# Admixture models

## Bayesian Methodology in Biostatistics (BST 249)

Jeffrey W. Miller

Department of Biostatistics
Harvard T.H. Chan School of Public Health

# Outline

# Outline

# Introduction: Admixtures

- Admixture models are a generalization of mixture models in which each datapoint is generated from multiple components.

- The two leading applications of admixtures are:
  1. Population structure models for genetic data, and
  2. Latent Dirichlet allocation for document/topic modeling.

- The papers introducing Bayesian admixture models for these applications are some of the most highly cited statistics papers of all time.
  - Structure: Pritchard et al. (2000) currently has over 30,445 citations according to Google Scholar.
  - LDA: Blei et al. (2003) currently has 36,355 citations according to Google Scholar.

# Outline

# Population structure: Background

- Genetic data are widely used in modern biology and medicine.
- Organisms tend to segregate into populations, such that individuals within a given population interbreed commonly, while breeding between populations is much less common.
- This causes the genetic data to be "admixed", that is, each individual has alleles coming from multiple populations.
- Failure to account for this population structure can lead to misleading results — for instance, disease association studies are often confounded by population structure.
- A common way to infer population structure from genotype data is to use an "admixture model".

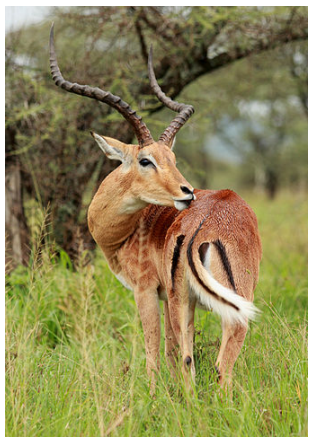
# Population structure: Basic idea of admixture models

- In a standard mixture model, each datapoint is generated by sampling from a single mixture component, chosen randomly.

- In an admixture, each datapoint consists of multiple parts, such as loci or words. For each part of a datapoint, a randomly selected mixture component is used to generate it.

- Thus, the different parts of each datapoint are generated from different mixture components in a way that varies from datapoint to datapoint.
  - Population structure: At each locus, each allele copy is drawn from a randomly selected population. The idea is that each individual has mixed heritage.
  - LDA: Each word in a document is drawn from a randomly selected topic. The idea is that each document covers a combination of topics.

# Population structure: Data



- As an example, Lorenzen et al. (2006) provide data of genotypes from $n = 216$ common impala and black-faced impala from Southern Africa.

- The common impala is widespread but the black-faced impala is an endangered subspecies.

# Population structure: Data

- For each animal $i = 1, \ldots, n$, its genotype was determined at $L = 8$ loci (i.e., 8 locations on the genome).

- At each locus $\ell = 1, \ldots, 8$, the genotype of animal $i$ consists of two allele copies, $x_{i\ell 1}, x_{i\ell 2} \in \{1, \ldots, V_\ell\}$ (since each animal has two copies of each chromosome, one from each parent).

- $x_{i\ell c}$ = variant observed at locus $\ell$, copy $c$. Missing values are denoted $-1$.

Genotype of animal $i = 86$

| $\ell$ | 1 | | 2 | | 3 | | 4 | | 5 | | 6 | | 7 | | 8 | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| $c$ | 1 | 2 | 1 | 2 | 1 | 2 | 1 | 2 | 1 | 2 | 1 | 2 | 1 | 2 | 1 | 2 |
| $x_{i\ell c}$ | 5 | 3 | 2 | 2 | –1 | –1 | 2 | 2 | 1 | 2 | 3 | 6 | 3 | 2 | 1 | 1 |

# Outline

# Population structure: Model

- (Visualize with diagram on whiteboard)
- Suppose there are $K$ populations, $L$ loci, and two allele copies at each locus.
- For individual $i$, let
  - $w_{ik}$ = proportion of genome that originates from population $k$.
  - $z_{i\ell c}$ = population of origin for locus $\ell$, copy $c$.
  - $x_{i\ell c}$ = variant observed for locus $\ell$, copy $c$.
- $\theta_{k\ell v}$ = frequency of variant $v$ at locus $\ell$ in population $k$.

