genogeographer – a tool for ancestry informative markers

useR! 2019 – Toulouse

Torben Tvedebrink[†,‡]
joint work with **Poul Svante Eriksen**[†]

† Department of Mathematical Sciences, Aalborg University

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

Forensic genetics

Genetic markers

The scientific evaluation of DNA and other genetic information to be used in the judiciary system is the domain of forensic geneticists. The typical use of DNA in crime cases relates to **identification cases** and **DNA mixtures**. Legal disputes involving DNA in civil cases are mostly **paternity testing** and **pedigree analysis**.

A modern DNA profile consists of either **STR**s (Short Tandem Repeat, which are repetitive genetic sequences), or **SNPs** (Single Nucleotide Polymorphisms, population variation on a single base).

Ancestry Informative Markers

AIMs

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

This presentation is not about identifying new markers, but to make **proper inference** of the results of a pre-selected **set of markers**.

Specifically, the SNP set considered here is the Applied Biosystems^{TM} **Precision ID Ancestry Panel** (containing **165 SNPs**).

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**, $\boldsymbol{x}_0$, has **0**, **1 or 2** copies of **allele 1**.

Well-defined hypotheses
Likelihood ratios

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

For the usual forensic cases, e.g. identification cases, 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.

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:

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

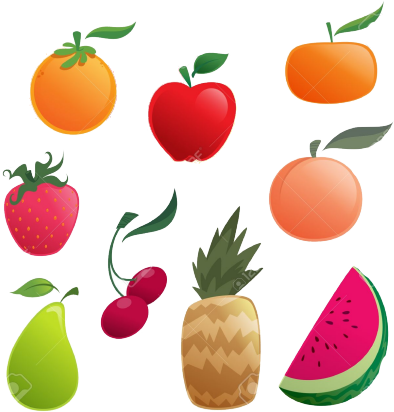
where $\hat{P}(\boldsymbol{x}_0 \mid H_i)$ is based on the estimates of allele frequencies, $\hat{p}_i$.

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 the frequency **estimates** being **close** to the **true frequencies**. For small sample sizes this is almost certainly **not** true!

Are we comparing nonsense with rubbish?

Exclusive – but not exhaustive



Are we comparing nonsense with rubbish?

Exclusive – but not exhaustive

We may have **exclusive** sample populations, but we do not have **exhaustive** databases of reference populations.

Hypothesis
The population of origin

To overcome the focus on relative frequencies of AIMs profiles, we propose to use a statistical likelihood ratio test framework. 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, $\boldsymbol{x_0}$, **originates** from population j

H_1: The AIMs profile, $\boldsymbol{x_0}$, **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.

Hence, we can standardise the LRT statistic to calculate a z-score, which indicates that a **large deviation** between the expected and observed genotype relative to the standard deviation is **evidence against the null hypothesis**.

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, where x_j is the allele count in sample j.

In the denominator, we assume **two different populations**. Hence, we estimate the allele frequencies separately:

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

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

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**.

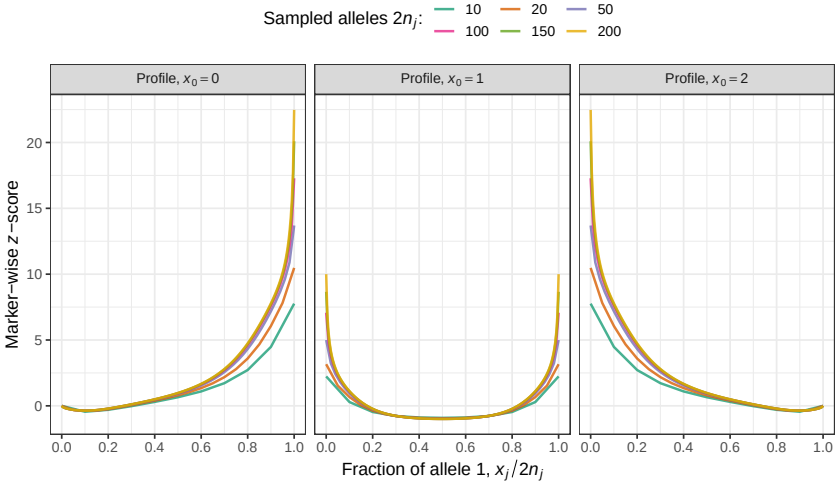
The distribution of $x_0 \mid x_+$ is **hyper-geometric**. Hence, the expectation and variance are easily computed over x_0 as this only takes the values of $\{0, 1, 2\}$.

