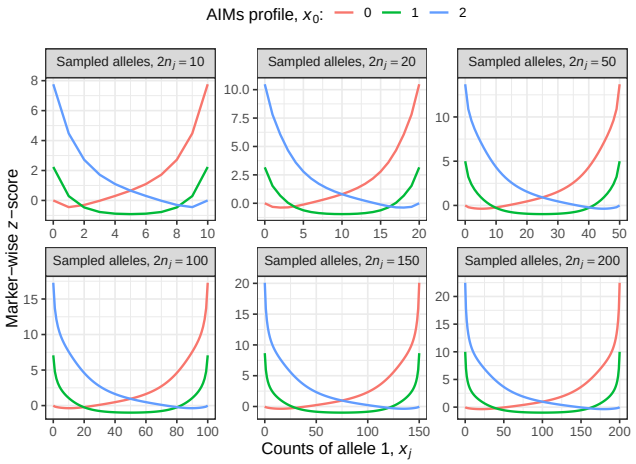
The distribution of $x_0 \mid x_+$ is **hyper-geometric**. Hence, the expectation and variance are easily computed over x_0 as this only takes the values of $\{0, 1, 2\}$.

We can standardise $-\log Q(x_0 \mid x_+)$ by **subtracting the expectation** and **dividing by the standard deviation**:

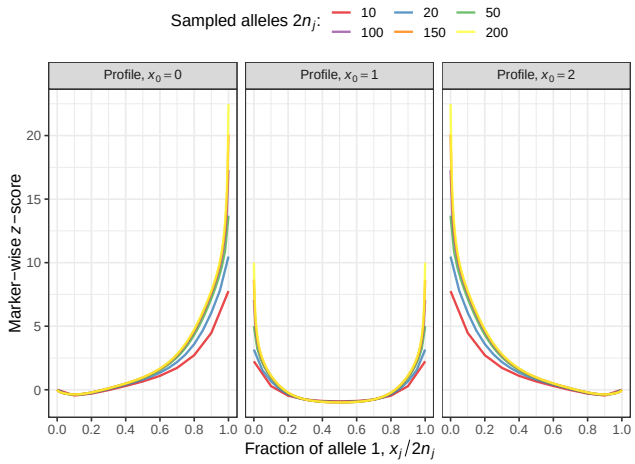
$$z = \frac{-\log Q(x_0 \mid x_+) + \mathbb{E}[\log Q(x_0 \mid x_+)]}{\sqrt{\mathbb{V}[\log Q(x_0 \mid x_+)]}}.$$

Marker-wise z-score

Visual representation

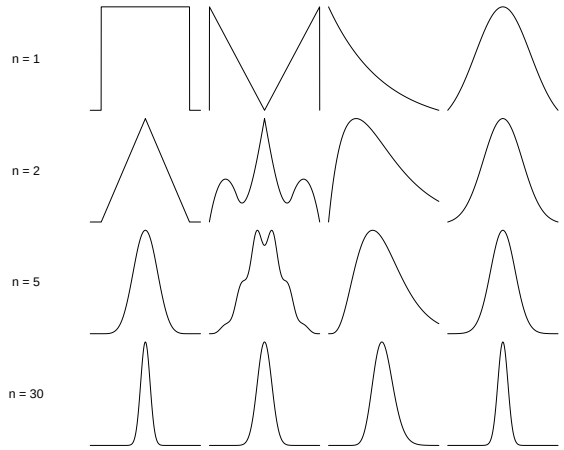


Marker-wise z-score



Central Limit Theorem (CLT)

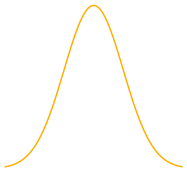
Visual proof



When the sample size, n, increases the distribution of the average

$$\bar{x} = \frac{1}{n}(x_1 + x_2 + \cdots + x_n)$$

will approach a normal distribution:



Summing over markers

Normal approximation

By **assuming independence** among markers, we sum over the L markers in order to **aggregate the evidence**:

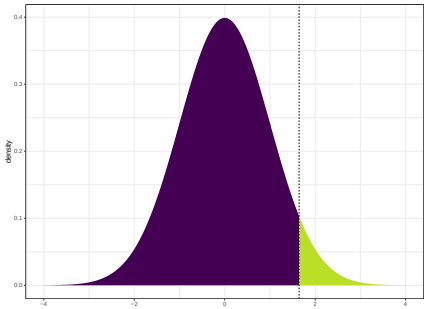
$$z = \frac{\sum_{l=1}^{L}\{-\log Q(x_{0l} \mid x_{+l}) + \mathbb{E}[\log Q(x_{0l} \mid x_{+l})]\}}{\sqrt{\sum_{l=1}^{L} \mathbb{V}[\log Q(x_{0l} \mid x_{+l})]}}.$$

Using a variant of the central limit theorem (CLT), we may assume that the profile-wise z-score approximately follows a **standard normal distribution**.

