

# Inferring population structure with dimensionality reduction and admixture analyses

Physalia course on lcWGS, Oct 2022

Nicolas Lou

# Intended learning outcomes

- By the end of this session, you will be able to
  - understand the theory underlying basic dimensionality reduction analyses (PCA and PCoA) with genomic data
  - appreciate how to extend such theory to low-coverage data
  - gain some intuition on the theory underlying admixture analysis with low-coverage data
  - recognize the strength and limitations of these analyses and the different software tools that implement them
  - implement a pipeline in ANGSD and PCAngsd to perform these analyses and interpret their results

# Population structure

- What is population structure?
- Why are we interested in population structure?
- Why are you interested in the population structure in your study system?

# Important applications of population structure analysis

- To characterize patterns of gene flow
- To discover signals of hybridization and introgression
- First step when inferring demographic history
- To delineating conservation units
- Important covariate to account for when conducting other analyses (e.g., GWAS, EAA)
- ...

# How do we approach the inference of population structure with genomic data?

- Challenges
  - High-dimensionality (hundreds of individuals, millions of loci)
  - The reality is always more complicated (after all, what is a population?)

# How do we approach the inference of population structure with genomic data?

- Possible solutions
  - Dimensionality reduction (e.g., PCA, PCoA)
    - Strengths: model free, assumption free (with exceptions); computationally simple
    - Limitations: the result is more abstract and difficult to interpret; uneven sample size skews result
  - Model-fitting (e.g., admixture analysis)
    - Strengths: the result is somewhat easier to interpret with clearer correspondence to tangible concepts and values (e.g., ancestral populations, admixture proportions)
    - Limitations: susceptible to misspecification and overfitting (e.g., IBD vs. discrete ancestral populations); computationally expensive

# Dimensionality reduction

- Common strategy
  - Start with an  $m$ -by- $n$  matrix with  $m$  genomic loci and  $n$  individuals
  - First collapse all genomic loci into an  $n$ -by- $n$  matrix with each entry representing the relationship between a pair of individuals
  - Then extract the first few axes that explain the most variation so that patterns of a variation can be visualized in a lower-dimensional space

# PCA vs. PCoA

- Principal component analysis (PCA)

- A covariance matrix is constructed

- The higher the value, the more similar two individuals are

$$C_{(w,y)} = \frac{1}{m} \sum_{s=1}^m \frac{(G_{(w,s)} - 2\hat{p}_s)(G_{(y,s)} - 2\hat{p}_s)}{\hat{p}_s(1 - \hat{p}_s)},$$

- Eigen decomposition is performed with the covariance matrix

- Principal coordinate analysis (PCoA)

- A distance matrix is constructed

- The higher the value, the more dissimilar two individuals are

$$d(i, j) = -\log\left(1 - \frac{1}{N} \sum_{s=1}^N \frac{|g(i, s) - g(j, s)|}{2}\right),$$

- Multi-dimensional scaling (MDS) is performed with the distance matrix



# PCA and PCoA with genotype data

- Let's do some math!
- Please bear with me and feel free to interrupt with questions at any time

# PCA and PCoA with unknown genotypes

- What are the challenges for performing dimensionality reduction when sequencing coverage is low?
- If you are a method developer, how would you approach this problem?

# What if the genotypes are not known?

- Strategy 1
  - Randomly sample an allele and use it to represent an individual's genotype
  - This is what **ANGSD -doIBS 1 -doCov 1** does
  - Strengths: simple, unbiased, less affected by difference in coverage
  - Limitations: doesn't account for the rest of data and possibility of sequencing error

# What if the genotypes are not known?

- Strategy 1

- Randomly sample an allele and use it to represent an individual's genotype
- This is what **ANGSD -doIBS 1 -doCov 1** does
- Strengths: simple, unbiased, less affected by difference in coverage
- Limitations: doesn't account for the rest of data and possibility of sequencing error

- Strategy 2

- Consider all possible genotypes and weigh them by their probability
- Two software tools: ngsCovar (part of ngsTools) and PCAngsd
- More math

