



RECAP

from the  
Health Data Science  
Sandbox



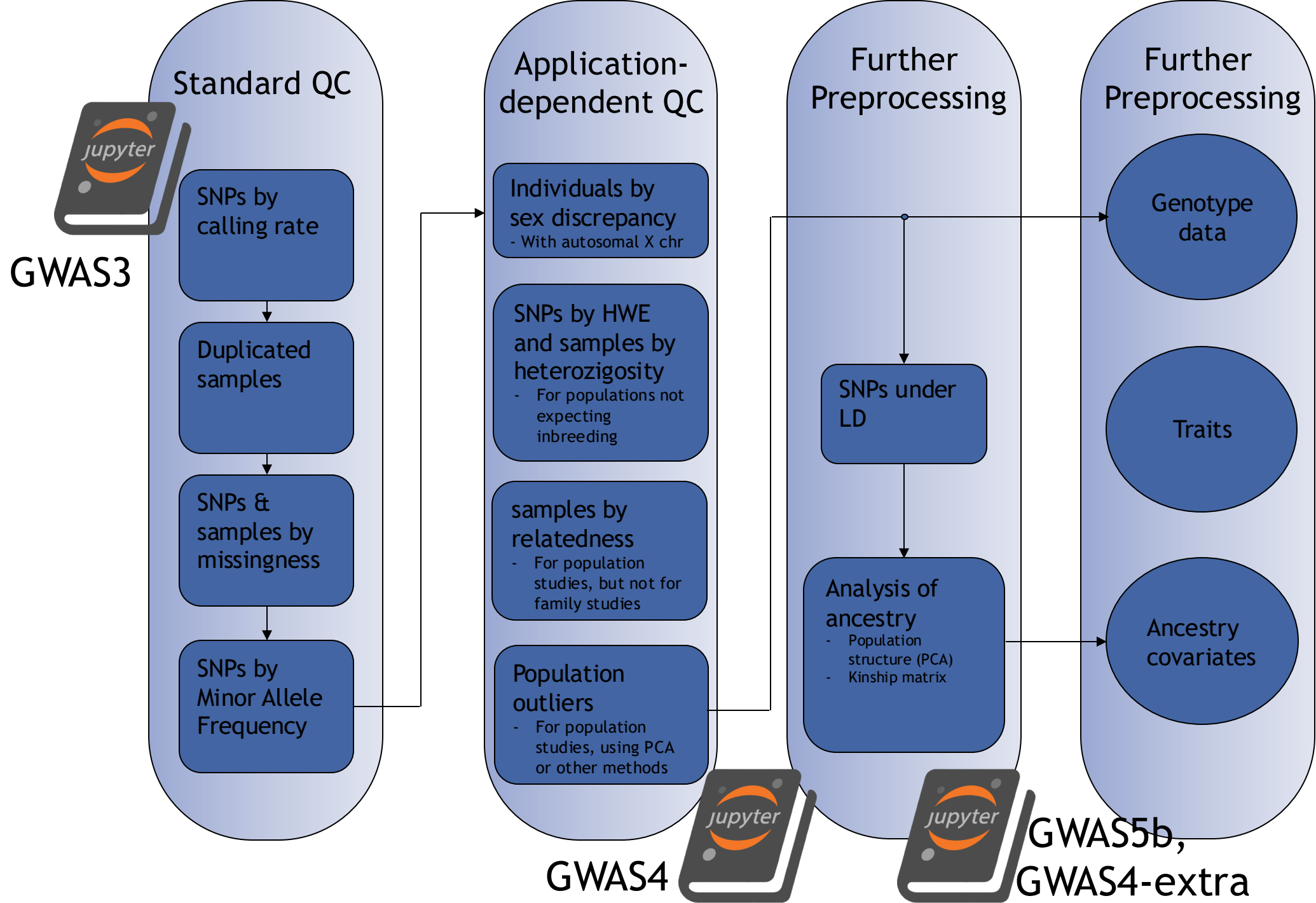
