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
- To trace the footprints of selection and adaptation
- 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

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

  - 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

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

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

PCA and PCoA with 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

ngsCovar

Quantifying Population Genetic Differentiation from Next-Generation Sequencing Data

Matteo Fumagalli,*[,1] Filipe G. Vieira,* Thorfinn Sand Korneliussen,[†,‡] Tyler Linderoth,* Emilia Huerta-Sánchez,* Anders Albrechtsen,[‡] and Rasmus Nielsen*[,‡,§]

*Department of Integrative Biology and [§]Department of Statistics, University of California, Berkeley, California 94720, [†]Centre for GeoGenetics, Natural History Museum of Denmark, Copenhagen, Denmark 2100, and [‡]Department of Biology, University of Copenhagen, Copenhagen, Denmark 2200

Fumagalli et al. 2013

ngsCovar

- Covariance matrix when individual genotypes are known

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

- Covariance matrix when individual genotypes are unknown

$$C_{(w,y)} = \frac{1}{\sum_{s=1}^{m} P_{\mathrm{var},s}} \sum_{s=1}^{m} \frac{\left(\sum_{G_{(w,s)}=0}^{2} \sum_{G_{(y,s)}=0}^{2} \left(G_{(w,s)} - 2\hat{p}_s\right)\left(G_{(y,s)} - 2\hat{p}_s\right) P\left(G_{(w,s)} \middle| X_{(w,s)}\right) P\left(G_{(y,s)} \middle| X_{(y,s)}\right)\right) P_{\mathrm{var},s}}{\hat{p}_s(1-\hat{p}_s)},$$