One-sided test

Large values are critical

The test statistic, Q, had a common population of origin in the numerator and different populations in the denominator. Thus, when **the data supports different populations**, Q will be **smaller than 1**.



However, taking the negative logarithm, it will be **large positive values that are critical to the null hypothesis**.

For a significance level of 5%, the critical value is **1.64**.

p-value

Normal approximation

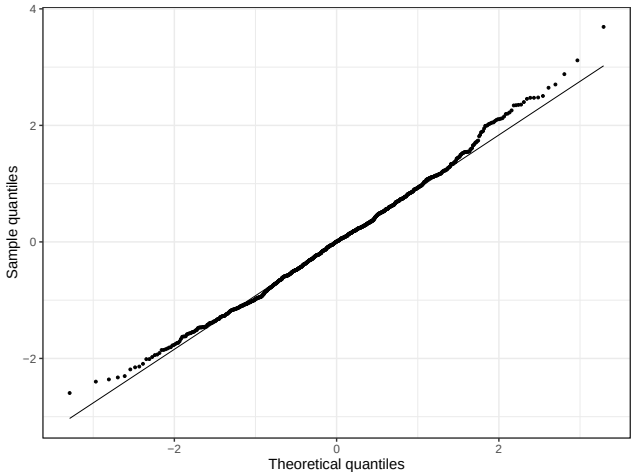
Hence, we can **compute a *p*-value** in order to **reject** or **accept** the **null hypothesis**.

However, simulations under the null hypothesis indicates that the **tails are slightly heavier** than a normal distribution (cf. next slide)

Furthermore, in cases where several alleles are close to being fixed in the population, the effective number of independent *stochastic* elements may be too small for the CLT. It may also be that only a **partial profile** is available causing the same challenge.

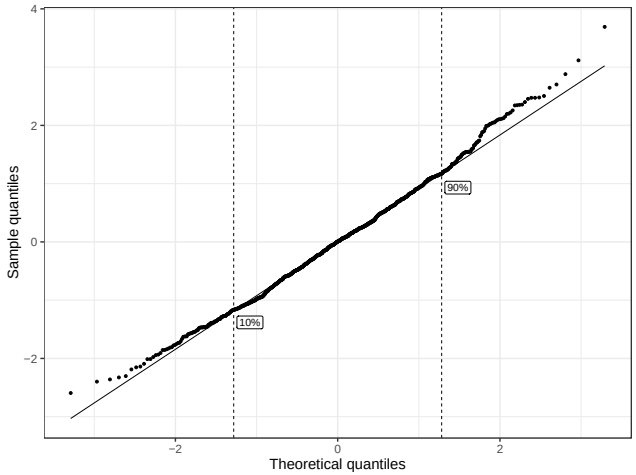
Approximately normal distributed

Slightly heavier trails – 1,000 simulated profiles



Approximately normal distributed

Slightly heavier trails – 1,000 simulated profiles



Importance sampling

Exponential tilting

Hence, if the p-value falls **below 0.10** we evaluate the p-value using **importance sampling**. Let $z(\boldsymbol{x}_0)$ denote z-score for a given profile, $\boldsymbol{x}_0$. The p-value of interest is

$$\hat{P}(z(\boldsymbol{x}_0^{\text{obs}}) \leq z(\boldsymbol{x}_0)) = \frac{1}{m}\sum_{i=1}^{m} \mathbb{I}(z(\boldsymbol{x}_0^{\text{obs}}) \leq z(\boldsymbol{x}_{0,i})) \frac{P(\boldsymbol{x}_{0,i} \mid \boldsymbol{x}_+)}{q(\boldsymbol{x}_{0,i} \mid \boldsymbol{x}_+; \theta)},$$

where q is the proposal distribution and $\boldsymbol{x}_{0,i} \sim q, i = 1, \ldots, m$.