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



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

The p-value can also be estimated using **importance sampling**, where **exponential tilting** is used as an efficient approach to derive a proposal distribution.

Due to the LRT approach, the test is **one-sided** with large values being critical to the null hypothesis.

Evaluating the weight of evidence

Decision rule

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

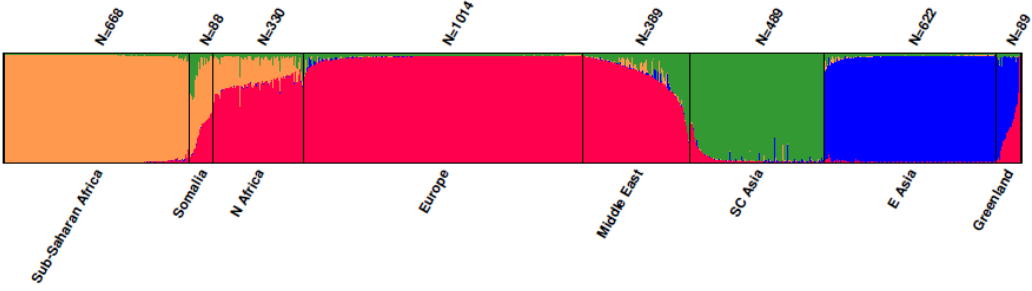
- If **all null hypotheses are rejected**, we take this as evidence of the fact that **there is no relevant population** among the reference populations.
- If **one or more hypotheses are accepted**, we compute LRs, where at **least one** of the two populations in the ratio **was accepted** (i.e. has a p-value above the significance level, e.g. 0.05).

Meta populations

Structure analysis

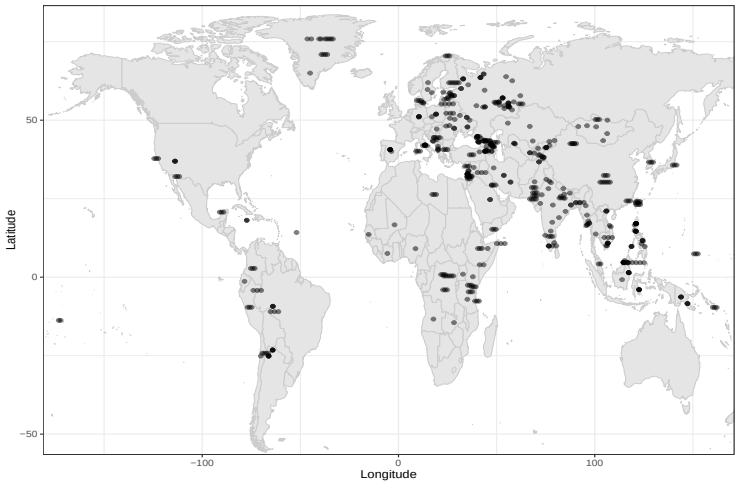
Our reference database consists of **publicly available** population samples from **36 populations**, which has been supplemented with SNP profiles from our own bio bank.

We used the software *Structure* to identify clusters among the samples. We identified **eight** clusters, also called **meta populations** with similar distribution of AISNPs.



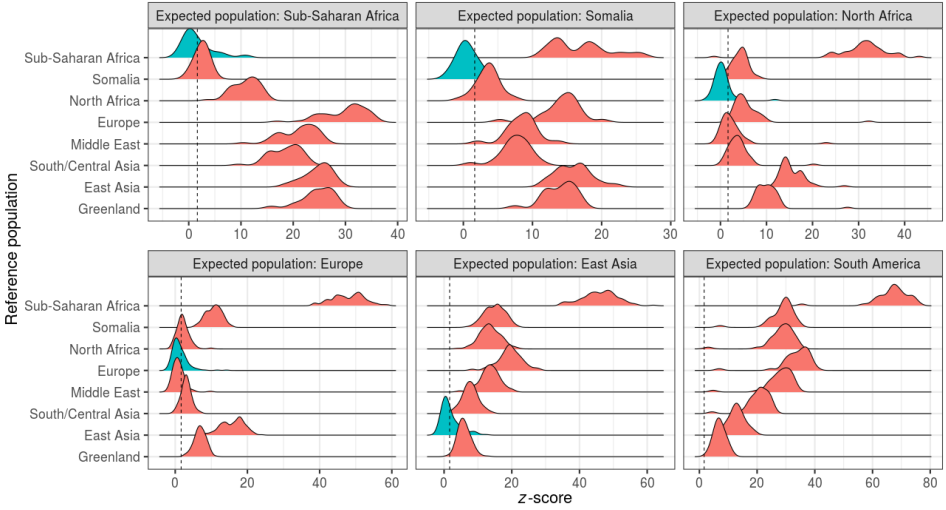
Validation study

Geographically scattered test samples (608 samples from 90 different countries)



