SLU

Basic
Biostatistics &
Bioinformatics

3 B’s Session 5

Population Structure

Lizel Potgieter
lizel.potgieter@slu.se
Department of Plant Breeding, SLU Alnarp

Who we are

SLU bioinformatics Infrastructure

Weekly online drop-in (Wednesdays at 13.00)

slubi@slu.se,

Alnarp: Lizel Potgieter (Dept. of Plant Breeding)

Statistics at SLU

SLU statistics center

Free consultations for all SLU staff

statistics@slu.se

Alnarp: Jan-Eric Englund and Adam Flöhr (Dept. of Biosystems and Technology)

Introduction

Population structure is essentially looking at differing levels of genetic relatedness among subgroups of a population

Infers the proportion of each individual's genome that came from ancestral populations = ancestry coefficients

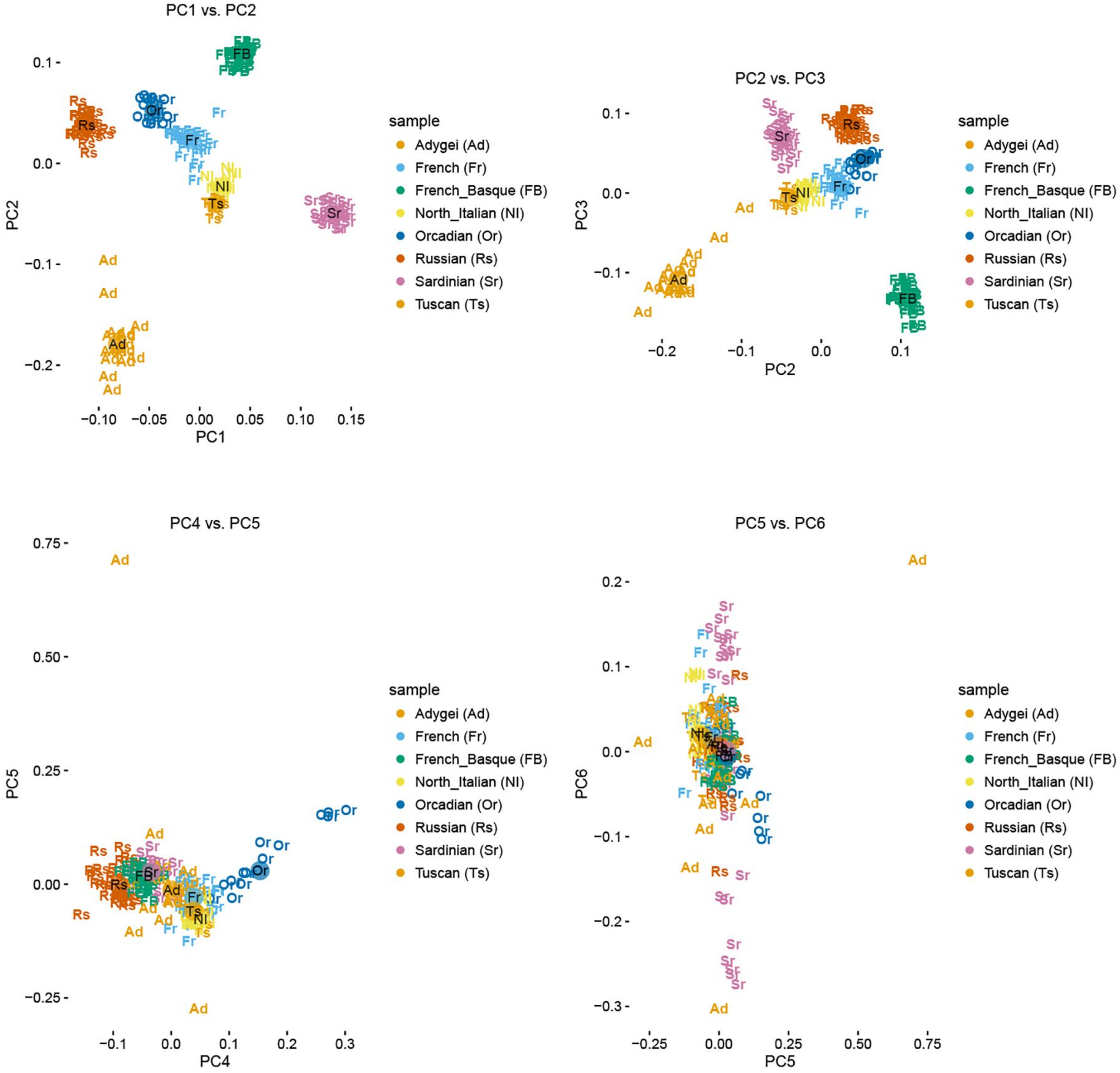
Two main tools: PCA and admixture proportion inference