We use **exponential tilting** to derive an efficient proposal distribution for simulating the p-value,

$$q(\boldsymbol{x}_0 \mid \boldsymbol{x}_+;\ \theta) = \exp\{\theta z(\boldsymbol{x}_0) - \kappa(\theta)\} P(\boldsymbol{x}_0 \mid \boldsymbol{x}_+),$$

where $\kappa(\theta)$ is the cumulant generating function.

Exponential tilting

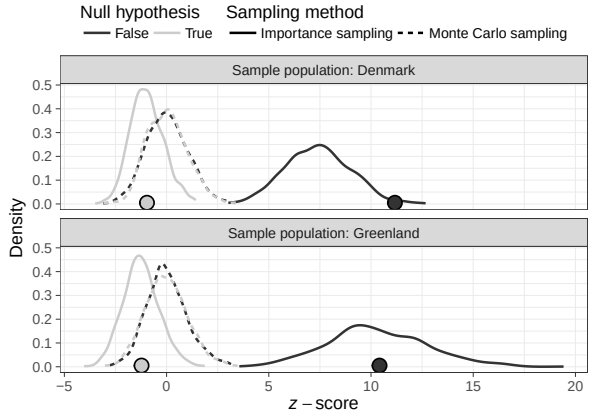
Example: non-zero p-values even in extreme cases

Profile sampled from "Sample population".

Density estimates of distribution used to estimate p-values.

Using the CLT approximation, the incorrect observation will get a p-value of 0.

Importance sampling fixes this issue.



GenoGeographer

Logic and Outcomes

From a logical point of view, an individual can belong to **at most one** of the exclusive (i.e. non-overlapping) reference populations.

For **each reference population**, the profile of interest is tested as belonging to that population.

However, based on the available information (i.e. genotyped loci and samples in the reference populations) it **may not be possible to reject** the incorrect null hypotheses. This implies that one or more populations can be the population of origin for the profile.

On the other hand, if the profile does not resemble any of the populations, then it will be **rejected in all of them**.

More accepted populations

LRs

In case more than one population has been accepted (null hypothesis not rejected), it implies that the profile is not too extreme relative to the observed allele frequencies of those populations.

However, one of them may still be more plausible than the others. In this case it is safe to use the pairwise likelihood ratios as the hypothesis tests have shown concordance between population samples and thee profile of interest.

Pairwise likelihood ratios

Variance of log likelihood ratios

To assess if the AIMs profile, $\boldsymbol{x}_0$, is more likely in population j than in population k, we compute the LR:

$$\widehat{LR}_{jk} = \frac{\hat{P}(\boldsymbol{x}_0 \mid H_j)}{\hat{P}(\boldsymbol{x}_0 \mid H_k)},$$

where $\hat{P}(\boldsymbol{x}_0 \mid H)$ is based on the estimates of allele frequencies, $\hat{p}_{jl}$.

Chakraborty et al. (1993) derived an expression of the variance of $\hat{P}(\boldsymbol{x}_0 \mid H_j)$, where the **variance increases** as the **sample size** n_j **decreases**.

The validity of the variance approximations depends on allele frequency **estimates** being **close** to the **true frequencies**.

We can use the variance estimates to construct approximate **95%-confidence intervals** (CI) for the LRs.

Evaluating the weight of evidence

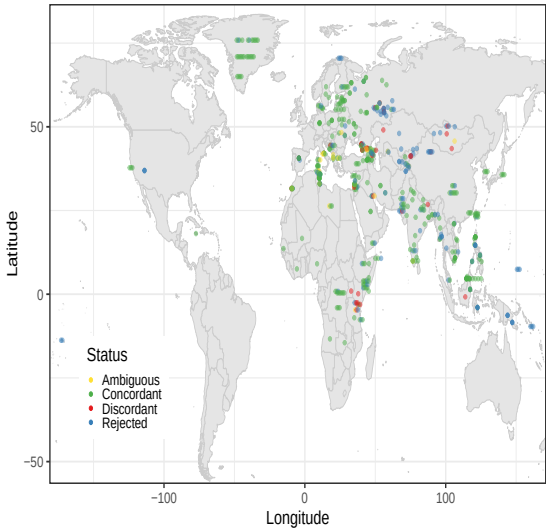
Decision rule

For **each population** among the reference populations, we **compute the z-score**.