**Alba/Samuele, PhD**

Sandbox Data scientist

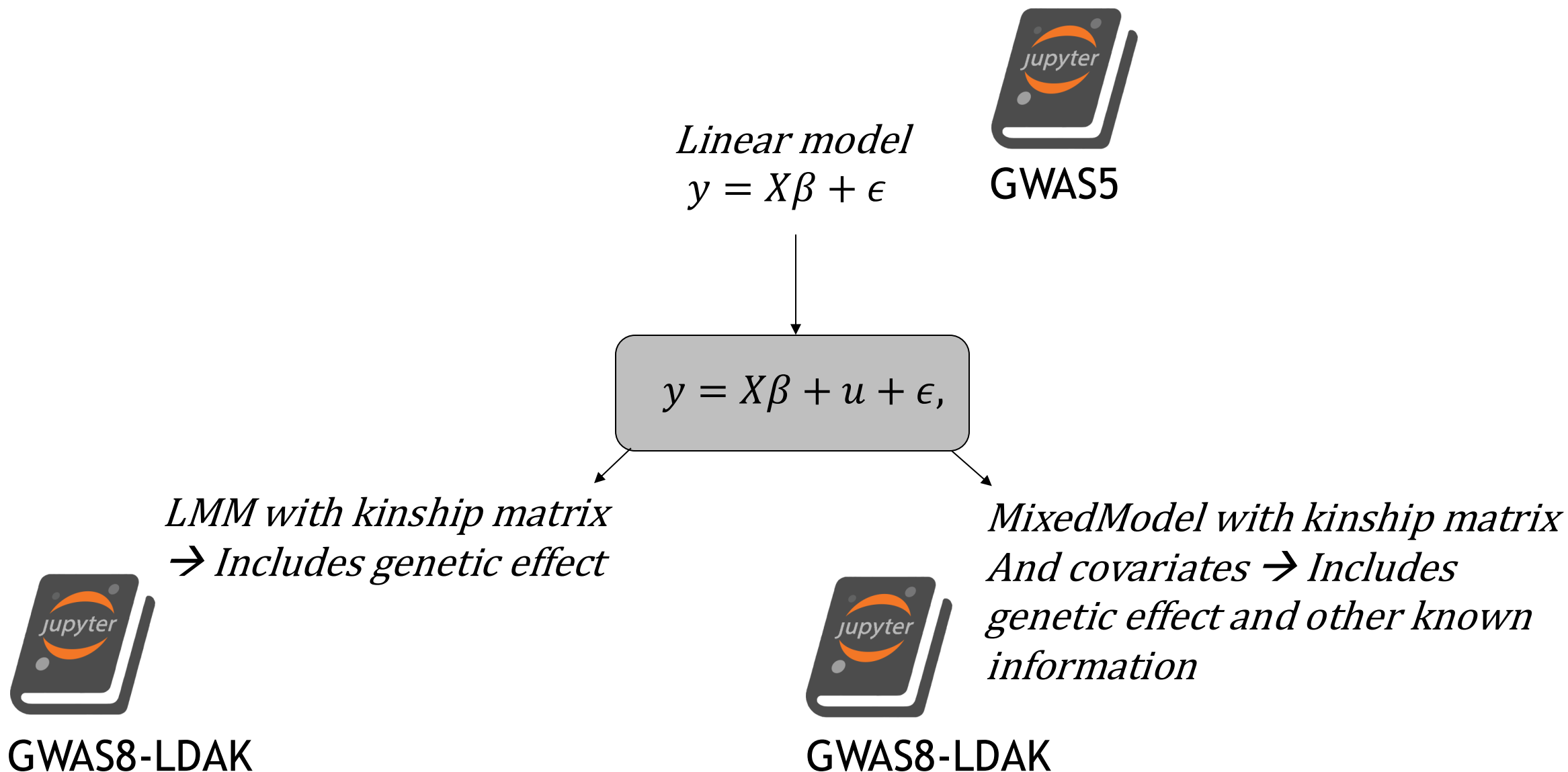
Center for Health Data Science (HeaDS)