# Population structure: Model

- Consider the following model:

$$\begin{aligned} Z_{i\ell c} \mid w &\sim \text{Categorical}(w_i) \\ X_{i\ell c} \mid \theta,\, Z_{i\ell c} = k &\sim \text{Categorical}(\theta_{k\ell}) \end{aligned}$$

  independently for $i \in \{1, \ldots, n\}$, $\ell \in \{1, \ldots, L\}$, $c \in \{1, 2\}$.

- Here, $w_i := (w_{i1}, \ldots, w_{iK})$ and $\theta_{k\ell} := (\theta_{k\ell 1}, \ldots, \theta_{k\ell V_\ell})$.

- Technically, the allele counts at each locus should be $\text{Multinomial}(2, \theta_{k\ell})$, since the order of the two allele copies is undetermined. However, it is fairly common to use this slightly simpler categorical model.

# Population structure: Prior

- For the prior, Pritchard et al. (2000) use

$$w_i \sim \text{Dirichlet}(\alpha, \ldots, \alpha)$$
$$\theta_{k\ell} \sim \text{Dirichlet}(\lambda, \ldots, \lambda)$$

  independently for each $i$, $k$, $\ell$.

- As a default, they set $\lambda = 1$, yielding a uniform prior on $\theta_{k\ell}$.

- As $\alpha \to 0$, this reduces to a standard mixture model since then each $w_i$ gives probability 1 to a single population $k$.

- As $\alpha \to \infty$, this forces $w_i = (1/K, \ldots, 1/K)$, making each individual equally admixed across all populations.

- To infer an appropriate $\alpha$, they put a hyperprior on it:

$$\alpha \sim \text{Uniform}(0, 10).$$

# Group activity: Quick check

Go to breakout rooms and work together to answer these questions:
https://forms.gle/9pKTd3Q64y95xnVr6

(Three people per room, randomly assigned. 5 minutes.)

# Population structure: MCMC

- Pritchard et al. (2000) use the following Gibbs scheme.
- Randomly initialize the $z$'s and $\alpha$, and iteratively:
  1. Update $w, \theta$ by sampling from the full conditional $w, \theta | x, z, \alpha$.
  2. Update $z$ by sampling from the full conditional $z | x, w, \theta, \alpha$.
  3. Update $\alpha$ using a Metropolis–Hastings step targeting its full conditional $\alpha | x, z, w, \theta$.

# Population structure: Issues that we will revisit

- There are some subtle issues that we will come back to:
  - How to choose $K$?
  - How to deal with label switching?
  - How to avoid overinterpreting the results?
- But first, let's look a couple applications.

# Outline

# Population structure: Application to thrush data

- Data was collected from the Taita thrush, *Turdus helleri*, an endangered bird species (Pritchard et al., 2000).

- $n = 155$ birds sampled at four locations in southeast Kenya:
  - Chawia (17), Ngangao (54), Mbololo (80), and Yale (4).

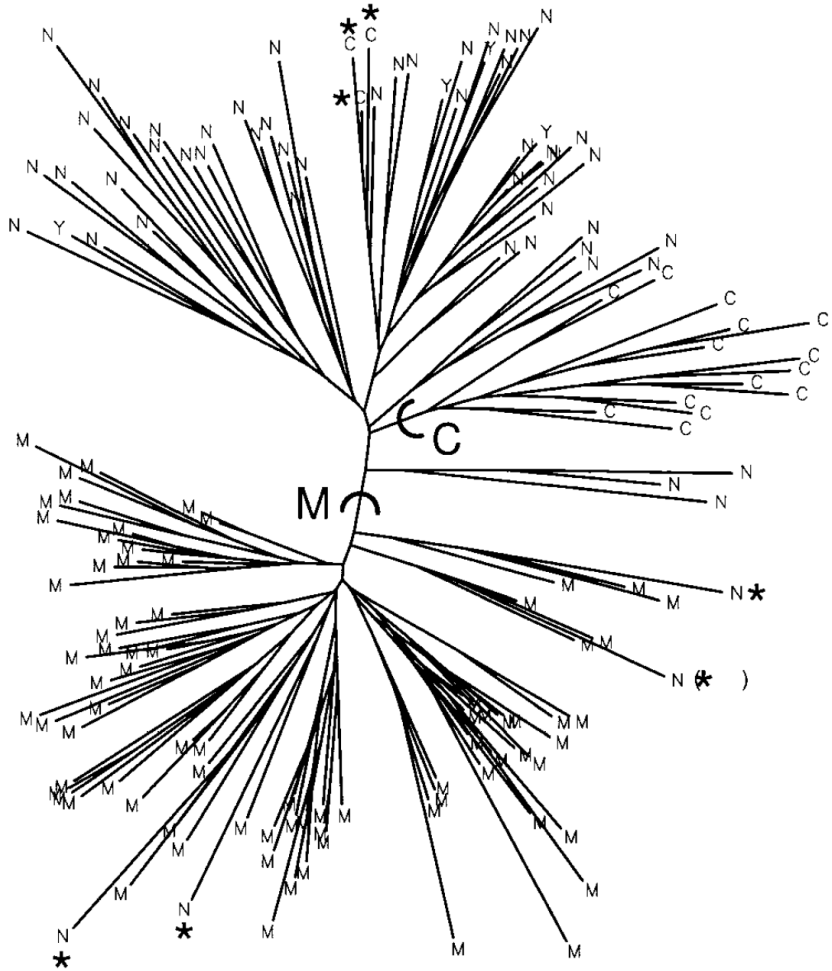
- Each individual was genotyped at seven microsatellite loci.