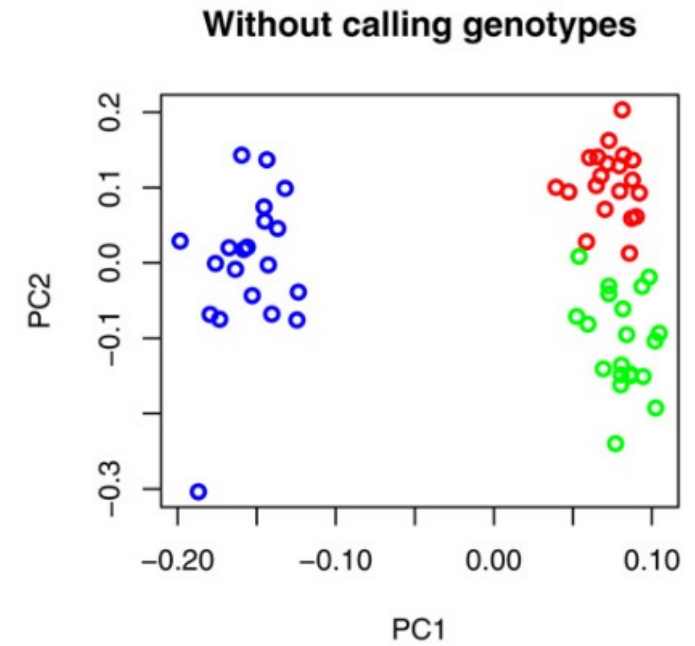
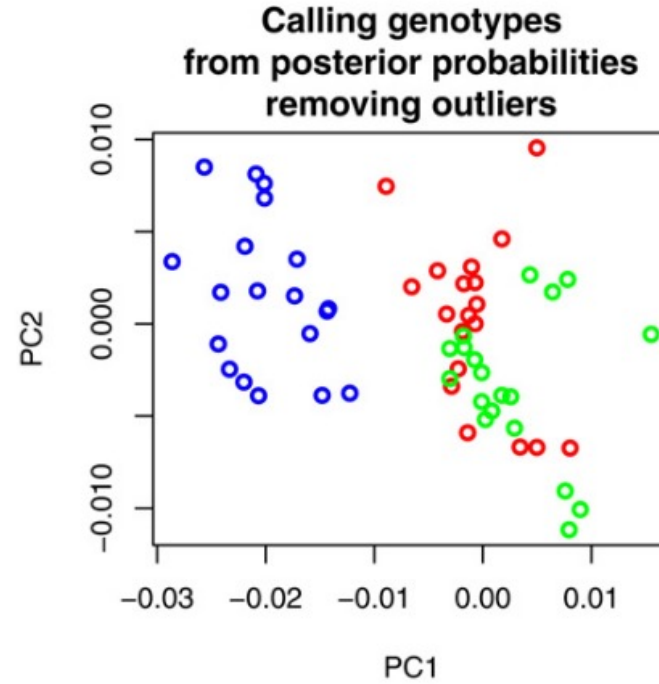
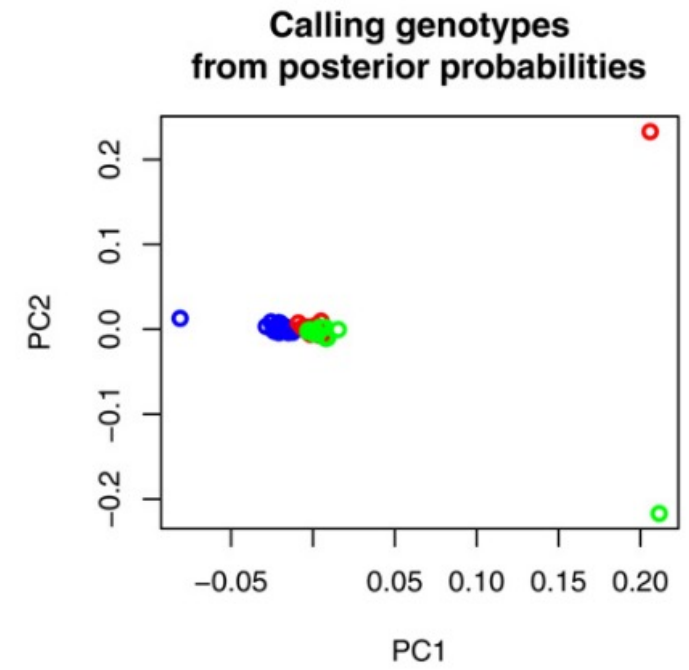
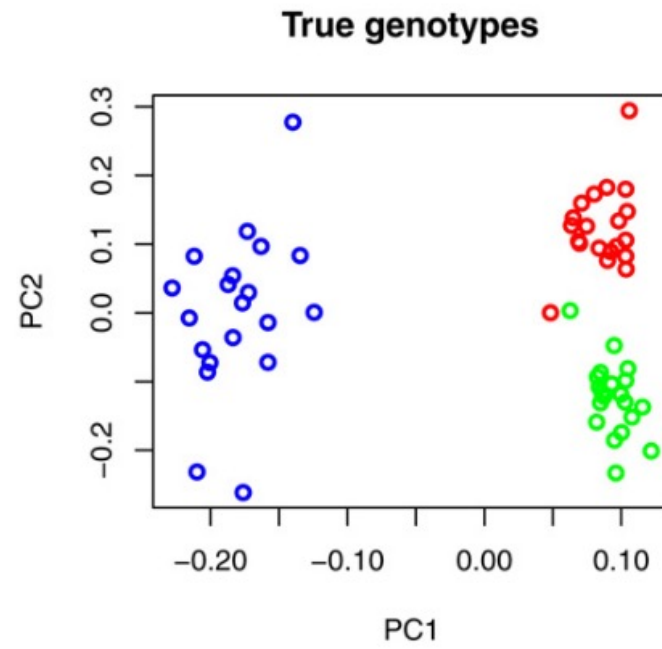
# ngsCovar vs. PCAngsd