Causes: physical separation followed by genetic drift, population bottlenecks or expansions, founder effects, evolutionary pressure, or simply random chance

PCA

can be confounded by demographic factors or irregular sampling designs

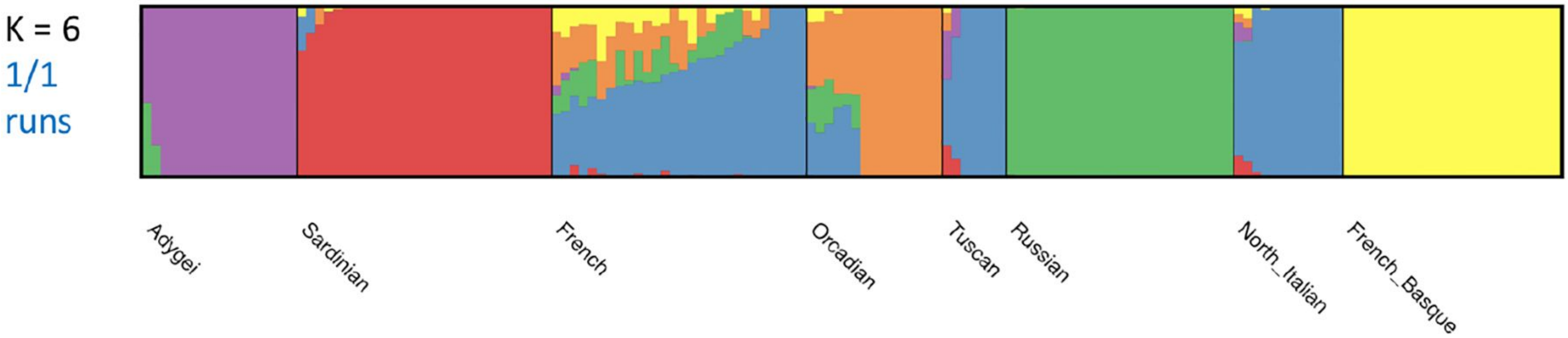
Best to be used in conjunction with other measures



See https://www.ncbi.nlm.nih.gov/pmc/articles/PMC8722024/ for more info

Admixture Proportion Inference

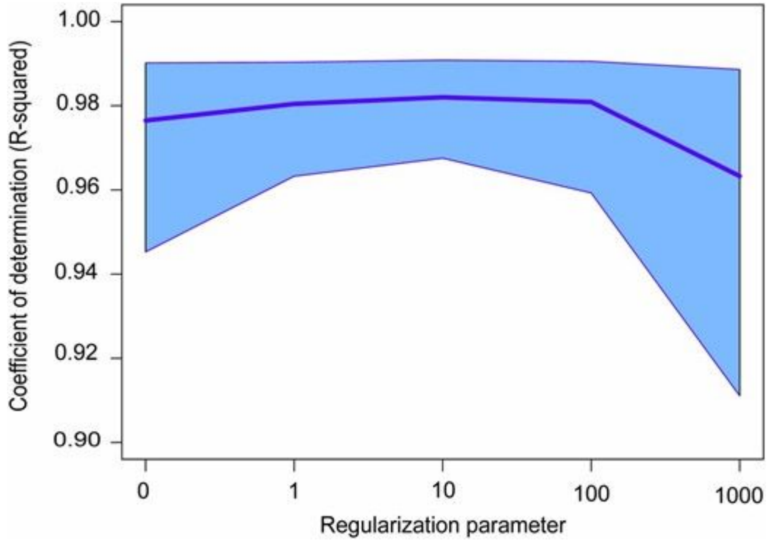
See https://www.ncbi.nlm.nih.gov/pmc/articles/PMC8722024/ for more info