- Objectives of analysis:
  1. Infer whether migration/interbreeding across locations has occurred.

  2. If the location labels were unknown, we could try to recover the labels.

# Thrush example: Hierarchical clustering method fails

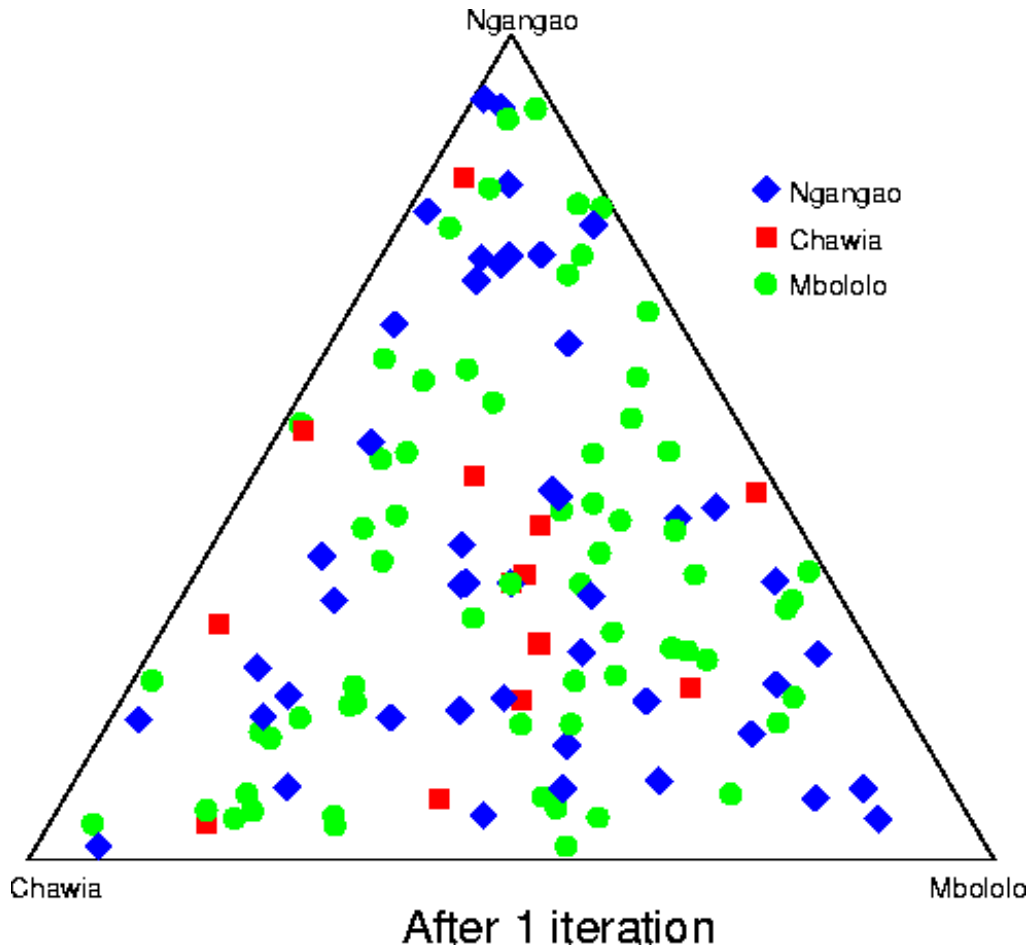
Neighor-joining tree does not cleanly separate populations (Pritchard et al., 2000)




- Before looking at the results of the admixture model, here is a competing method. Neighbor-joining trees are a commonly used clustering method for this type of data.