- A circular problem
  - We would like to know the **population structure**
  - To understand population structure, we need **posterior** genotype probabilities
  - To calculate posterior probabilities, we need **prior** probabilities
  - The prior is obtained from global allele frequencies, and **Hardy-Weinberg equilibrium** is assumed when translating allele frequency to genotype probabilities
  - HWE assumption isn't true when there is **population structure**
  - But there isn't a better prior as we don't know the population structure; that's the whole point of us doing this in the first place 😞!

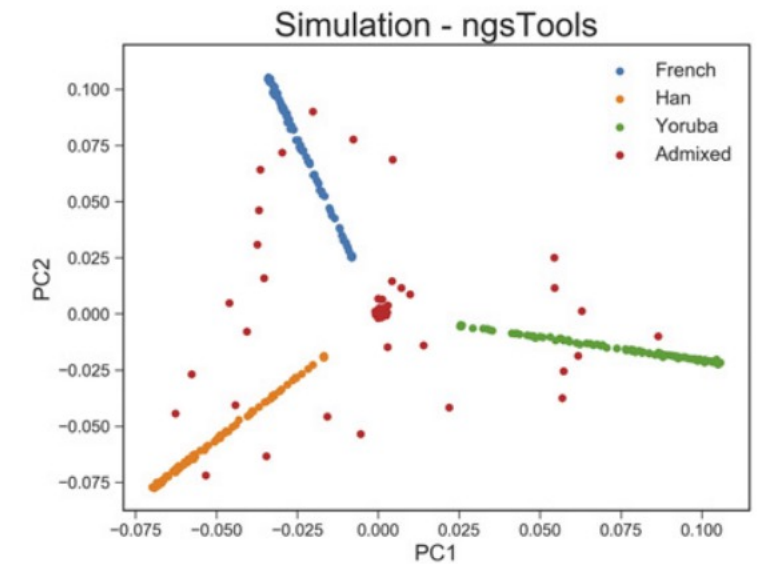
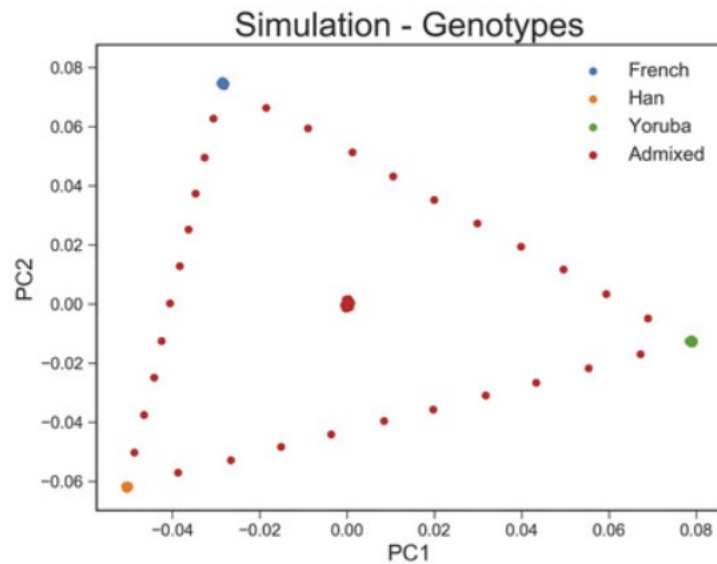
# ngsCovar vs. PCAngsd

- ngsCovar
  - ngsCovar accepts the invalid assumption as it is
  - Its result tends to still make sense in general
  - But there can be weirdness at times, especially when there is uneven coverage
  - Lower-coverage individuals are more influenced by the prior, and thus appear to be admixed even when they are not