UNIVERSITY OF  
COPENHAGEN





# Association testing



# Polygenic Risk Score



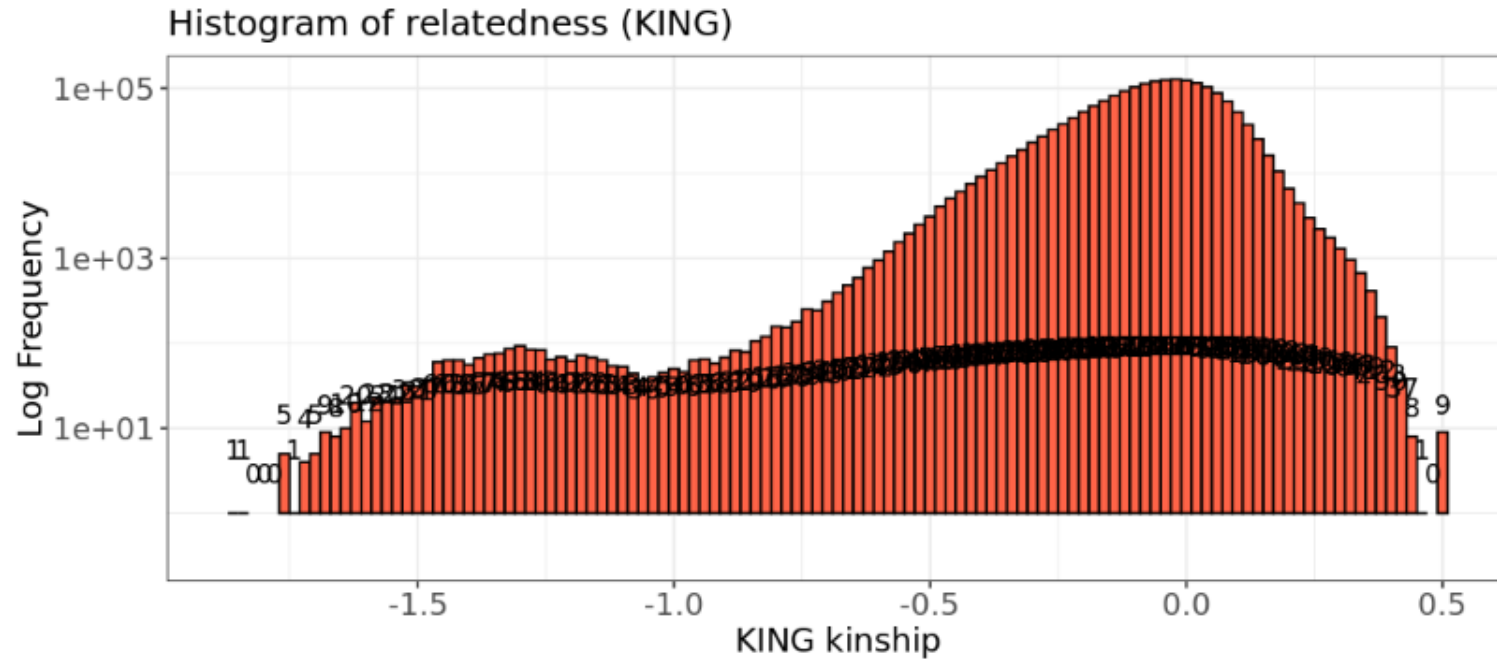
GWA6



GWAS7

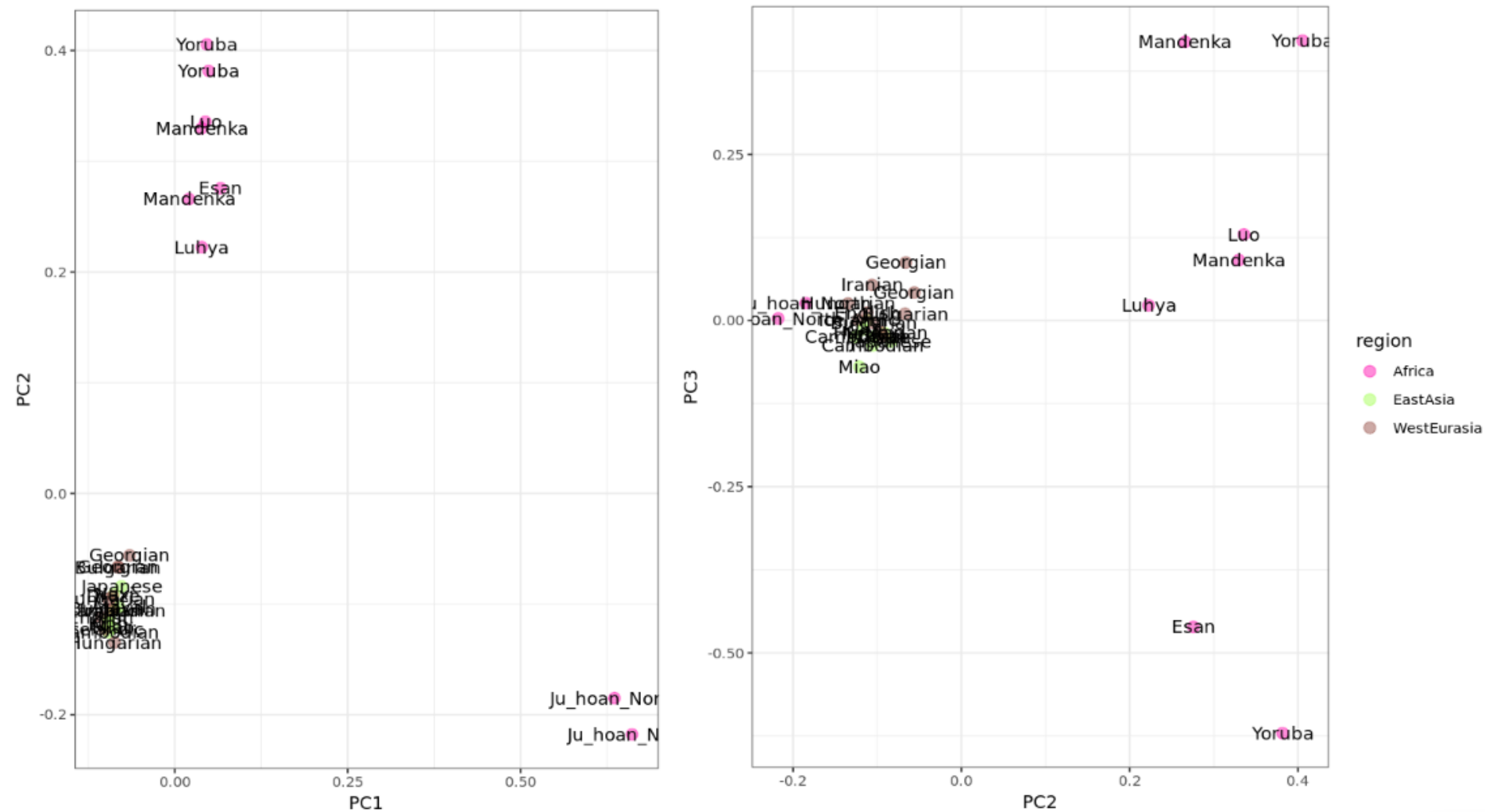