- If **all null hypotheses are rejected**, we take this as evidence of the fact that **there is no relevant population** among the reference populations.
- If **one or more hypotheses are accepted**, we compute LRs, where at **least one** of the two populations in the ratio **was accepted** (i.e. has a p-value above the significance level, e.g. 0.05).
- In case of two (or more) accepted populations, the CI's can be used to assess if one population is significantly more plausible than the others.

Validation study

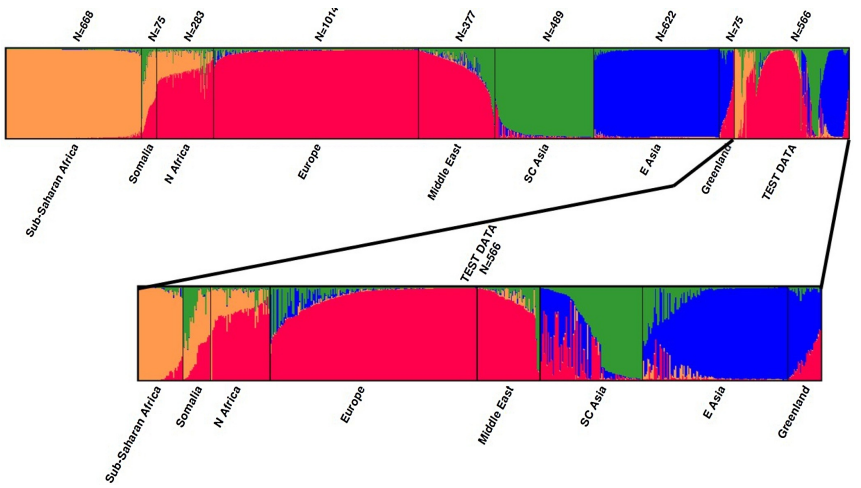
Geographically scattered test samples (566 samples from 90 different countries)



Joint work with Helle S. Mogensen, Claus Børsting, Vania Pereira and Niels Morling, published in FSI:Genetics (Mogensen et al., 2020).

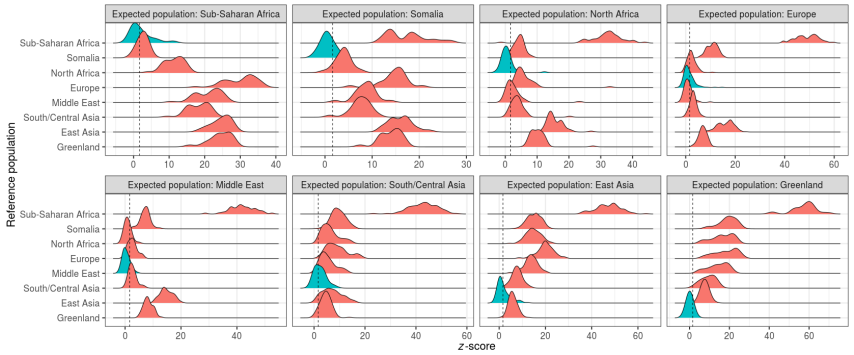
Validation study

STRUCTURE analysis



Validation study

Density estimates of z-scores. Large z-scores critical to hypothesis



Brier score

Deeper insight into STRUCTURE (1/3)

For each of the reference samples, STRUCTURE estimates the admixture components, $\boldsymbol{q}_i = (q_i^{(1)}, \ldots, q_i^{(K)})$, where $\sum_{k=1}^{K} q_i^{(k)} = 1$ by definition.

Based on these, we may compute the average admixture component for each of the reference metapopulations, by taking the average over the admixture components for the samples belonging to this meta-population:

$$\bar{\boldsymbol{q}}_j = (\bar{q}_j^{(1)}, \ldots, \bar{q}_j^{(K)}), \quad \text{with} \quad \bar{q}_j^{(k)} = n_j^{-1} \sum_{i \in R_j} q_i^{(k)}, \quad k = 1, \ldots, K,$$

where R_j is the set of the n_j samples from meta-population j.