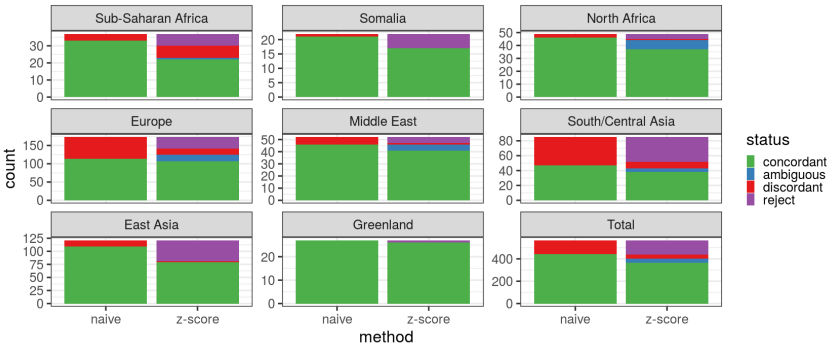
Validation study

Density estimates of z-scores (excerpt)



Validation study

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



For the **naïve method** the **error rate is 26.2%**. When when using the ***z*-score approach** with the rejection and ambiguous options, the **error rate is 8.1%**.

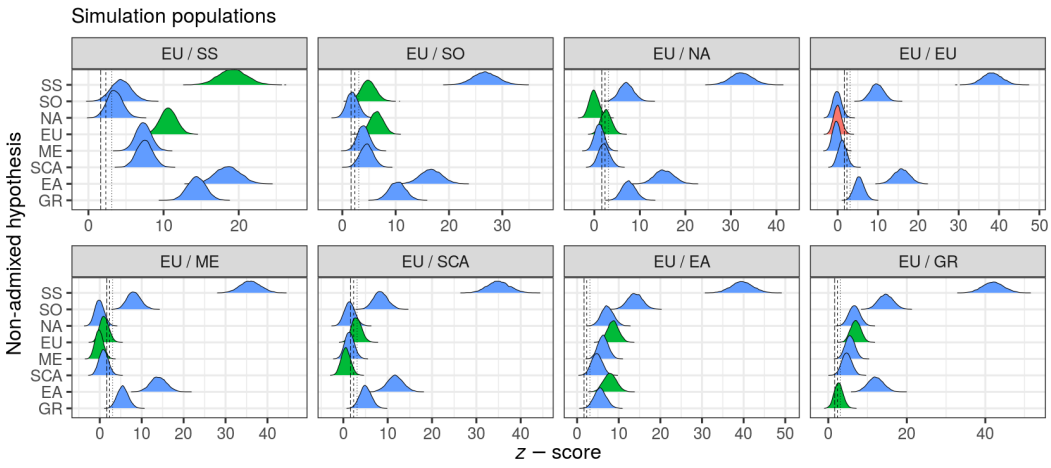
Admixed profiles

Parents from different populations

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

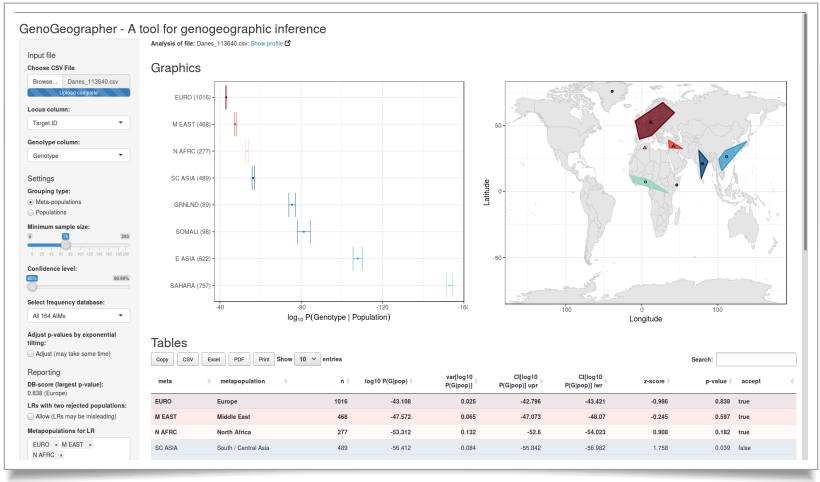
To account for this we reformulate our LRT approach to handle two reference populations. The only methodological adjustment was to use the **EM algorithm** to assign the **ambiguous alleles** at **heterozygous markers** to a single population by a latent variable.