# Thrush example: Visualizing MCMC for admixture model

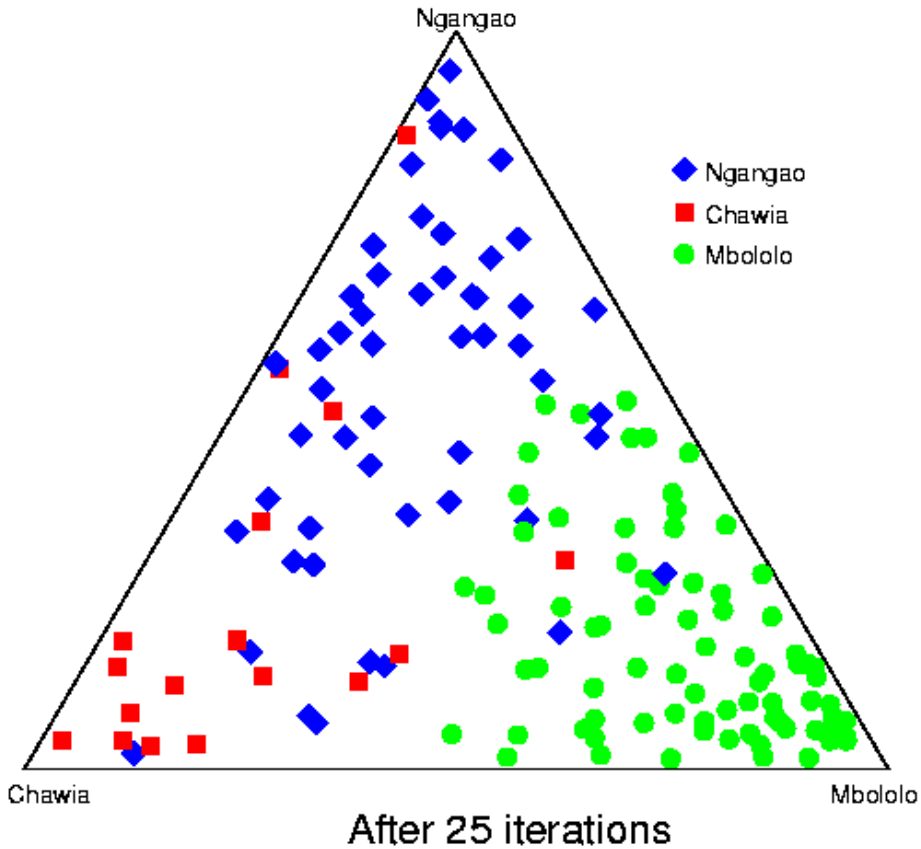
Random initialization of $w_i$ for each individual $i$
(estimated proportion of genome from each population)




(figures from `https://web.stanford.edu/group/pritchardlab/software/timelapse.html`)