Brier score

Deeper insight into STRUCTURE (2/3)

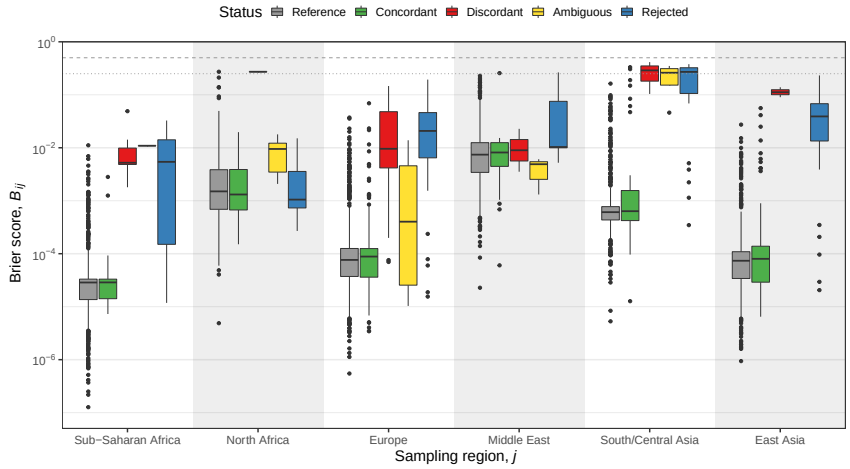
STRUCTURE admixture components was also computed for each of the test samples. For each test sample i, we computed a Brier score deviation between the test sample and the average for the reference metapopulation j:

$$B_{ij} = \frac{1}{K}\sum_{k=1}^{K}(q_i^{(k)} - \bar{q}_j^{(k)})^2.$$

The closer B_{ij} is 0 the more similar is sample i to metapopulation j in terms of STRUCTURE admixture components.

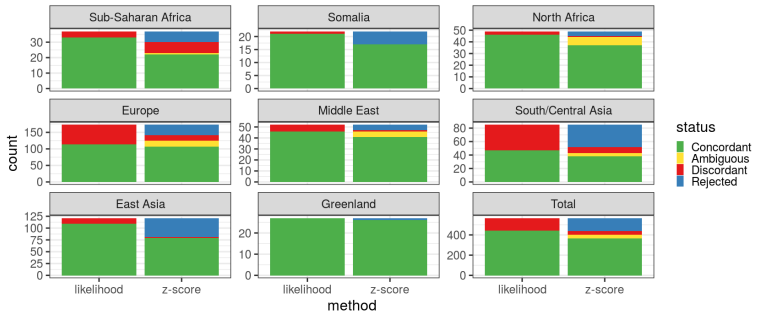
Brier score

Deeper insight into STRUCTURE (3/3)



Validation study

– A reduction in the error rate by a factor of three!



Based on the **likelihood** the **error rate is 21.9%**. When when using the ***z*-score approach** with the rejection and ambiguous options, the **error rate is 8.2%**.

Admixed profiles

1st order: Parents from different populations

Our validation study indicated that some of the profiles may have had **admixed origin** – i.e. parents from different populations.

The only methodological adjustment is to allocate the **ambiguous alleles** at **heterozygous markers** to one of the parental populations (for homozygous markers same allele is inherited from each population):

$$P(\text{Allele comes from population } j \mid x_0 = 1) = \frac{p_j + p_\bullet}{p_+ - 2p_\bullet},$$

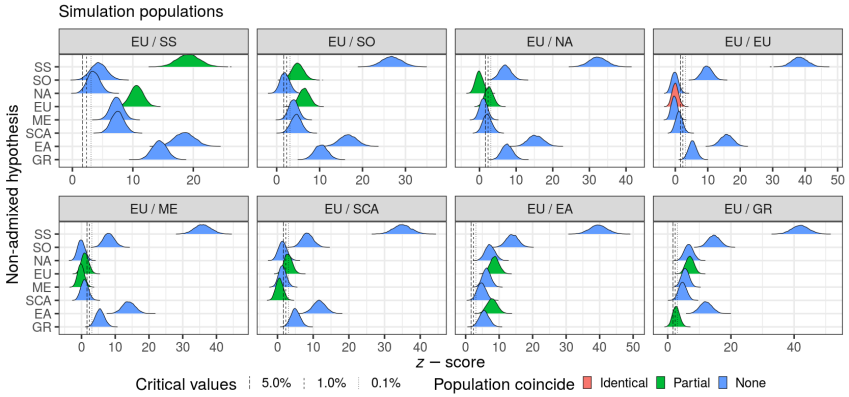
where $p_\bullet = p_1 p_2$ and $p_+ = p_1 + p_2$ and p_j is the allele frequency in population j.

