UNIVERSITY OF HOHENHEIM

Population structure inference

3502-470 Plant Genetic Resources

Prof. Karl Schmid

SS 2025

Institute of Plant Breeding, Seed Science and Population Genetics
University of Hohenheim

How can individuals or populations be grouped?

- Geographic origin
- Phenotypic similarity
- Genetic similarity

But: Geographic origin and phenotypic similarity can be frequently misleading.

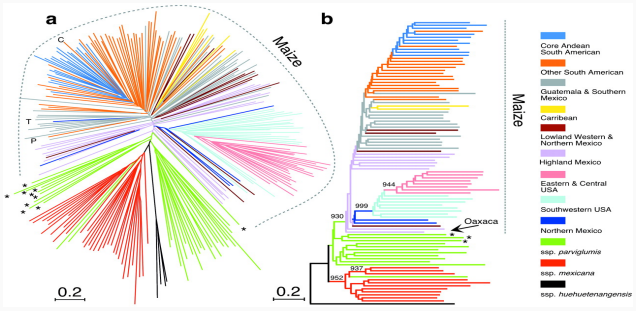
Caveats with genetic similarity

- Often a large proportion of genetic variation segregates within a (sub)population
- Only few genes may show significant structuring (i.e., adaptive trait genes)
- Many markers are needed to differentiate between genome-wide and gene-specific clustering

Methods for population structure inference

- Phylogenetic analysis
- Principal components analysis
- Model-based analysis

Phylogenetic analysis of maize varieties



Matsuoka et al., PNAS, 2002

Phylogenetic analysis

Pros

- Reflects phylogenetic relationships
- Fast calculation (for distance-based methods)
- Gives measure of genetic distance
- Straightforward interpretation

Cons

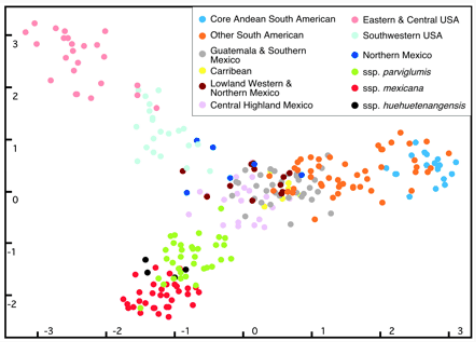
- What is the correct number of clusters?
- What if there is gene flow or reticulate evolution?

Principal components analysis (PCA)

General idea:

- Reduce the dimensions of multivariate data. Here, each marker is a dimension
- PCA is model-free and parameter-free
- Approach: Find different combinations of markers such that these combinations are uncorrelated. These combinations are called principal components
- The total variation in the data is explained by all principal components
- To reduce complexity, only retain a small number of components (those with highest variances) which explain a certain amount of the total variance or just set a number of dimensions arbitrarly to see the most important principal components
- Each data point gets projected on the chosen principal components, distinguish clusters by their values on the principal components

Principal component analysis (PCA) of maize varieties



Matsuoka et al., PNAS, 2002

Pros and Cons of PCA

Pros

- Reduction of complexity in data
- Fast calculation
- Principal components can be used for 'downstream applications'
- Identify the important processes that affect data

Cons

- Determination of distinct clusters is difficult
- No easy interpretation of evolutionary processes
- Gene flow and reticulate evolution are difficult to identify

Model-based structure inference

- Use an explicit population genetic model
- Assumptions:
  - Each subpopulation is in Hardy-Weinberg equilibrium (HWE)
  - All markers are in complete linkage equilibrium *within* populations
- Any deviation from assumption is due to the presence of population structure

Implementation:

- STRUCTURE was the most popular program (Pritchard et al., 2000), but use the faster implementation *admixture* (Alexander et al., 2009)
- In STRUCTURE, two models are possible: with and without admixture
- STRUCTURE is a Bayesian method
- A similar method is implemented in the R package 'LEA'

How does `admixture` work?

Assumptions:

- Each subpopulation is in Hardy-Weinberg equilibrium (HWE)
- All markers are in complete linkage equilibrium *within* populations

Maximum likelihood approach (with admixture)

- Assume that the sample contains genetic material inherited from K (ancestral) populations, which had allele frequencies $A = (A_{i,1} \ldots, A_{i,K})$ at the observed loci i
- For each individual, assume that a fraction of q_j its DNA can be traced back to the jth (ancestral) populations
- Compute the likelihood of the genetic data of the individuals being produced from K subpopulations with the assumed q's and A's for every possible choice of q's and A's
- Take the ancestry fractions q that produce the highest likelihood

How to choose K?