Fumagalli et al. 2013

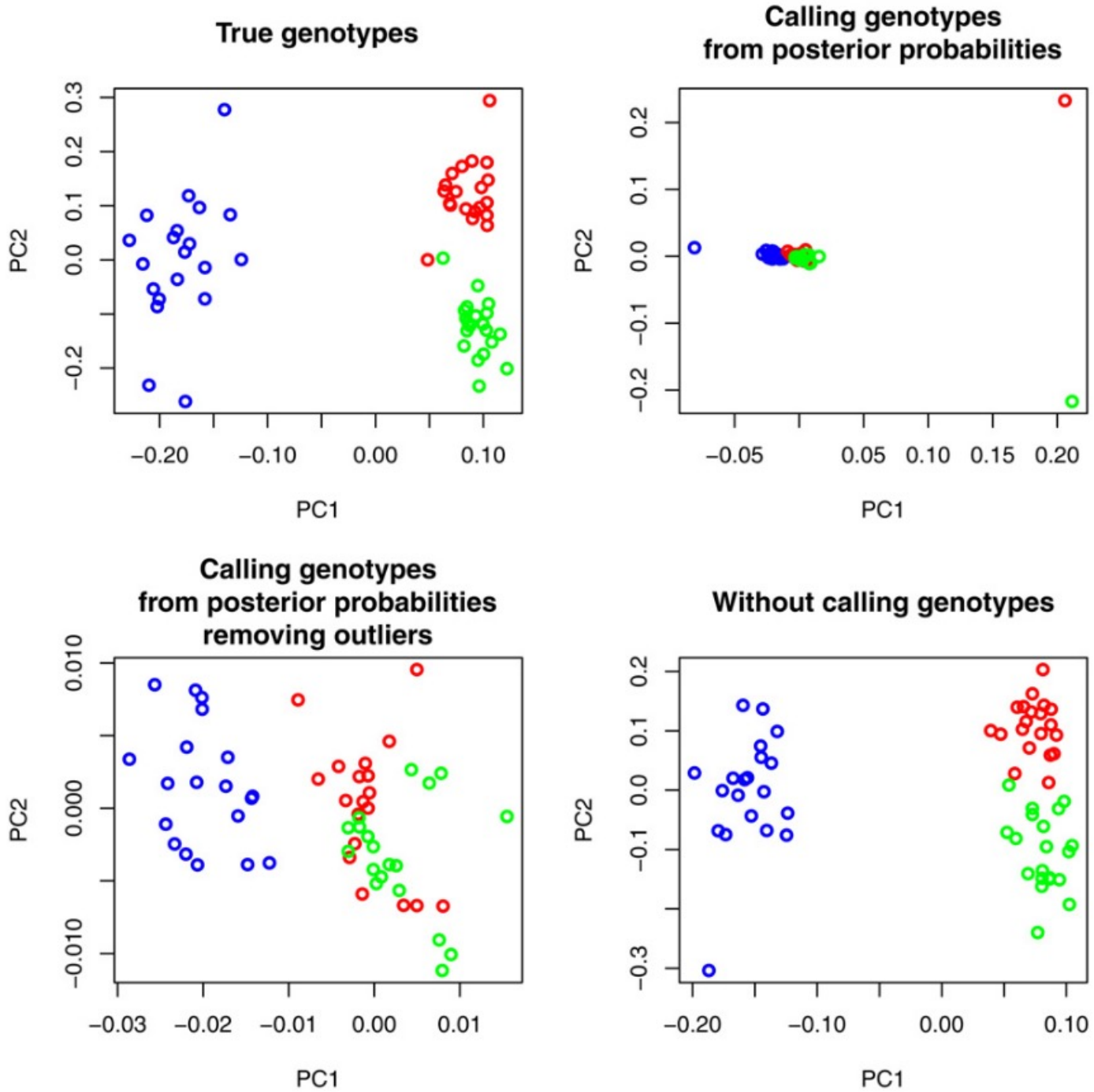
ngsCovar vs. PCAngsd

- 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

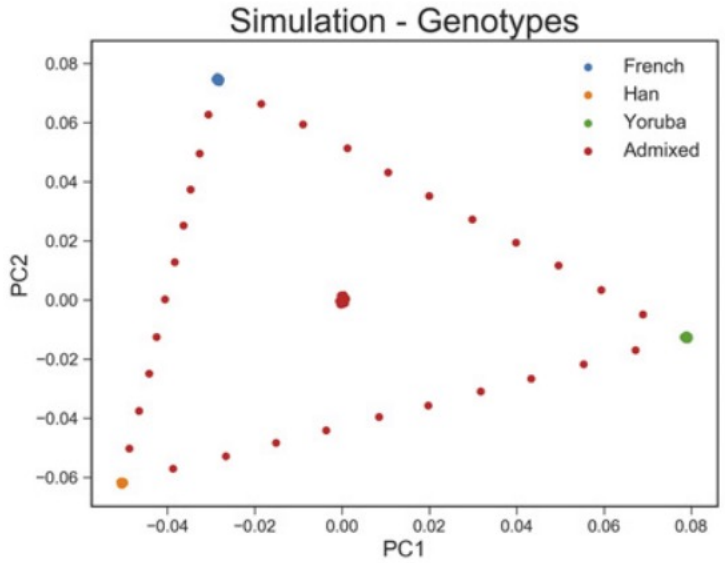
- ngsCovar
  - ngsCovar accepts the invalid assumption as it is
  - Its result tends to still make sense in general
  - But there can be weirdness at times, especially when there is uneven coverage
  - Lower-coverage individuals are more influenced by the prior, and thus appear to be admixed even when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

ngsCovar vs. hard calling genotypes



Fumagalli et al. 2013

ngsCovar vs. PCAngsd



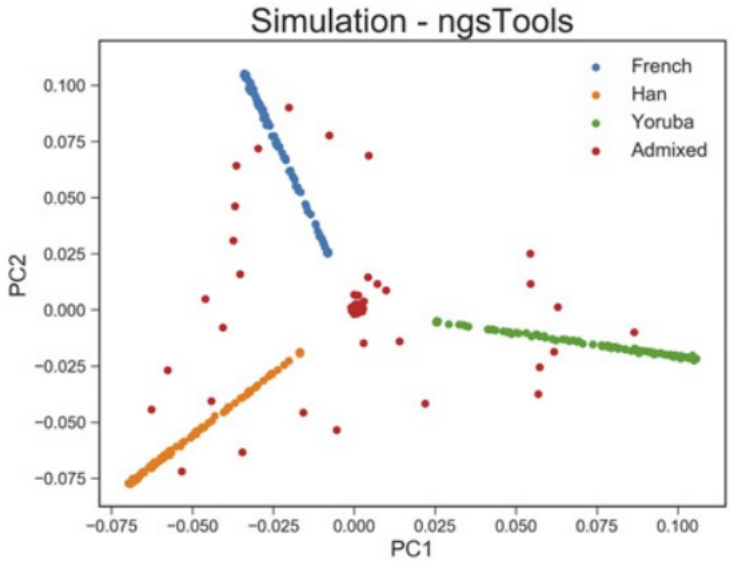


Figure 1 PCA plots of the top two principal components in the simulated dataset consisting of 380 individuals and 0.4 million variable sites. The left-hand plot shows the PCA performed on the known genotypes using Equation 2. The middle plot shows the PCA performed by PCAngsd, and the right-hand plot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