Thus, if $p_2 \approx 0$, we have that $p_\bullet \approx 0$ and $p_+ \approx p_1$ suggesting that $P(\text{Allele comes from population } 1 \mid x_0 = 1) \approx 1$ (and conversely for $j = 2$).

Also, if $p_1 \approx p_2$ we have that $p_\bullet \approx p_1^2$ and $p_+ \approx 2p_1$ yielding $P(\text{Allele comes from population } 1 \mid x_0 = 1) \approx 1/2$ as expected.

1st order admixed profiles

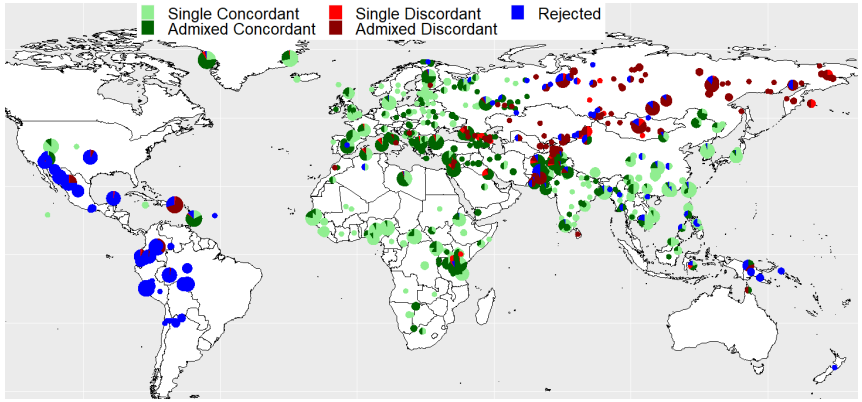
Simulation study - z-scores (excerpt of European admixtures)



SS: Sub-Saharan Africa
SO: Somalia
NA: North Africa
EU: Europe
ME: Middle East
SCA: South/Central Asia
EA: East Asia
GR: Greenland.

Recent validation study (Poster P138)

Update of AIMS population data and test with the GenoGeographer admixture module



Summary

- **Pairwise likelihood ratios are not sufficient** for assessing the weight of evidence for AISNP profiles
- The likelihood ratio test (z-score) **is not dependent on known allele frequencies**
- Its similarity to **Fisher's exact test** ensures a sound statistical approach
- The `GenoGeographer.org` enables fast and flexible analysis using a well-defined framework
- The use of **metapopulations** reduces the risk of making too specific statements about the country/area/population of origin
- **1st order admixtures** (and in theory, higher order) can be handled within the same framework.

References (1/8)

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

Chakraborty, R., M. R. Srinivasan, and S. F. Daiger (1993). Evaluation of standard error and confidence intervals of estimated multilocus genotype probabilities and their implications in dna forensics. *Am J Hum Genet 52*, 60–70.

Cheung, E. Y., M. E. Gahan, and D. McNevin (2017). Prediction of biogeographical ancestry from genotype: a comparison of classifiers. *Int J Legal Med 131*, 901–912.

Cheung, E. Y. Y., M. E. Gahan, and D. McNevin (2018). Prediction of biogeographical ancestry in admixed individuals. *Forensic Sci Int Genet 36*, 104–111.

References (2/8)

Falush, D., M. Stephens, and J. K. Pritchard (2003). Inference of population structure using multilocus genotype data: Linked loci and correlated allele frequencies. *Genetics 164*, 1567–1587.

Falush, D., M. Stephens, and J. K. Pritchard (2007). Inference of population structure using multilocus genotype data: dominant markers and null alleles. *Mol. Ecol. Notes 7*, 574–578.

Hellenthal, G., G. B. J. Busby, G. Band, J. F. Wilson, C. Capelli, D. Falush, and S. Myers (2014). A genetic atlas of human admixture history. *Science 343*(6172), 747–751.

Jombart, T., S. Devillard, and F. Balloux (2010). Discriminant analysis of principal components: a new method for the analysis of genetically structured populations. *BMC Genetics 11*(94), 1–15.

Kidd, K. K. et al. (2014). Progress toward an efficient panel of SNPs for ancestry inference. *Forensic Sci Int Genet 10*, 23–32.