# ngsCovar vs. hard calling genotypes



# ngsCovar vs. PCAngsd



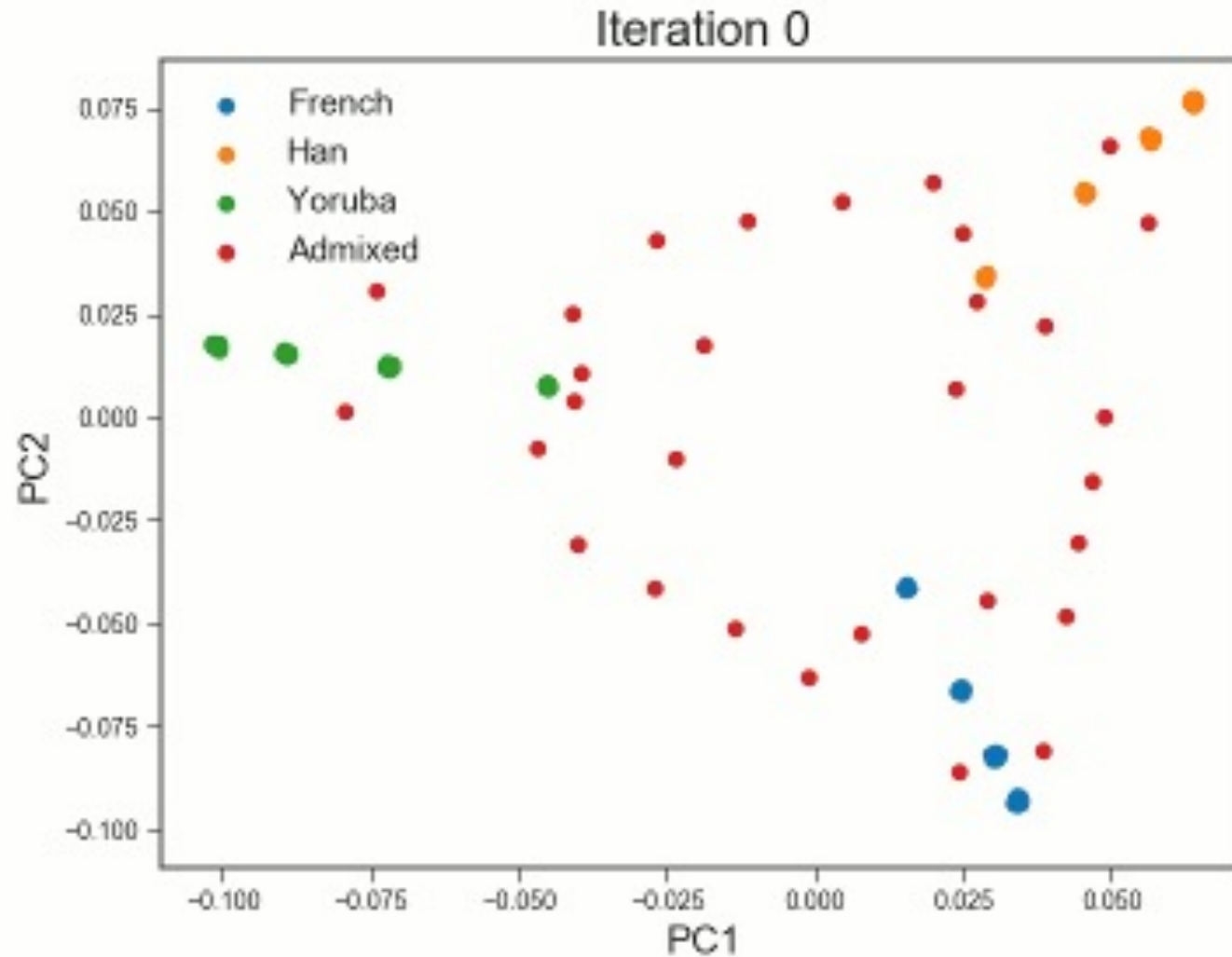
**Figure 1** PCA plots of the top two principal components in the simulated dataset consisting of 380 individuals and 0.4 million variable sites. The left-hand plot shows the PCA performed on the known genotypes using Equation 2. The middle plot shows the PCA performed by PCAngsd, and the right-hand plot displays the PCA performed by the ngsTools model (Equation 3).