Meisner & Albrechtsen 2018

ngsCovar vs. PCAngsd

Inferring Population Structure and Admixture Proportions in Low-Depth NGS Data

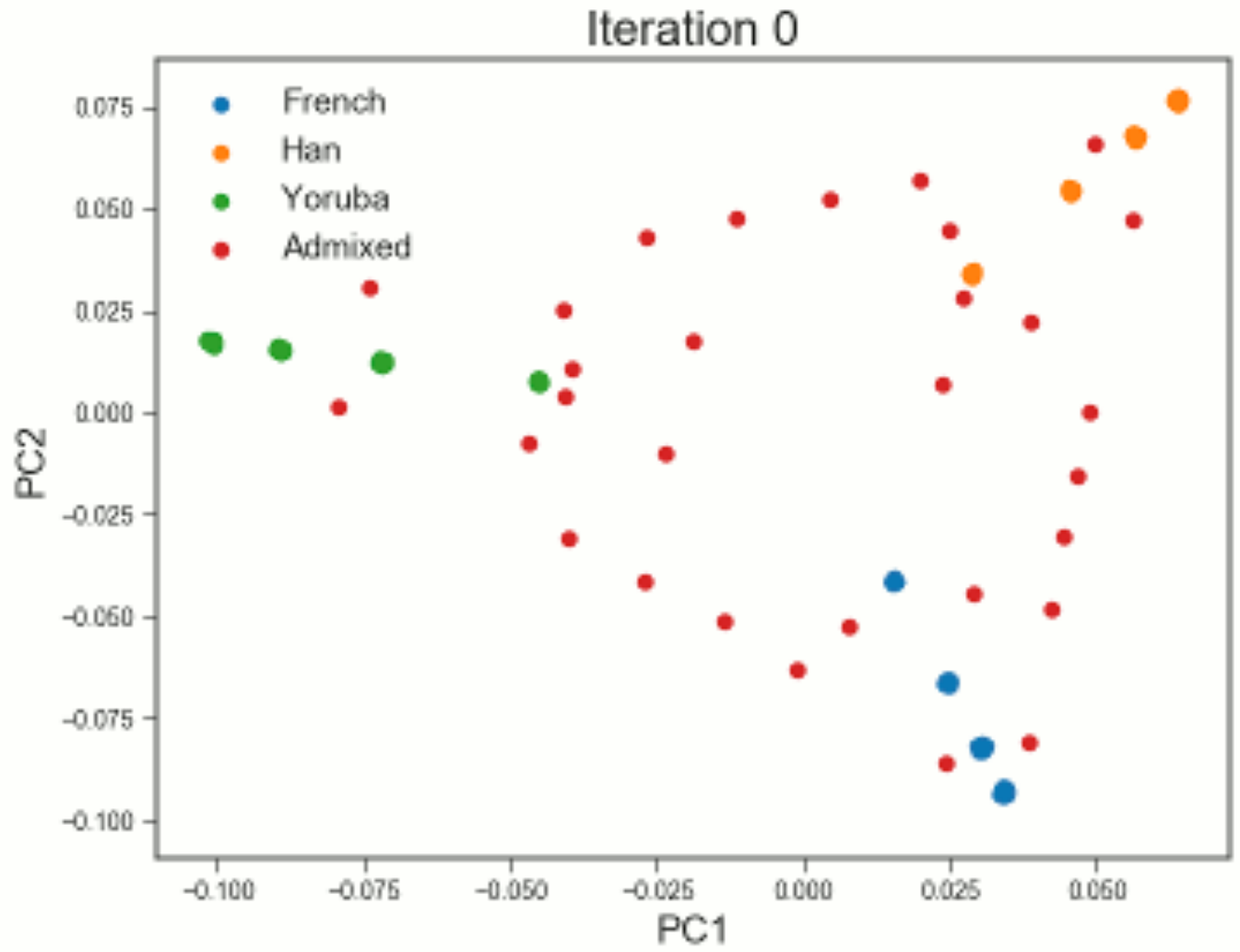
Jonas Meisner[1] and Anders Albrechtsen
The Bioinformatics Centre, Department of Biology, University of Copenhagen, DK-2200, Denmark
ORCID IDs: 0000-0002-9540-6673 (J.M.); 0000-0001-7306-031X (A.A.)

Meisner & Albrechtsen 2018

ngsCovar vs. PCAngsd

- PCAngsd
  - PCAngsd takes this to another level by employing 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