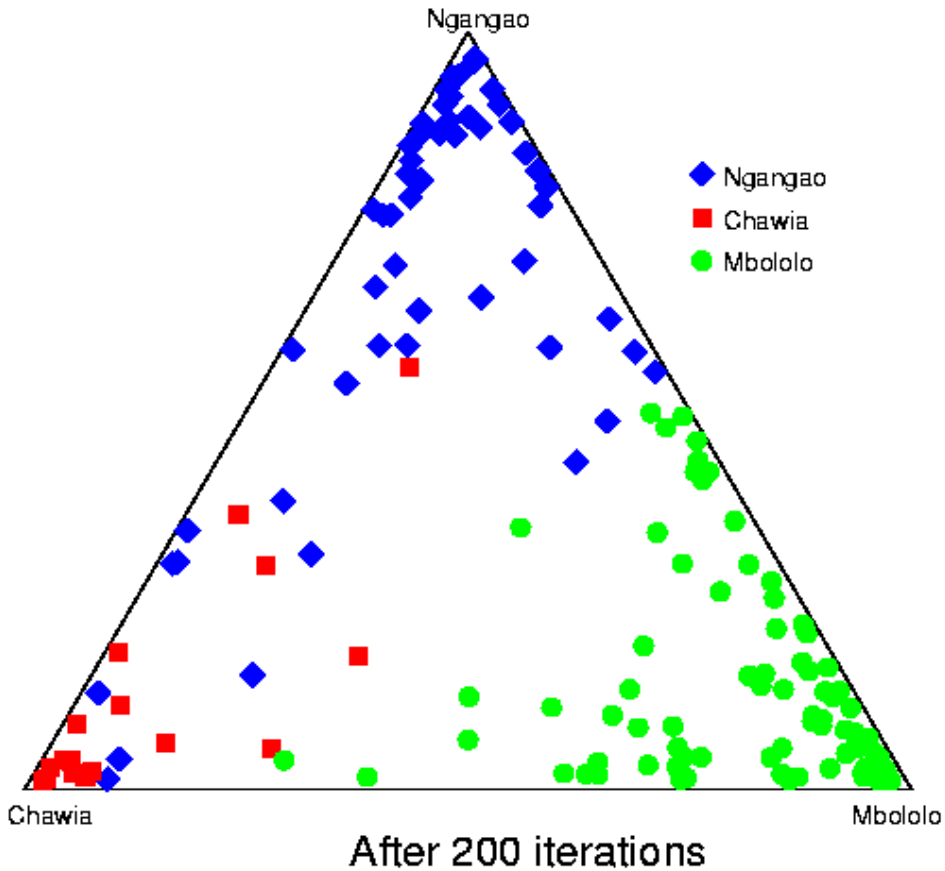
# Thrush example: Visualizing MCMC for admixture model

MCMC sample after 25 iterations: $w_i$ for each individual $i$
(estimated proportion of genome from each population)



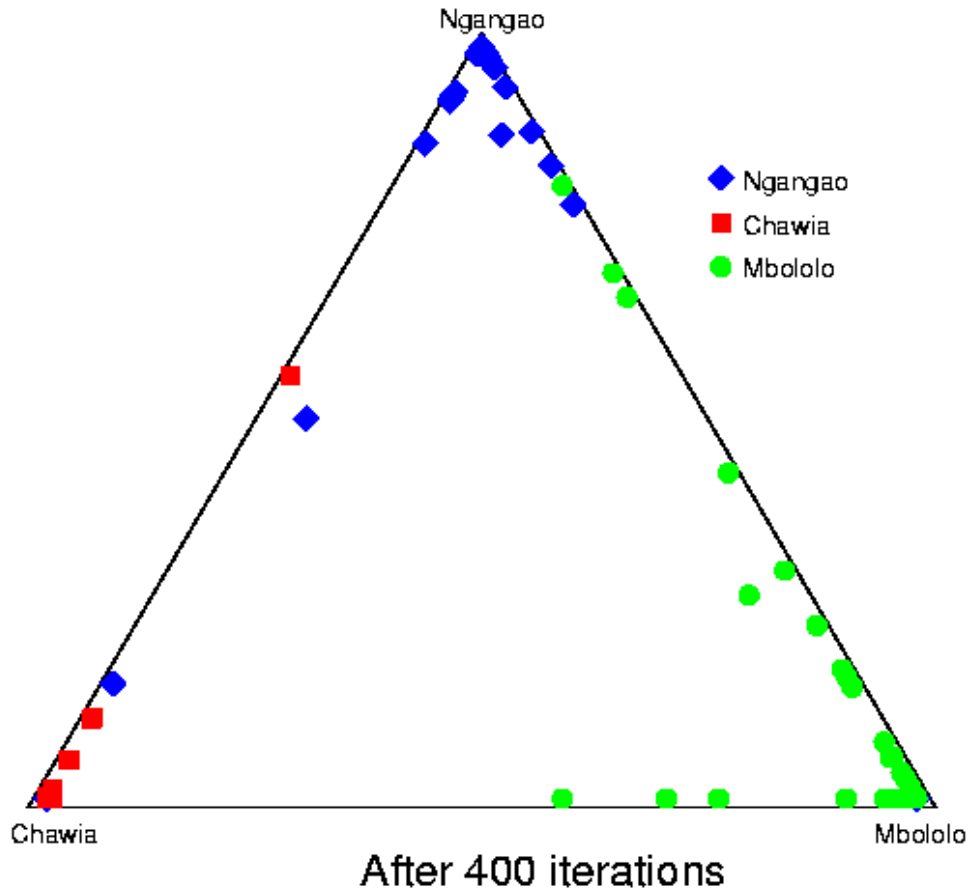
# Thrush example: Visualizing MCMC for admixture model