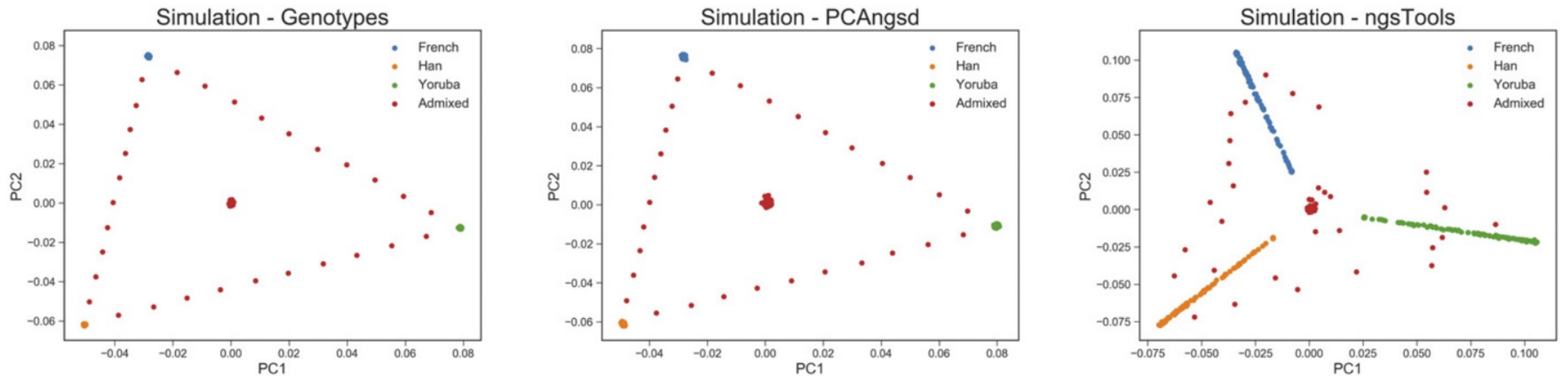
# ngsCovar vs. PCAngsd

- PCAngsd
  - PCAngsd takes this to another level by taking an iterative approach
  - The key idea is that although it could be inaccurate, the result from ngsCovar gives us important insight into the underlying population structure, and therefore a better prior
  - Its first iteration is essentially the same as ngsCovar
  - But in later iterations, it uses the result from the previous iteration to account for the population structure within the samples and gradually correct for biases in the prior genotype probabilities
  - This is the general intuition, but the algorithmic details are more complicated