Admixed profiles
Simulation study



GenoGeographer.org

Freely available R implementation with Shiny application front-end



Also available as a R package on CRAN: `genogeographer`

genogeographer package

z-score calculations, genogeographer::genogeo()


```
> z_score <- genogeographer::genogeo(x0, grouping = "meta", ...)
> z_score %>% select(...)
# A tibble: 8 x 7
  metapopulation         logP logP_lwr logP_upr  z_score  p_value accept
  <chr>                 <dbl>    <dbl>    <dbl>    <dbl>    <dbl> <lgl>
1 Middle East           -42.7    -43.2    -42.2 -1.05    8.53e- 1 TRUE
2 Europe                -43.4    -43.7    -43.0 -0.0766  5.31e- 1 TRUE
3 South / Central Asia  -47.7    -48.2    -47.2  0.00372 4.99e- 1 TRUE
4 North Africa          -48.5    -49.3    -47.7  0.253   4.00e- 1 TRUE
5 Greenland             -65.7    -67.2    -64.2  4.82    7.05e- 7 FALSE
6 Somalia               -72.2    -74.8    -69.7  8.17    1.52e-16 FALSE
7 East Asia             -83.2    -84.3    -82.0 11.0     1.35e-28 FALSE
8 Sub-Saharan Africa   -163.    -166.    -160.  45.2     0.       FALSE
```

genogeographer package

LR calculations, genogeographer::LR_table()


```
> genogeographer::LR_table(z_score)
# A tibble: 22 x 7
   numerator   denominator            logLR var_logLR   CI_lwr CI_upr null_in_CI
   <chr>       <chr>                  <dbl>     <dbl>    <dbl>  <dbl> <lgl>
 1 Middle East Europe                 0.650    0.0917   0.0565   1.24 FALSE
 2 Middle East South / Central Asia   5.02     0.124    4.33     5.71 FALSE
 3 Middle East North Africa           5.76     0.222    4.84     6.68 FALSE
 4 Middle East Greenland             23.0      0.632   21.4     24.6  FALSE
 5 Middle East Somalia               29.5      1.81    26.9     32.2  FALSE
 6 Middle East East Asia             40.4      0.386   39.2     41.7  FALSE
 7 Middle East Sub-Saharan Africa   120.       2.07   117.     123.   FALSE
 8 Europe      South / Central Asia   4.37     0.0901   3.78     4.96 FALSE
 9 Europe      North Africa           5.11     0.187    4.26     5.96 FALSE
10 Europe      Greenland             22.3      0.598   20.8     23.9  FALSE
# ... with 12 more rows
```

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 **meta populations** reduces the risk of making too specific statements about the country/area/population of origin

Future research
Some perspectives

- **Higher-order admixture** - How to extend current method to e.g. 2nd order-admixture?
- **Relaxation of independence** assumption. Our current PhD student is currently looking implementing outlier detection models for dependent markers.
- **Cascade analysis** - for some hypotheses only subsets of the original SNPs are informative.
- How to deal with **DNA mixtures** (cases with DNA from more than one contributor)?

Thank you for your attention!

. . . and References

Tvedebrink T, Eriksen PS, Mogensen HS and Morling N (2018). *Weight of the Evidence of Genetic Investigations of Ancestry Informative Markers*. Theoretical Population Biology 120: 1-10.

Tvedebrink T and Eriksen PS (2019). *Inference of admixed ancestry with Ancestry Informative Markers*. Forensic Science International: Genetics, *(in press)*.

Kosoy R et al. (2009). *Ancestry informative marker sets for determining continental origin and admixture proportions in common populations in America*. Human Mutation 30: 69-78.

Kidd KK et al. (2014). *Progress toward an efficient panel of SNPs for ancestry inference*. Forensic Science International: Genetics 10: 23-32.

Chakraborty R et al. (1993). *Evaluation of standard error and confidence intervals of estimated multilocus genotype probabilities and their implications in DNA*. American Journal of Human Genetics 52: 60-70.

Pakstis A, et al. (2017). *Increasing the reference populations for the 55 AISNP panel: the need and benefits*. Int. Journal of Legal Medicine 131(4): 913-917.

Pagani L, et al. (2016). *Genomic analyses inform on migration events during the peopling of Eurasia*. Nature 538: 238-242.