MCMC sample after 200 iterations: $w_i$ for each individual $i$
(estimated proportion of genome from each population)



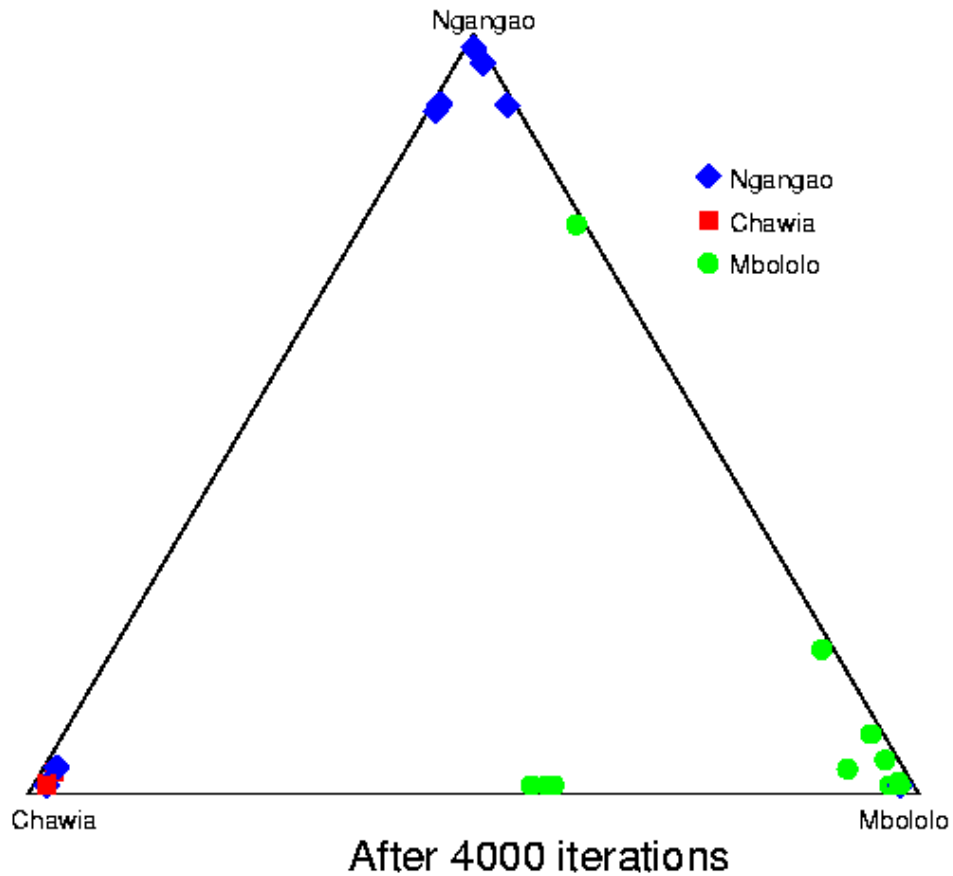
# Thrush example: Visualizing MCMC for admixture model

MCMC sample after 400 iterations: $w_i$ for each individual $i$
(estimated proportion of genome from each population)