Meisner & Albrechtsen 2018

ngsCovar vs. PCAngsd

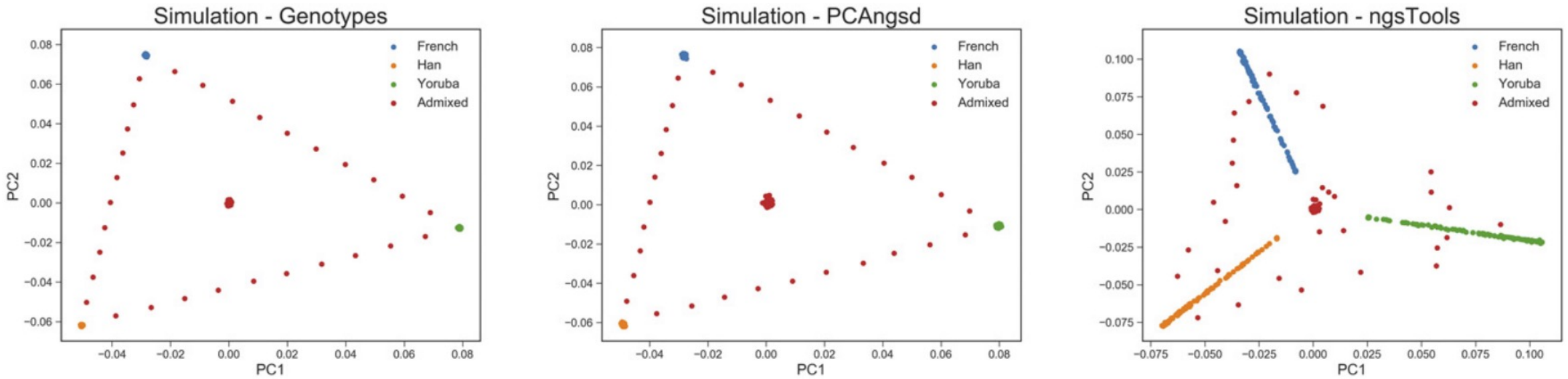
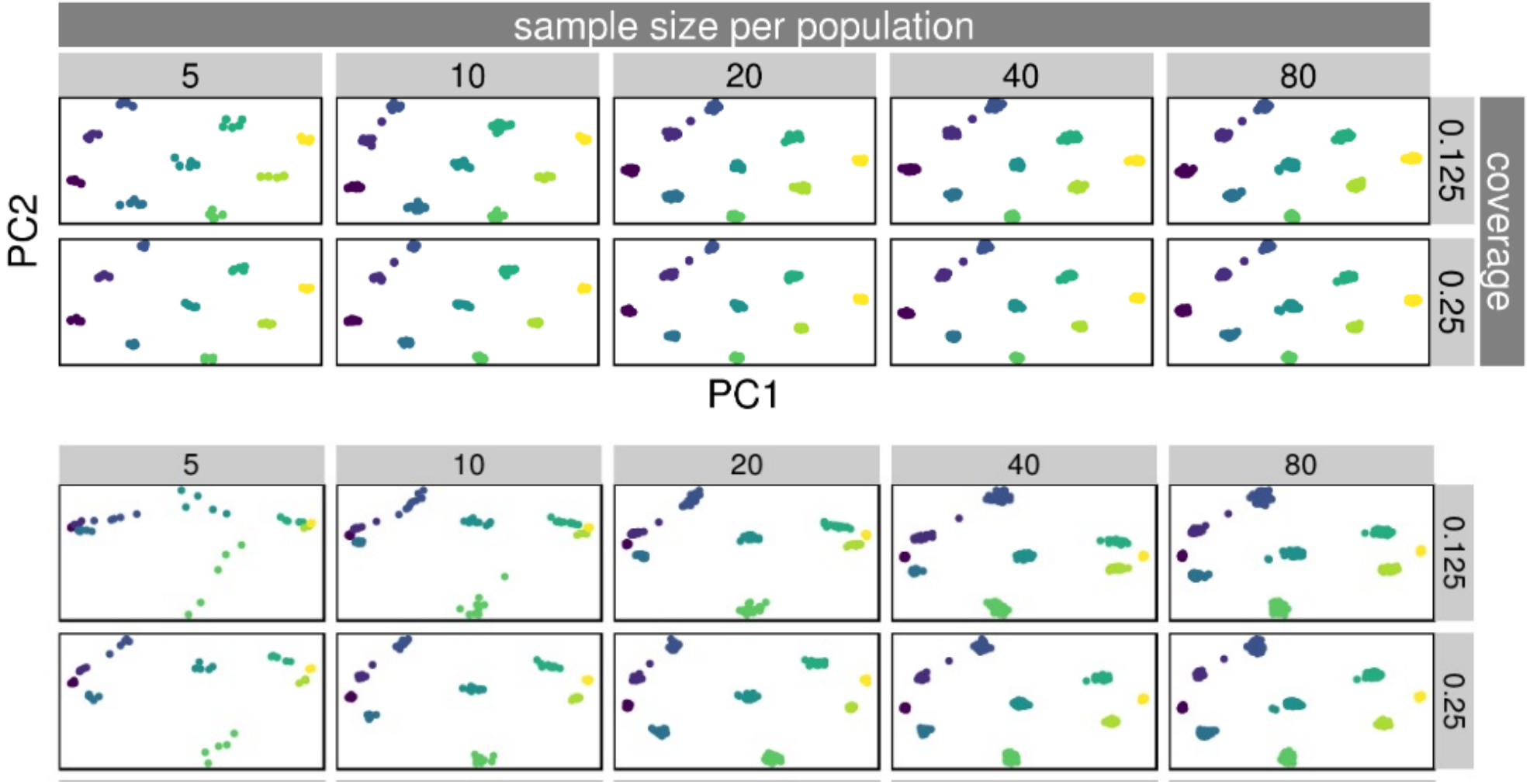