References (3/8)

Kosoy, R., et. al, and M. F. Seldin (2009). Ancestry Informative Marker Sets for determining continental origin and admixture proportions in common populations in America. *Hum Mutat 30*(1), 69–78.

Lawson, D. (2018). *badMIXTURE: Validating Structure With Chromosome Painting*. R package version 0.0.0.9000.

Lawson, D. J., G. Hellenthal, S. Myers, and D. Falush (2012, 01). Inference of population structure using dense haplotype data. *PLOS Genetics 8*(1), 1–16.

Lawson, D. J., L. van Dorp, and D. Falush (2018). A tutorial on how not to over-interpret STRUCTURE and ADMIXTURE bar plots. *Nat Commun 9*(3258).

References (4/8)

McNevin, D., C. Santos, A. Gómez-Tato, J. Álvarez-Dios, M. Casares de Cal, R. Daniel, C. Phillips, and M. V. Lareu (2013). Anassessment of bayesian and multinomial logistic regression classification systems to analyse admixed individuals. *Forensic Science International: Genetics Supplement Series 4*, e63–e64.

McVean, G. (2009). A genealogical interpretation of principal components analysis. *PLOS Genetics 5*(10), 1–10.

Menozzi, P., A. Piazza, and L. L. Cavalli-Sforza (1978). Synthetic maps of human gene frequencies in europeans. *Science 201*, 786–792.

Miller, J. M., C. I. Cullingham, and R. M. Peery (2020). The influence of a priori grouping on inference of genetic clusters: simulation study and literature review of the DAPC method. *Heredity 125*(5), 269–280.

References (5/8)

Mogensen, H. S., T. Tvedebrink, C. Børsting, V. Pereira, and N. Morling (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.

Novembre, J. (2014). Variations on a common STRUCTURE: New algorithms for a valuable model. *Genetics 197*(3), 809–811.

Novembre, J. (2016). Pritchard, stephens, and donnelly on population structure. *Genetics 204*(2), 391–393.

Novembre, J. et al. (2008). Interpreting principal component analyses of spatial population genetic variation. *Nat Genet 40*, 646–649.

Patterson, N., A. L. Price, and D. Reich (2006). Population structure and eigenanalysis. *PLoS Genetics 2*(12), e190.

References (6/8)

Porras-Hurtado, L., Y. Ruiz, C. Santos, C. Phillips, Á. Carracedo, and M. Lareu (2013). An overview of STRUCTURE: applications, parameter settings, and supporting software. *Frontiers in Genetics 4*, 98.

Price, A., N. Patterson, R. Plenge, et al. (2006). Principal components analysis corrects for stratification in genome-wide association studies. *Nat Genet 38*, 904–909.

Pritchard, J. K., M. Stephens, and P. J. Donnelly (2000). Inference of population structure using multilocus genotype data. *Genetics 155*, 945–959.

Qu, Y., D. Tran, and W. Ma (2019). Deep learning approach to biogeographical ancestry inference. *Procedia Computer Science 159*, 552–561.

References (7/8)

Raj, A., M. Stephens, and J. K. Pritchard (2014). faststructure: Variational inference of population structure in large snp data sets. *Genetics 197*, 573–589.

Rosenberg, N., L. Li, R. Ward, and J. Pritchard (2003). Informativeness of genetic markers for inference of ancestry. *American Journal of Human Genetics 73*, 1402–1422.

Rosenberg, N. A. (2005). Algorithms for selecting informative marker panels for population assignment. *Journal of Computational Biology 12*(9), 1183–1201.

Santos, C., C. Phillips, A. Gomez-Tato, J. Alvarez-Dios, Á. Carracedo, and M. V. Lareu (2016). Inference of ancestry in forensic analysis ii: Analysis of genetic data. In W. Goodwin (Ed.), *Forensic DNA Typing Protocols*, Volume 1420 of *Methods in Molecular Biology*, pp. 255–285. Springer.

References (8/8)

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

Tvedebrink, T. (2022). Review of the forensic applicability of biostatistical methods for inferring ancestry from autosomal genetic markers. *Genes 13*(1), 141.

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

Wangkumhang, P. et al. (2018). Statistical methods for detecting admixture. *Current Opinion in Genetics & Development 53*, 121–127.