# ngsCovar vs. PCAngsd

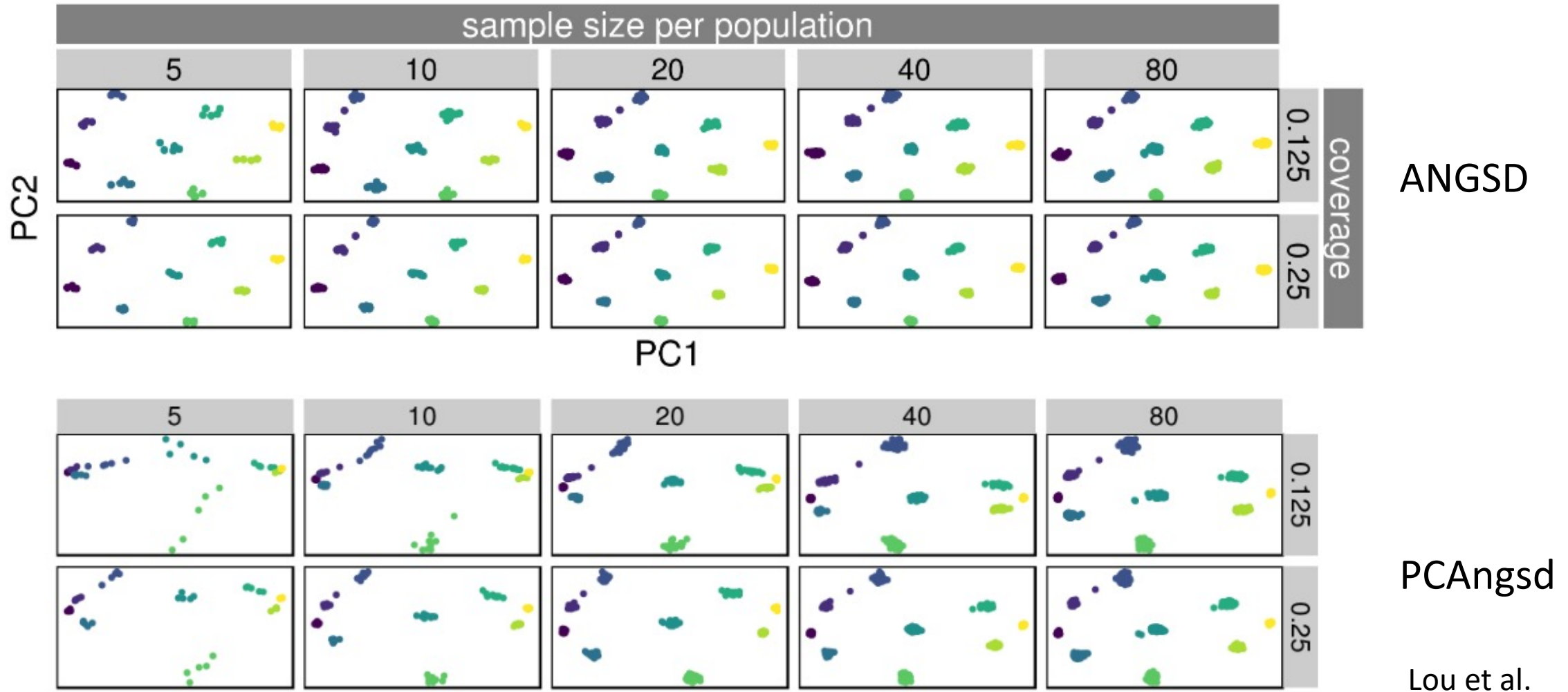


# ngsCovar vs. PCAngsd

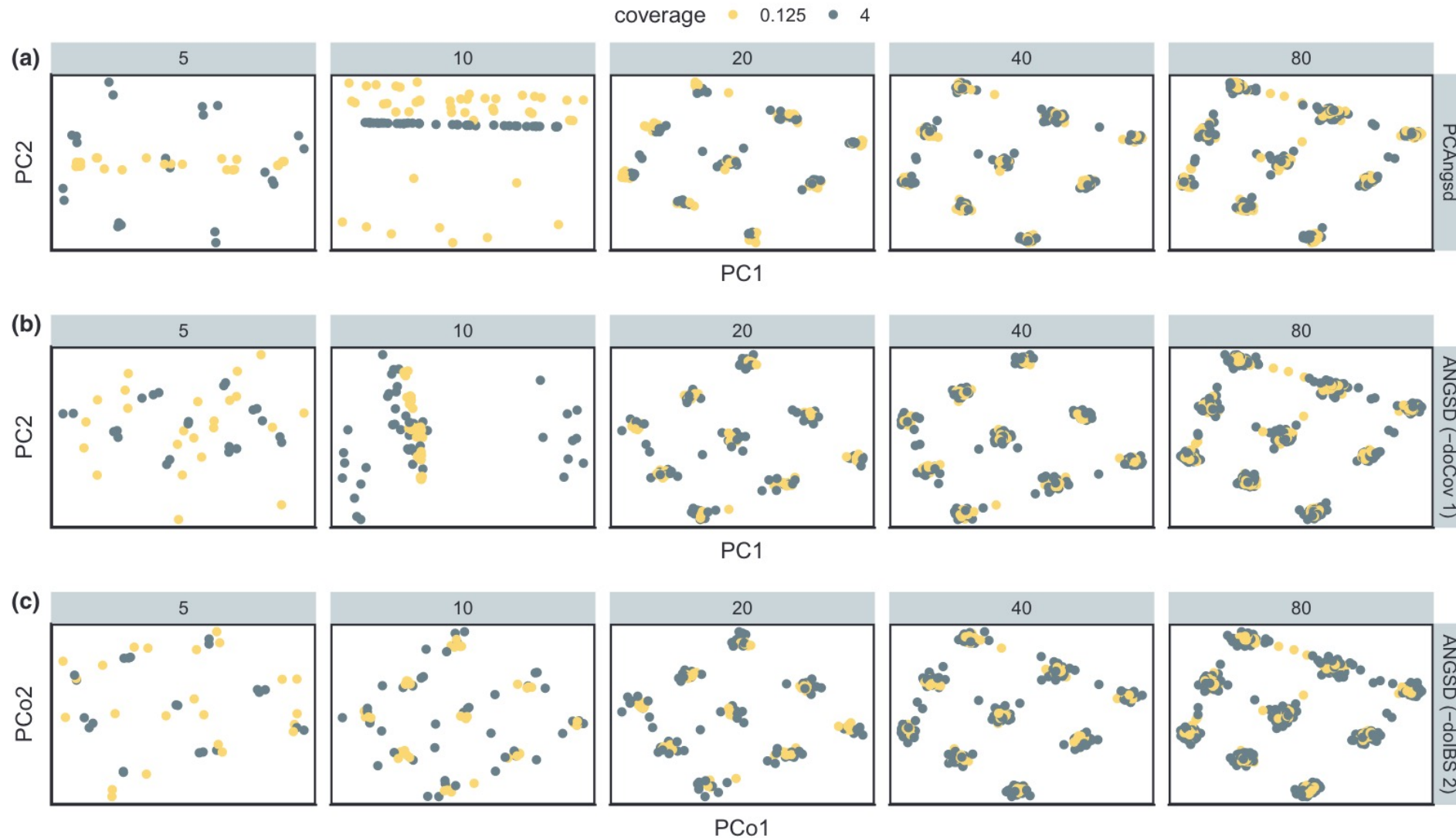


**Figure 1** PCA plots of the top two principal components in the simulated dataset consisting of 380 individuals and 0.4 million variable sites. The left-hand plot shows the PCA performed on the known genotypes using Equation 2. The middle plot shows the PCA performed by PCAngsd, and the right-hand plot displays the PCA performed by the ngsTools model (Equation 3).