Figure 1 PCA plots of the top two principal components in the simulated dataset consisting of 380 individuals and 0.4 million variable sites. The left-hand plot shows the PCA performed on the known genotypes using Equation 2. The middle plot shows the PCA performed by PCAngsd, and the right-hand plot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

Meisner & Albrechtsen 2018

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

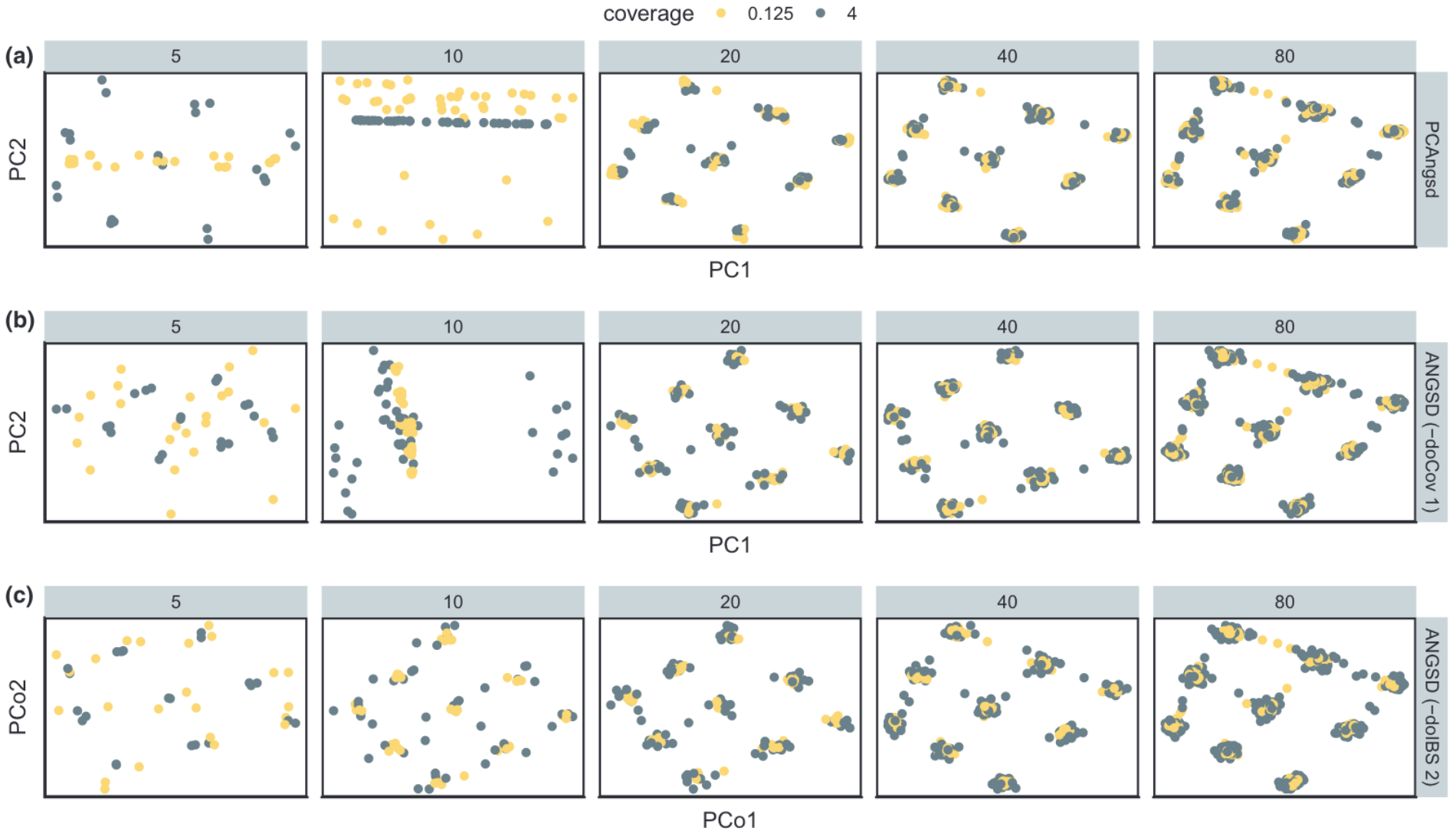


ANGSD

PCAngsd

Lou et al. 2021