## Relatedness in mice data

*Only 8 founders for ~1900 mice!*

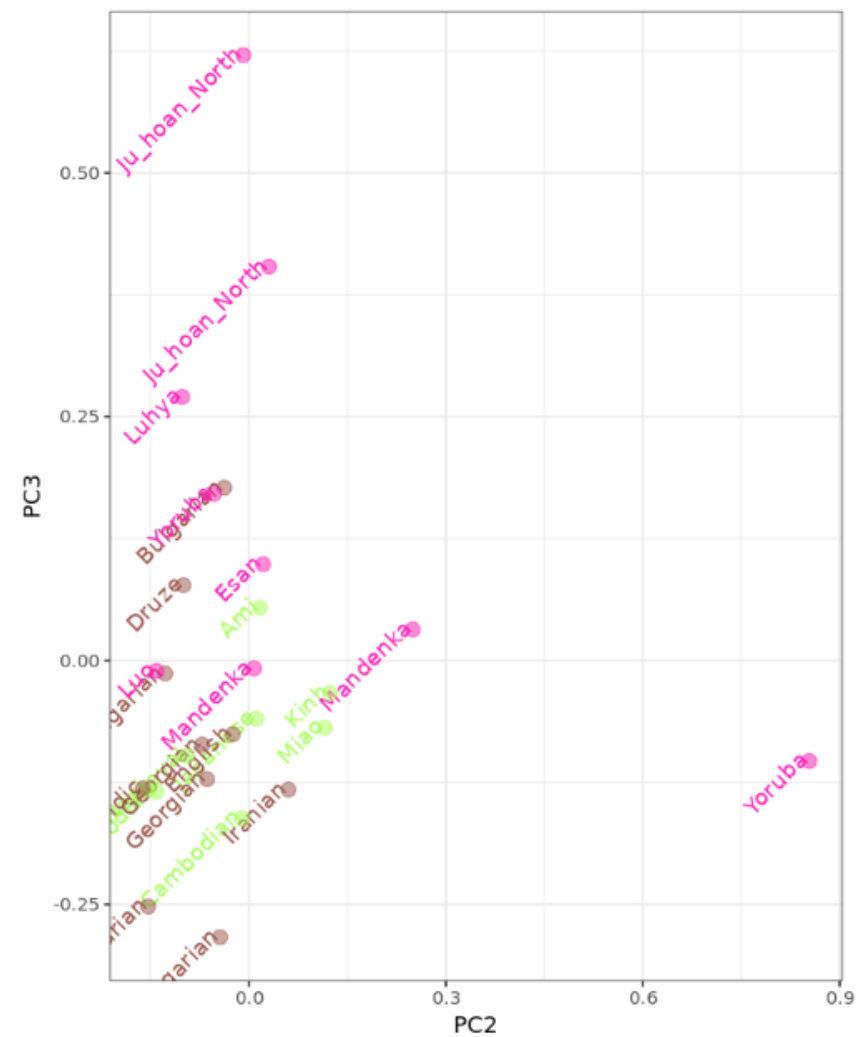


- KING's algorithm is for **outbred** populations
- Founders related to a lot of data can cause huge shift in  $\text{KING} < 0$