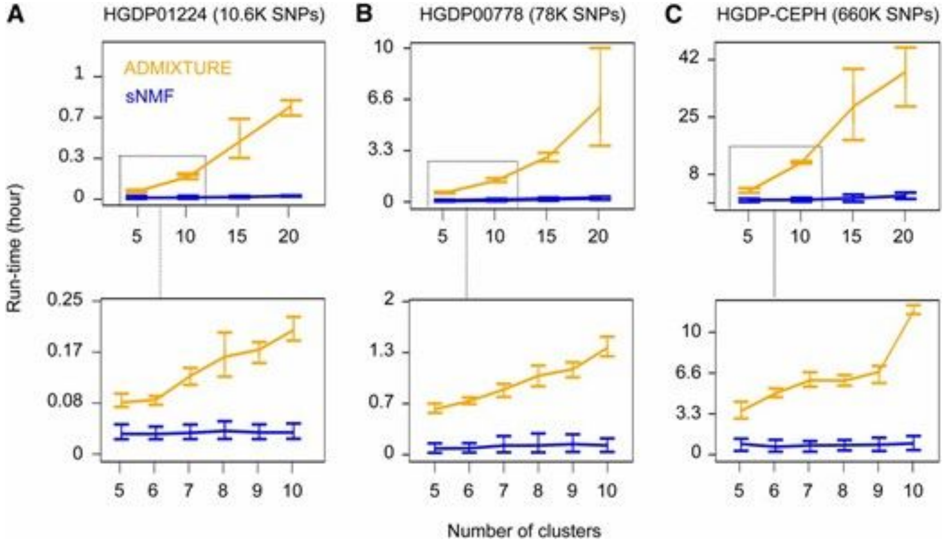


Most Common Unsupervised Tools

Unsupervised: Likelihood based methods on data

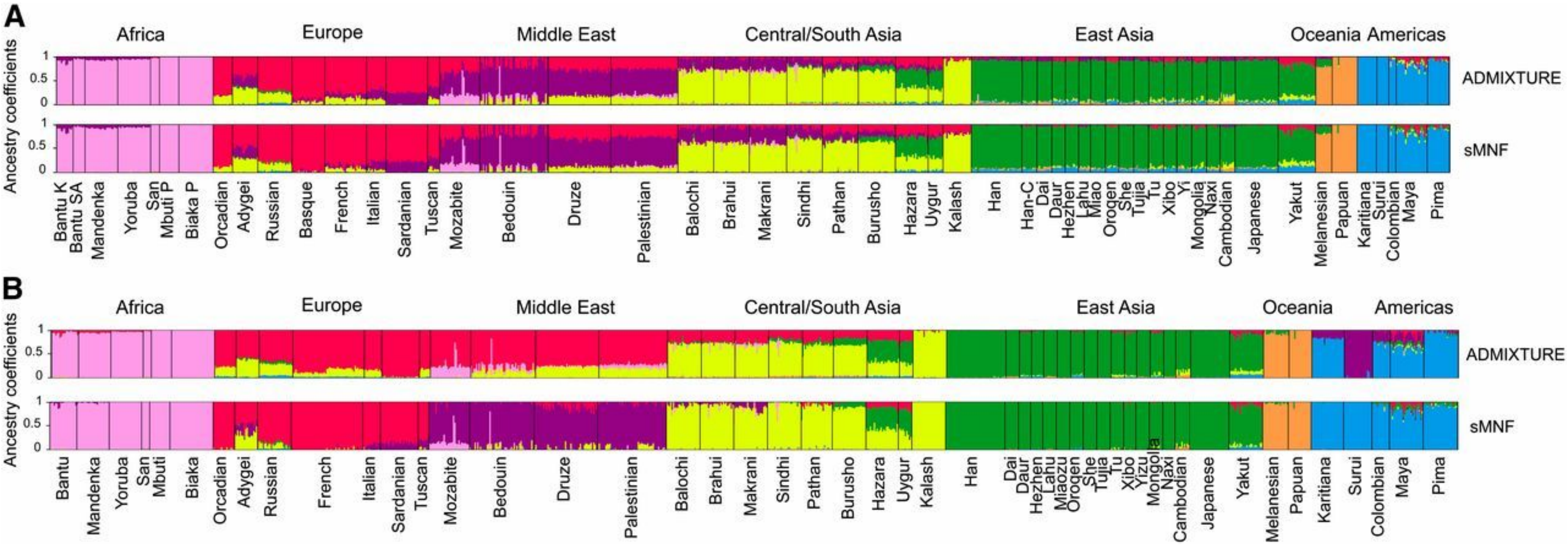
- ADMIXTURE
- STRUCTURE
- sNMF
  - In the tutorial we will use sNMF
  - it has been shown to be as accurate as the other two major programs with significantly reduced run-time (10-30x faster)
  - You can run sNMF from your R terminal with a .vcf as input





See https://doi.org/10.1534/genetics.113.160572 for sNMF benchmarking

sNMF vs ADMIXTURE



See https://doi.org/10.1534/genetics.113.160572 for sNMF benchmarking

Linkage Disequilibrium

LD pruning is crucial

Linked SNPs contain redundant information

In some cases, regions of the genome have higher LD than others and have a disproportionate influence and result in distortion

You can determine LD with Plink (plink v1.9 is older, but more stable)

See here for details on how to do it: https://www.biostars.org/p/300381/

Other Measures to Consider

Fst: Fixation index

Fst is small: allelic frequencies among populations are comparable

Fst is large: allelic frequencies among populations are large

Average number of pairwise differences between two individuals sampled from different sub-populations (between) or from the same sub-population (within)

Easy to compute with VCFtools (haploid version exists, too)

$$F_{ST} = \frac{\pi_{\text{Between}} - \pi_{\text{Within}}}{\pi_{\text{Between}}}$$

See https://www.nature.com/articles/nrg2611 for discussion on Fst and structure

SLU

SCIENCE AND EDUCATION **FOR SUSTAINABLE LIFE**