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



Lou & Therkildsen 2022

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

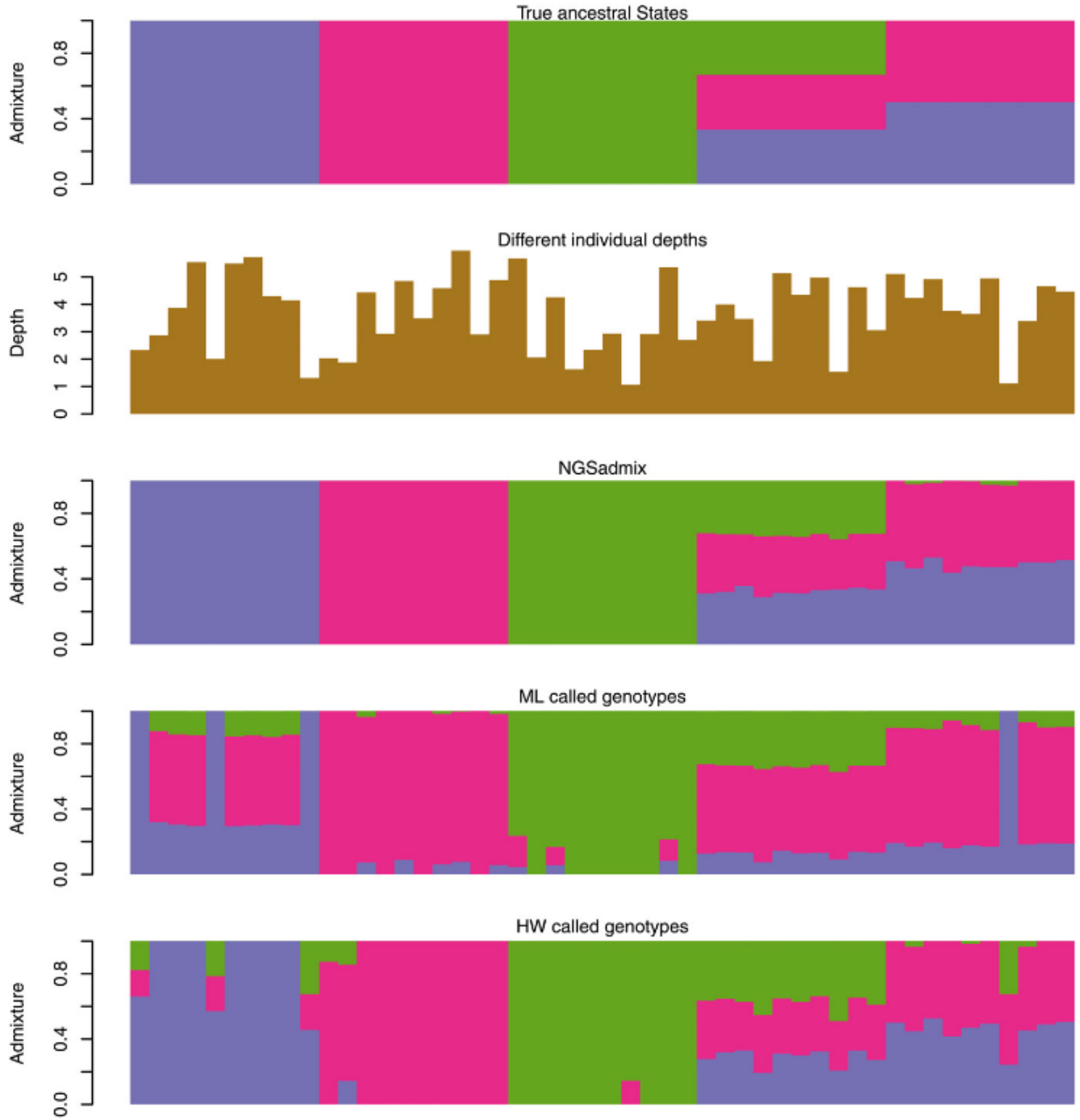
- Assume a model of K source populations each under HWE
- Find the combination of allele frequencies in source populations and individual admixture proportions that maximizes the likelihood of data
- When genotypes are known, the likelihood of data is the likelihood of genotypes
- When genotypes are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