Choosing the number of populations in `admixture`

How to choose the best K? E.g. via cross-validation (see Alexander and Lange), other methods available

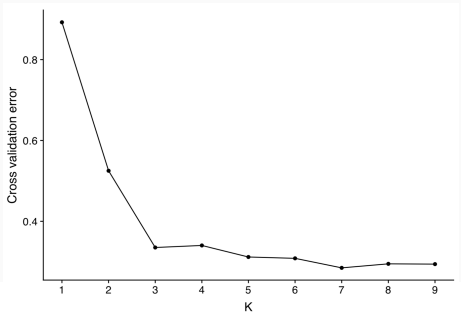
For each K, do multiple times:

- Mask a proportion of the SNP matrix (e.g. SNP position 5 in individual 3,...) at random
- Estimate the ancestries and ancestral allele frequencies as on the previous slide without the masked information
- Predict the masked entries, record your error

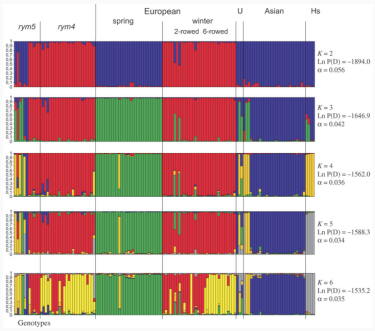
Take the K that has smallest error averaged over the multiple cross-validations

Example of cross-validation

Population structure analysis of wild and domesticated amaranths (Stetter et al., MBE (2020)



Example of a STRUCTURE analysis in cultivated barley (Stracke et al.)



Pros and cons of model-based inference

Pros

- Is based on an explicit population genetic model
- Distinct populations can be identified
- Admixture can be detected
- Straight-forward interpretation

Cons

- Running times can be long
- No (direct) distance measure between populations
- Assumptions of model are frequently violated*
- Strong influence of sampling structure, missing data (also in other methods)

*: Some violations can be circumvented, e.g. run the method on a subset of SNPs essentially without LD

Combining different methods: Back to Quingke barley

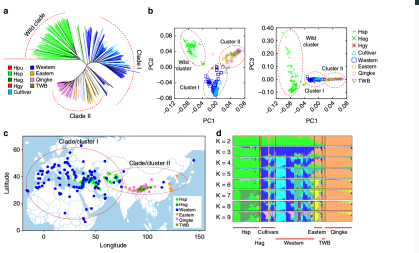


Fig. 2 Population structure of global barley accessions. Hpu *H. pubiflorum*, Hsp *H. spontaneum*, Hag *H. agriocrithon*, western: landraces in clade/cluster I, eastern landraces in clade/cluster II, TWB Tibetan weedy barley. **a** Neighbor-joining clustering based on genetic distance. **b** Principal component analyses. **c** Global distribution of barley clade/clusters revealed by (**a**, **b**). **d** Individual ancestry coefficients from $K = 2$ to $K = 9$. The red blocks below the plots

Zeng et al.

$K = 9$ chosen based on earlier results on worldwide barley diversity,

Summary

- The analysis of population structure is an important aspect of the study of genetic resources.
- A **phylogenetic approach** is usually fast and robust, however it does not provide well defined criteria for population subdivision and the assignment of individuals to populations.
- **Principle component analysis** is a fast and reliable methods, that is not based on a model of evolution.
- **Model-based approaches** such as the STRUCTURE program allow to estimate the number of populations and to assign individuals to different populations. May have long running times for large data sets.

Literature i

Cited studies using population structure inference

- Matsuoka et al. (2002) A single domestication for maize shown by multilocus microsatellite genotyping *Proc. Nat. Acad. Sc. USA* 99 (9): 6080-6084
- Zeng, Xingquan, et al. "Origin and evolution of qingke barley in Tibet." Nature Communications 9.1 (2018): 5433.
- S. Stracke, T. Presterl, N. Stein, D. Perovic, F. Ordon and A. Graner (2007) Effects of Introgression and Recombination on Haplotype Structure and Linkage Disequilibrium Surrounding a Locus Encoding Bymovirus Resistance in Barley. *Genetics* 175:805-817

Literature ii

- Jombart, Thibaut, Sébastien Devillard, and François Balloux. "Discriminant analysis of principal components: a new method for the analysis of genetically structured populations." *BMC genetics* 11.1 (2010): 94.
- Pritchard, Jonathan K., Matthew Stephens, and Peter Donnelly. "Inference of population structure using multilocus genotype data." *Genetics* 155.2 (2000): 945-959.
- D.H. Alexander, J. Novembre, and K. Lange. Fast model-based estimation of ancestry in unrelated individuals. *Genome Research* 19 (2009): 1655–1664

Literature iii

- Gauch Jr, H. G., Qian, S., Piepho, H. P., Zhou, L., & Chen, R. (2018). Effective principal components analysis of SNP data. bioRxiv, 393611.
- Li, Han, and Peter Ralph. "Local PCA shows how the effect of population structure differs along the genome." Genetics 211.1 (2019): 289-304.
- Duforet-Frebourg N, Luu K, Laval G, Bazin E, Blum MG. Detecting genomic signatures of natural selection with principal component analysis: application to the 1000 genomes data
- Alexander, David H., and Kenneth Lange. "Enhancements to the ADMIXTURE algorithm for individual ancestry estimation." BMC bioinformatics 12.1 (2011): 246.

References i

Matsuoka Y, Vigouroux Y, Goodman M, Sanchez G, Buckler E, Doebley J (2002) A single domestication for maize shown by multilocus microsatellite genotyping. Proc Natl Acad Sci USA 99(9):6080–6084, DOI 10.1073/pnas.052125199