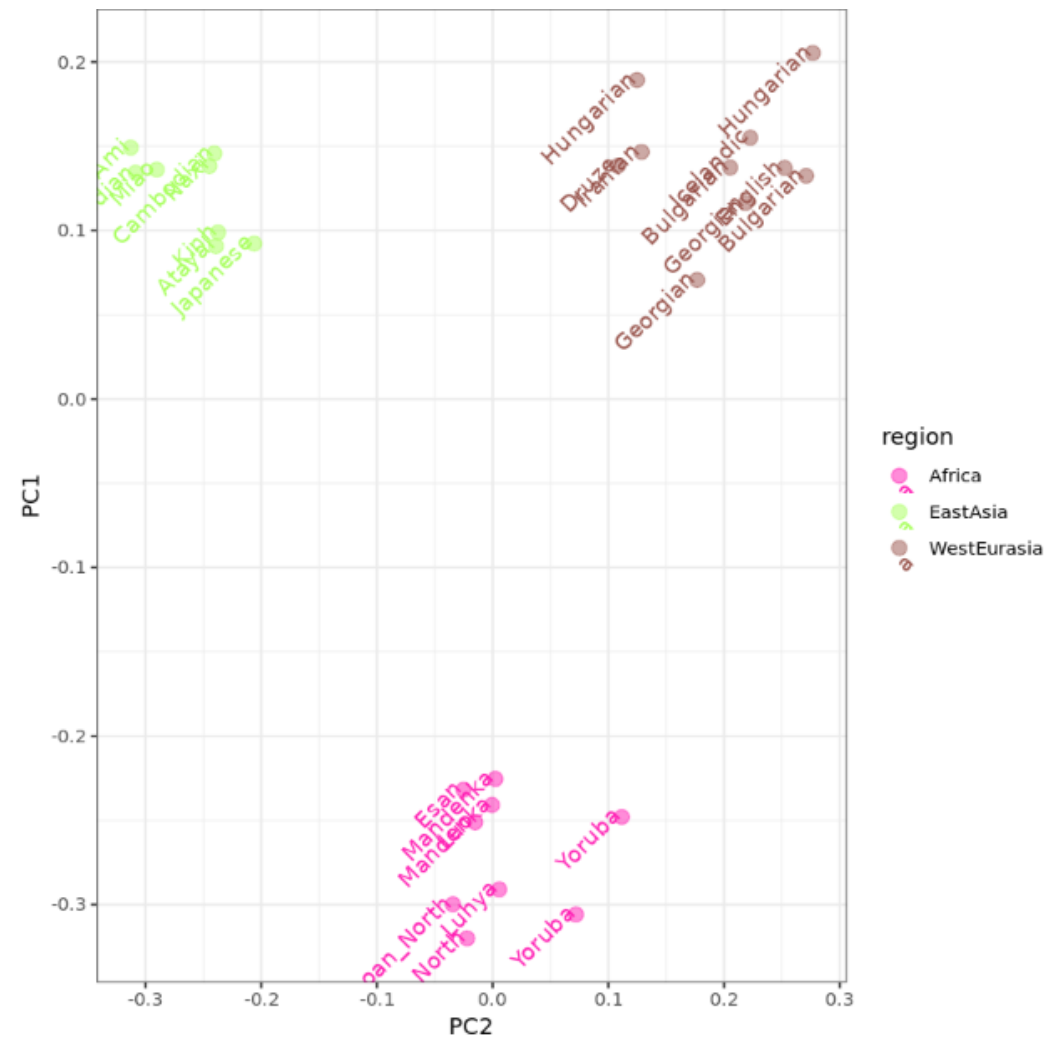
# Pop structure and LD



# Pop structure and LD

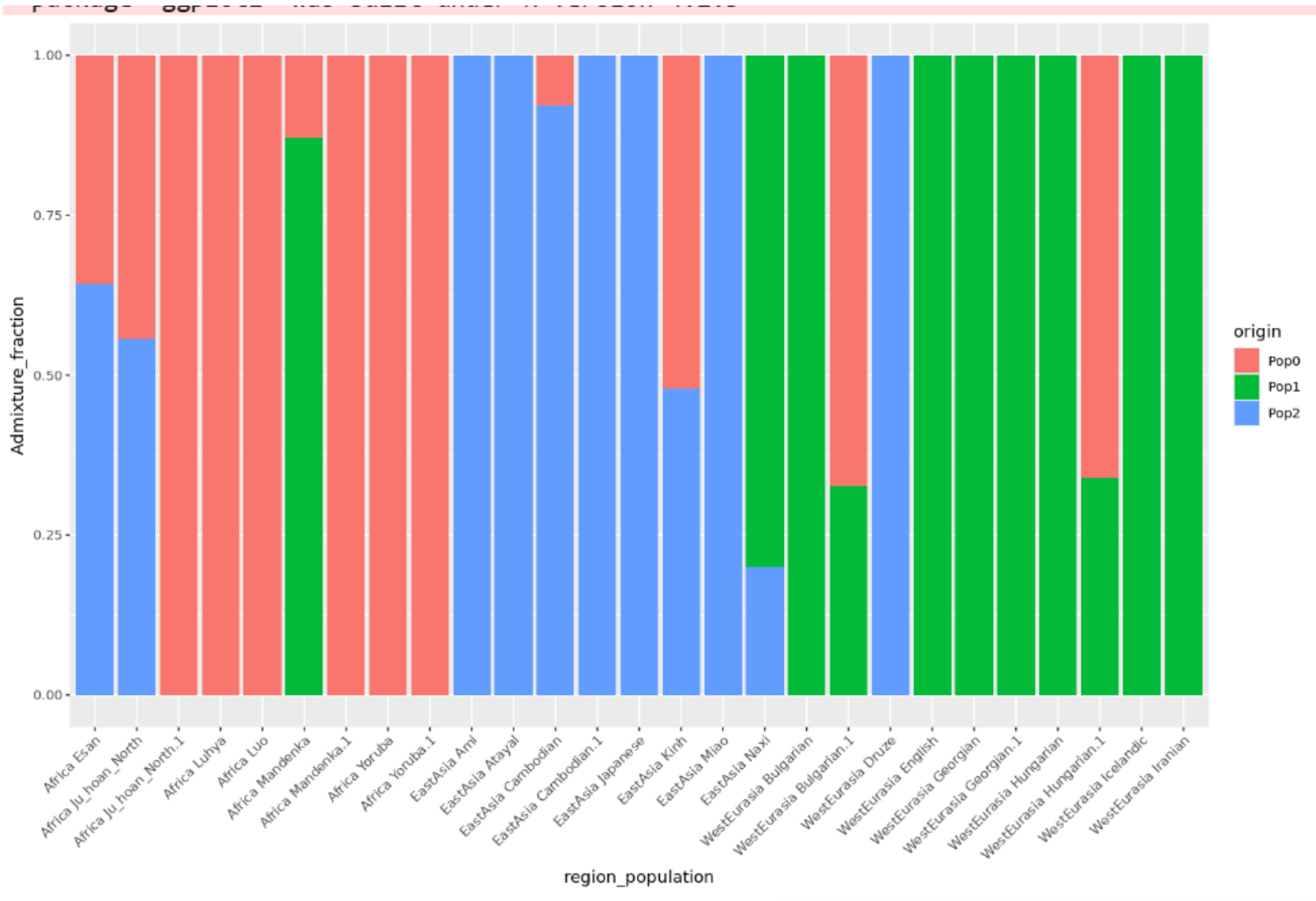


LD<.2



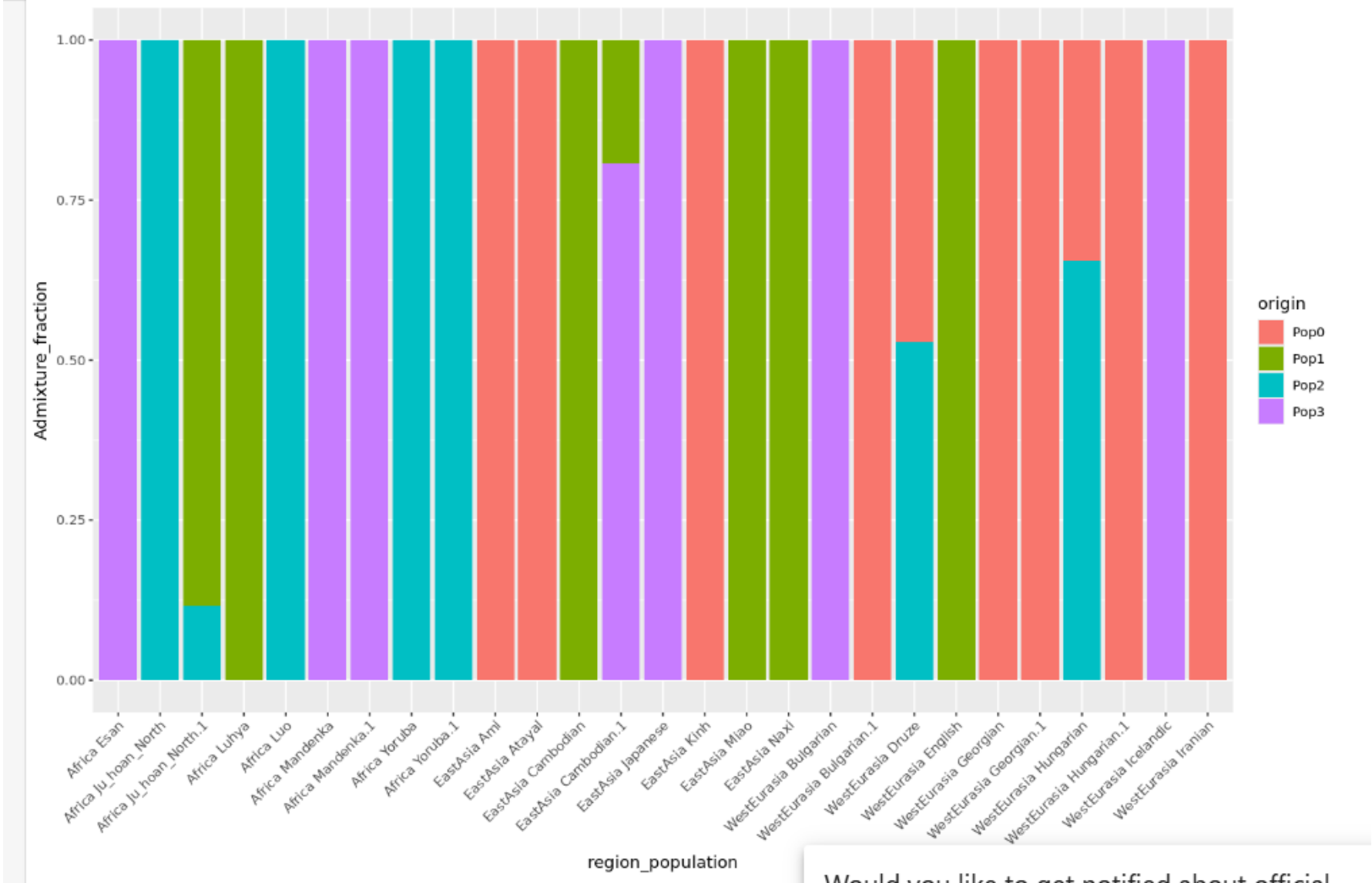
LD<.5

# Pop structure and LD

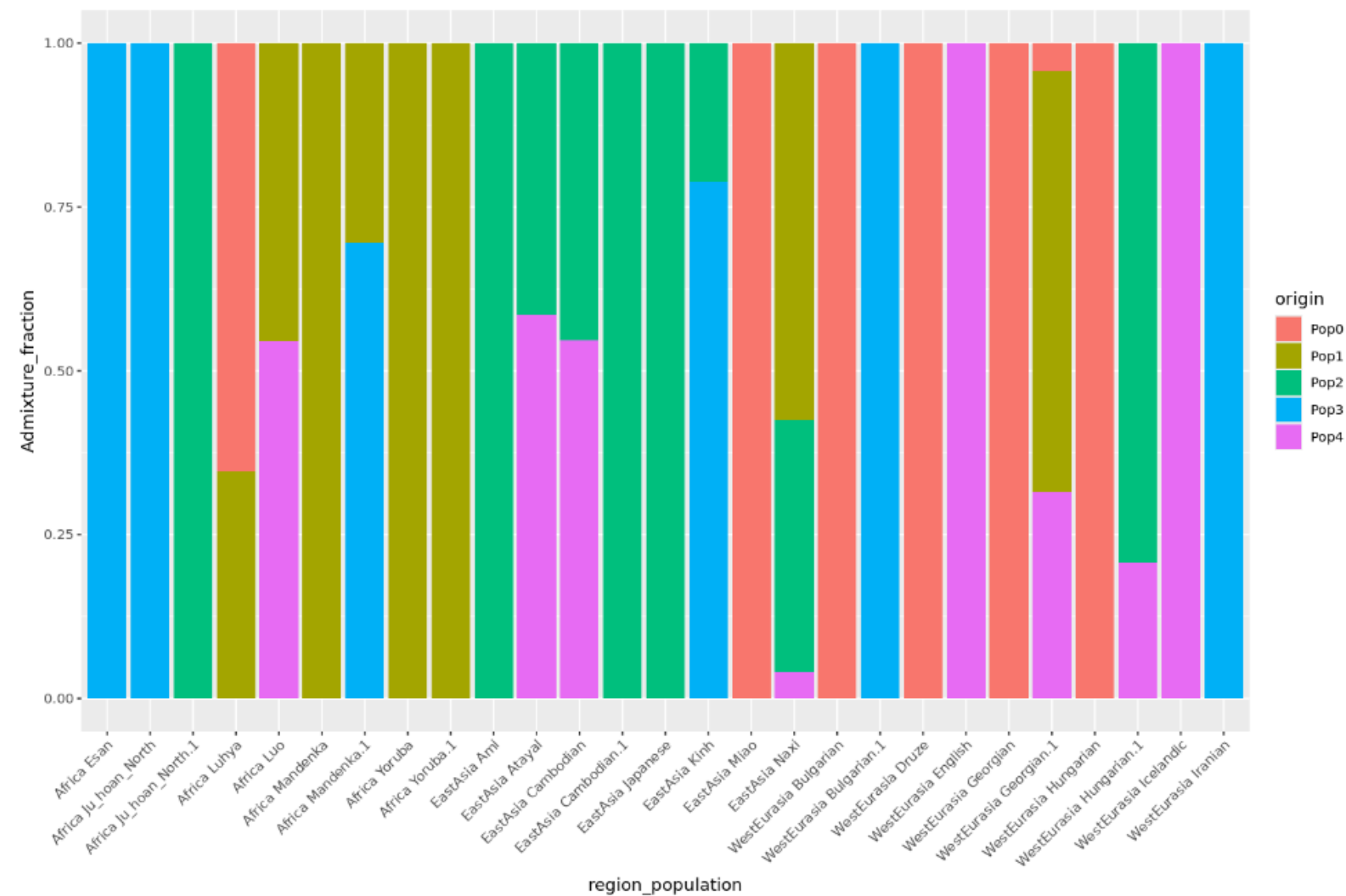




# Pop structure and LD



# Pop structure and LD



# LD pruning + clumping

Highly recommend to read this paper. Check the effect that LD pruning and clumping has in the lead SNPs and the association signal



GENETICS

GENETICS, 2025, iyaf009

<https://doi.org/10.1093/genetics/iyaf009>

Advance Access Publication Date: 5 February 2025

Investigation

## Measuring linkage disequilibrium and improvement of pruning and clumping in structured populations

Ulises Bercovich ,<sup>1,†</sup> Malthe Sebro Rasmussen ,<sup>2,†</sup> Zilong Li ,<sup>2</sup> Carsten Wiuf ,<sup>1</sup> Anders Albrechtsen<sup>2,\*</sup>

<sup>1</sup>Department of Mathematical Sciences, University of Copenhagen, Copenhagen 2100, Denmark