ngsAdmix

Estimating Individual Admixture Proportions from Next Generation Sequencing Data

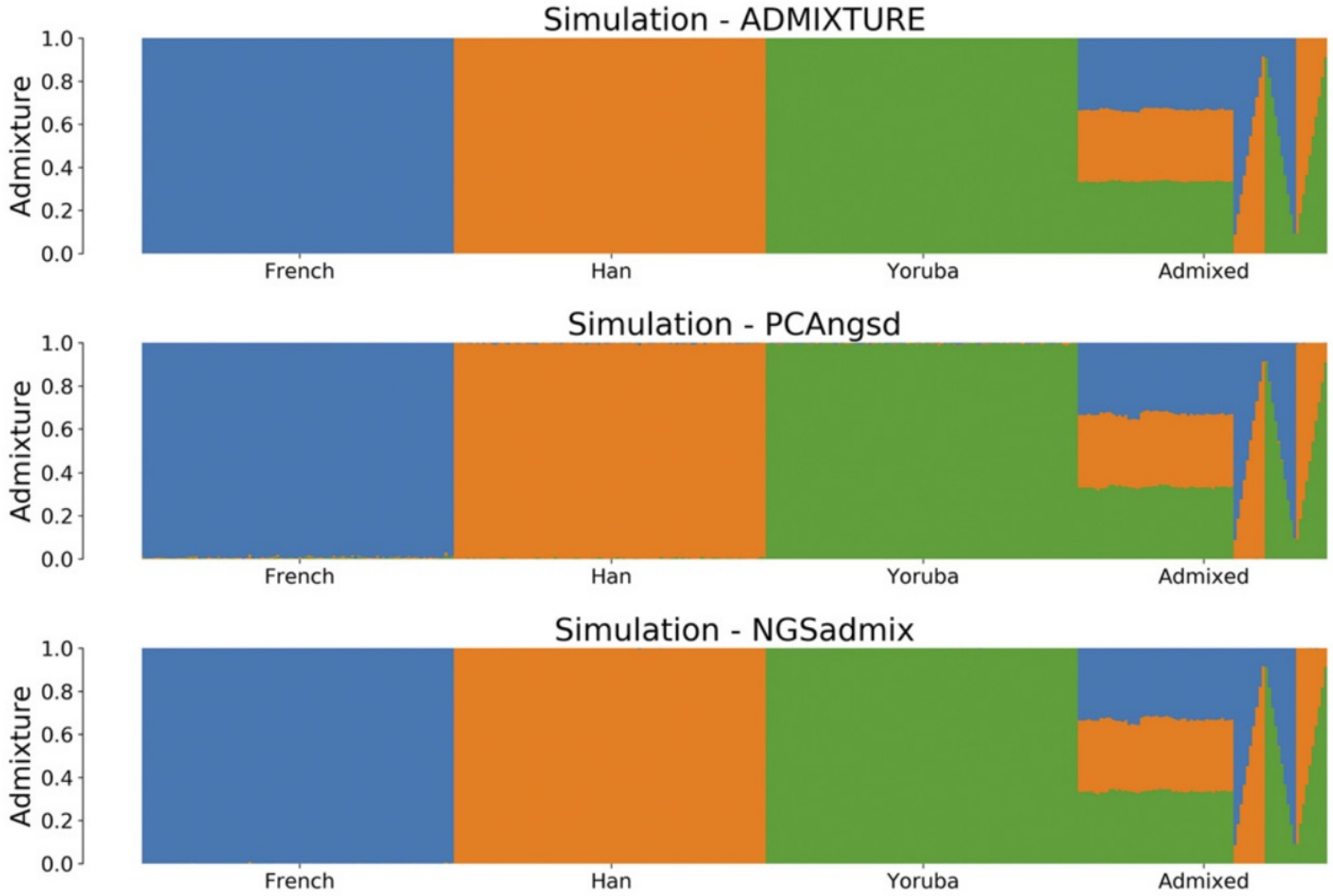
Line Skotte,*[,1,2] Thorfinn Sand Korneliussen,[†,1] and Anders Albrechtsen*
*The Bioinformatics Centre, Department of Biology, University of Copenhagen, DK-2200 Copenhagen N, and [†]Center for GeoGenetics, National History Museum of Denmark, DK-1350 Copenhagen K, Denmark