# ANGSD random read sampling vs. PCAngsd when given an extreme dataset



# ANGSD random read sampling vs. PCAngsd when given another extreme dataset



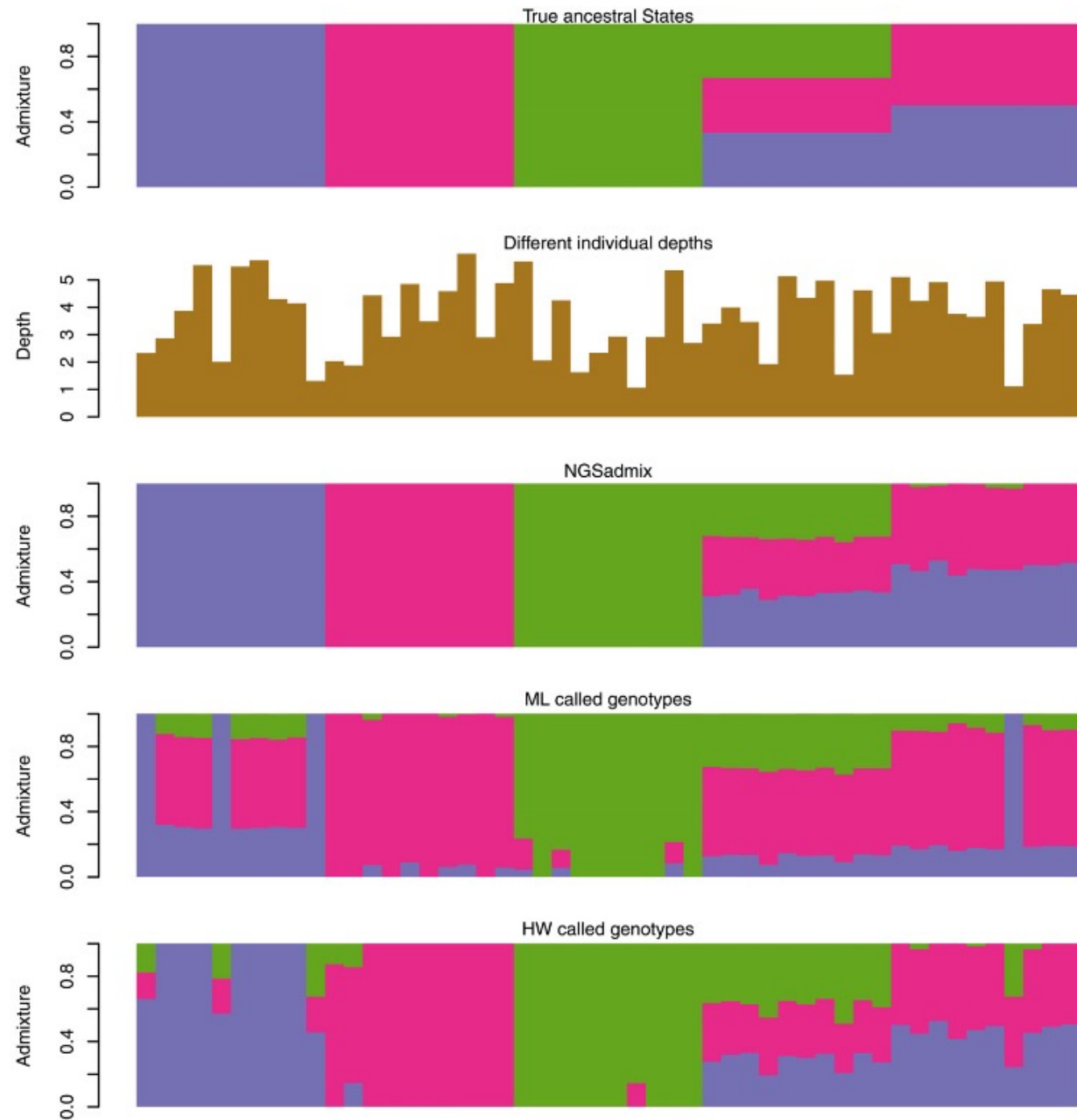
# Practical recommendations

- Use both ANGSD random read sampling and PCAngsd
- Compare their results
  - If different, look for potential artifacts
- Check correlation between PC axes with artificial differences among samples (e.g., batches, sequencing depth, etc.)