<sup>2</sup>Department of Biology, University of Copenhagen, Ole Maaløes Vej 5, Copenhagen 2200, Denmark

\*Corresponding author: Department of Biology, University of Copenhagen, Ole Maaløes Vej 5, Copenhagen 2200, Denmark. Email: aalbrechtsen@bio.ku.dk

<sup>†</sup>These authors contributed equally to this work.

Standard measures of linkage disequilibrium (LD) are affected by admixture and population structure, such that loci that are not in LD within each ancestral population appear linked when considered jointly across the populations. The influence of population structure on LD can cause problems for downstream analysis methods, in particular those that rely on LD pruning or clumping. To address this issue, we propose a measure of LD that accommodates population structure using the top inferred principal components. We estimate LD from the correlation of genotype residuals and prove that this LD measure remains unaffected by population structure when analyzing multiple populations jointly, even with admixed individuals. Based on this adjusted measure of LD, we can perform LD pruning to remove the correlation between markers for downstream analysis. Traditional LD pruning is more likely to remove markers with high differences in allele frequencies between populations, which biases measures for genetic differentiation and removes markers that are not in LD in the ancestral populations. Using data from moderately differentiated human populations and highly differentiated giraffe populations we show that traditional LD pruning biases  $F_{ST}$  and principal component analysis (PCA), which can be alleviated with the adjusted LD measure. In addition, we show that the adjusted LD leads to better PCA when pruning and that LD clumping retains more sites with the retained sites having stronger associations.

**Keywords:** linkage disequilibrium; heterogeneous populations; principal component analysis; Pearson's  $r^2$ ; SNP markers

# 1. Extra exercises on the website

[https://hds-sandbox.github.io/GWAS\\_course/develop/workshop.html](https://hds-sandbox.github.io/GWAS_course/develop/workshop.html)

## 2. LDAK (3 association testing approaches)

### 3. Bring your data

4. Download the slides

DOWNLOAD SLIDES

5. EXTRA EXERCISE - GWAS4

You can download an extra notebook and data with population structure and effect of LD on that. Upload the files through jupyterlab, and copy the data in the data folder and the notebook in the notebooks folder:

DOWNLOAD DATA

DOWNLOAD NOTEBOOK

6. EXTRA EXERCISE - LDAK

A notebook with an analysis using LDAK for preprocessing and some of the implemented association testings and heritability estimates

DOWNLOAD NOTEBOOK

7. Course evaluation survey - The Novo Nordisk Foundation funds the Sandbox project and is interested in the outcomes of our training activities, so we really appreciate your responses!

Evaluation survey link