ngsAdmix vs. genotype calling



Skotte et al. 2013

NGSadmix vs. PCAngsd



Meisner & Albrechtsen 2018

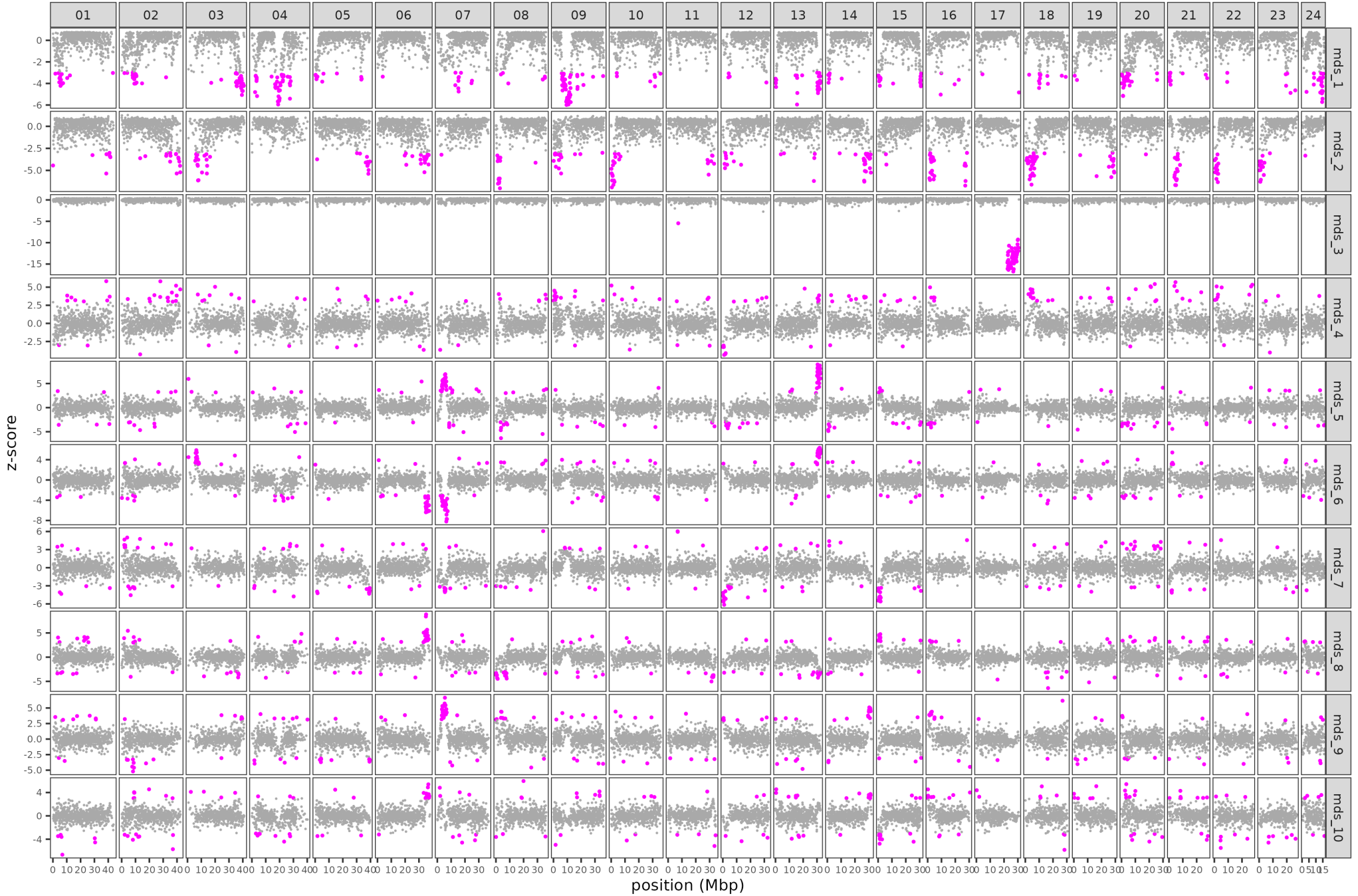
Practical recommendations

- Try both nsgAdmix and PCAngsd
- ngsAdmix is computationally intensive, so allocate enough memory and time, and prepare to heavily down-sample
- 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 third option that is promising, and it has some additional functionalities, but a systematic comparison with the other two isn’t yet available

Other applications of PCA and Admixture analysis

- DAPC
- Selection scan
  - PCAngsd (--selection)
  - Ohana (selscan)
  - Local PCA (implemented in lostruct and adapted for low-coverage data in our snakemake pipeline)

Local PCA


01
02
03
04
05
06
07
08
09
10
11
12
13
14
15
16
17
18
19
20
21
22
23
24
mds_1
mds_2
mds_3
mds_4
mds_5
mds_6
mds_7
mds_8
mds_9
mds_10
z-score
position (Mbp)

Local PCA


black-and-yellow
gopher
non.outlier (3.2%, 1.7%)
mds_1.minus (12.4%, 1.9%)
mds_2.minus (5.1%, 1.8%)
mds_3.minus (16.1%, 2.3%)
mds_4.plus (4.5%, 2.3%)
mds_4.minus (5.3%, 2.2%)
mds_5.plus (11.5%, 2.2%)
mds_5.minus (6.4%, 2%)
mds_6.plus (9.7%, 2%)
mds_6.minus (8%, 2.2%)
mds_7.plus (6.1%, 1.9%)
mds_7.minus (6.8%, 2%)
mds_8.plus (6.2%, 2.1%)
mds_8.minus (7.9%, 2.4%)
mds_9.plus (8.2%, 2.2%)
mds_9.minus (5.9%, 1.9%)
mds_10.plus (5.3%, 2%)
mds_10.minus (5.8%, 1.9%)
PC2
PC1

Questions before we go into the practical?