# Admixture analysis

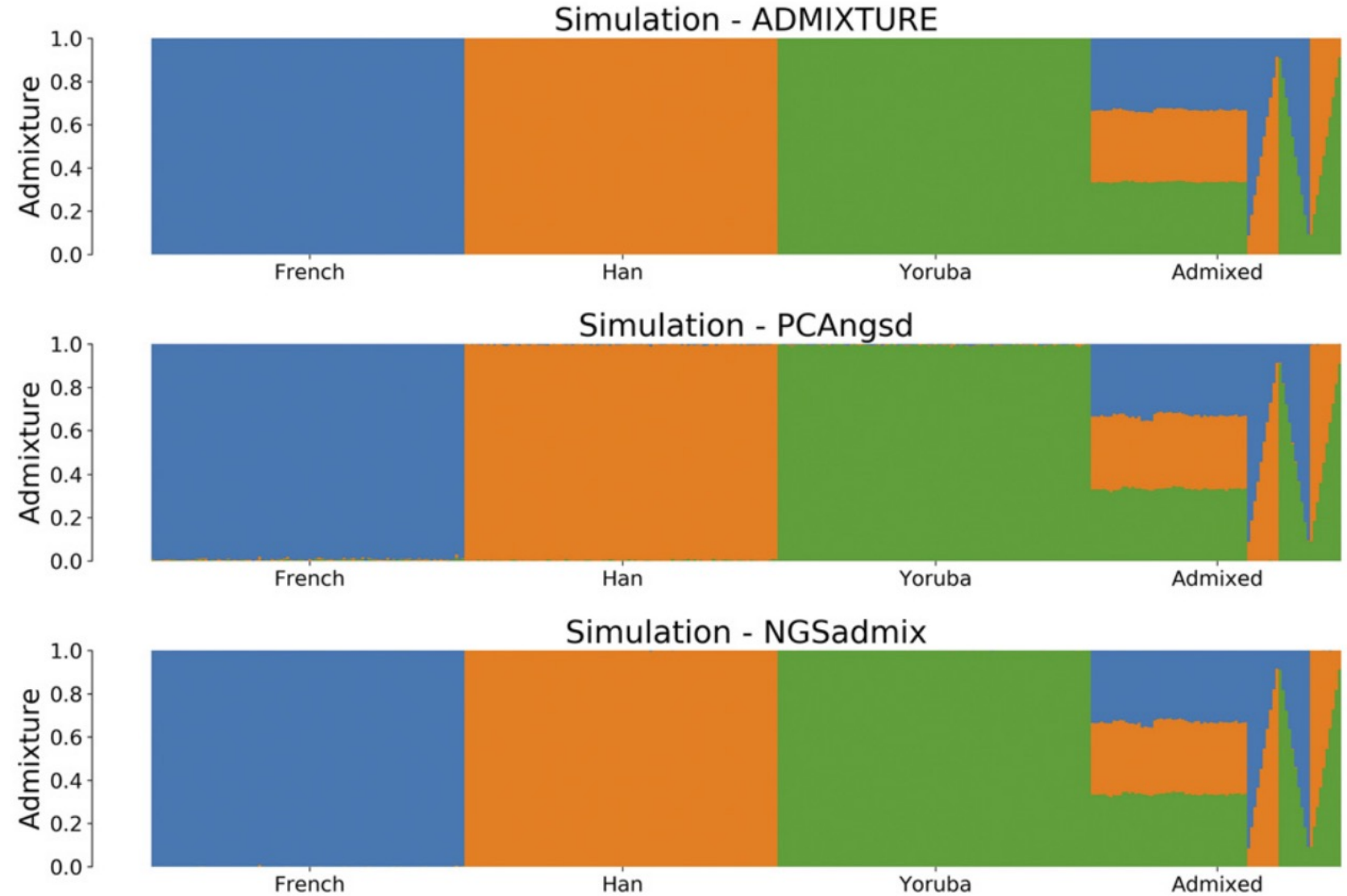
- Assume a model of  $K$  source populations each under HWE
- Find the combination of allele frequencies in source populations and individual admixture proportions that maximizes the likelihood of data
- When genotypes are known, the likelihood of data is the likelihood of genotypes
- When genotypes are unknown, integrate over all possible genotypes

# ngsAdmix vs. genotype calling





# NGSadmixmap vs. PCAngsd



# Practical recommendations

- Try both nsgAdmix and PCAngsd
- ngsAdmix is computationally intense, so allocate enough memory and time, and prepare to heavily downsample
- Although less reliant on the HWE assumption compared to ngsCovar, other model assumptions need to be kept in mind, e.g.,
  - independence of loci
  - discrete source populations
  - number of source populations
- Ohana is another software that seems to be promising, but we have not tried it yet

Questions before we go into the practical?