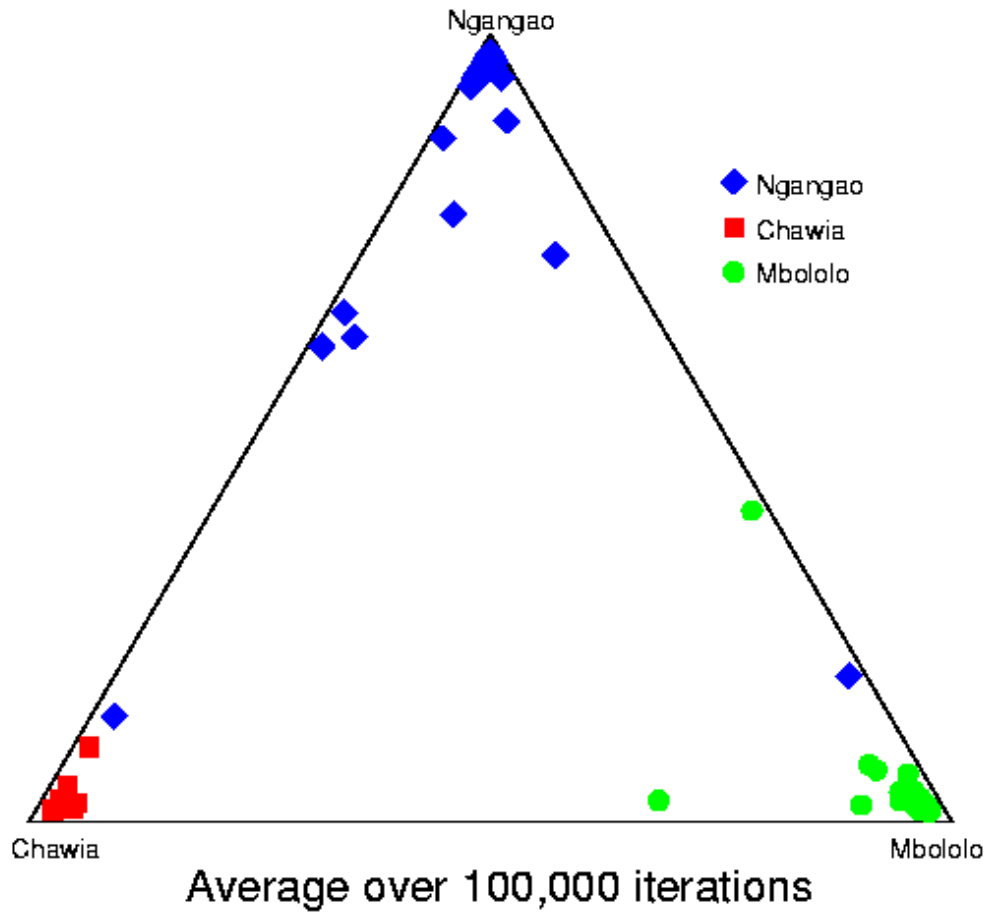
# Thrush example: Visualizing MCMC for admixture model

MCMC sample after 4000 iterations: $w_i$ for each individual $i$
(estimated proportion of genome from each population)



# Thrush example: Visualizing MCMC for admixture model

Averaging over 100,000 MCMC iterations: $w_i$ for each individual $i$
(estimated proportion of genome from each population)



# Thrush example: Choosing number of populations $K$

- As in finite mixtures, choosing $K$ in admixtures is tricky.
- Pritchard et al. (2000) propose a rough approximation to the marginal likelihood $p(x|K)$. (See homework 4.)
- It is intended only to help choose $K$, not to provide an accurate approximation to $p(x|K)$.
- The high value of $p(K = 3 \mid x)$ should not be taken too seriously, even if $p(x|K)$ were exact. (Critical thinking: Why?)

**Inferring the value of *K*, the number of populations, for the *T. helleri* data**

| $K$ | log $P(X\|K)$ | $P(K\|X)$ |
|---|---|---|
| 1 | −3144 | ∼0 |
| 2 | −2769 | ∼0 |
| 3 | −2678 | 0.993 |
| 4 | −2683 | 0.007 |
| 5 | −2688 | 0.00005 |

The values in the last column assume a uniform prior for $K$ ($K \in \{1, \ldots, 5\}$).

(figure from Pritchard et al., 2000)

# Population structure: Application to human genetics data

- Rosenberg et al. (2002) used this admixture model to study the genetic structure of human populations.
- They used genotype data from $n = 1056$ individuals from 52 pre-defined groups.
- Each individual was genotyped at $L = 377$ autosomal microsatellite loci.
- The pre-defined group labels were not used as inputs to the model.

# Human genetics example: Results on all data

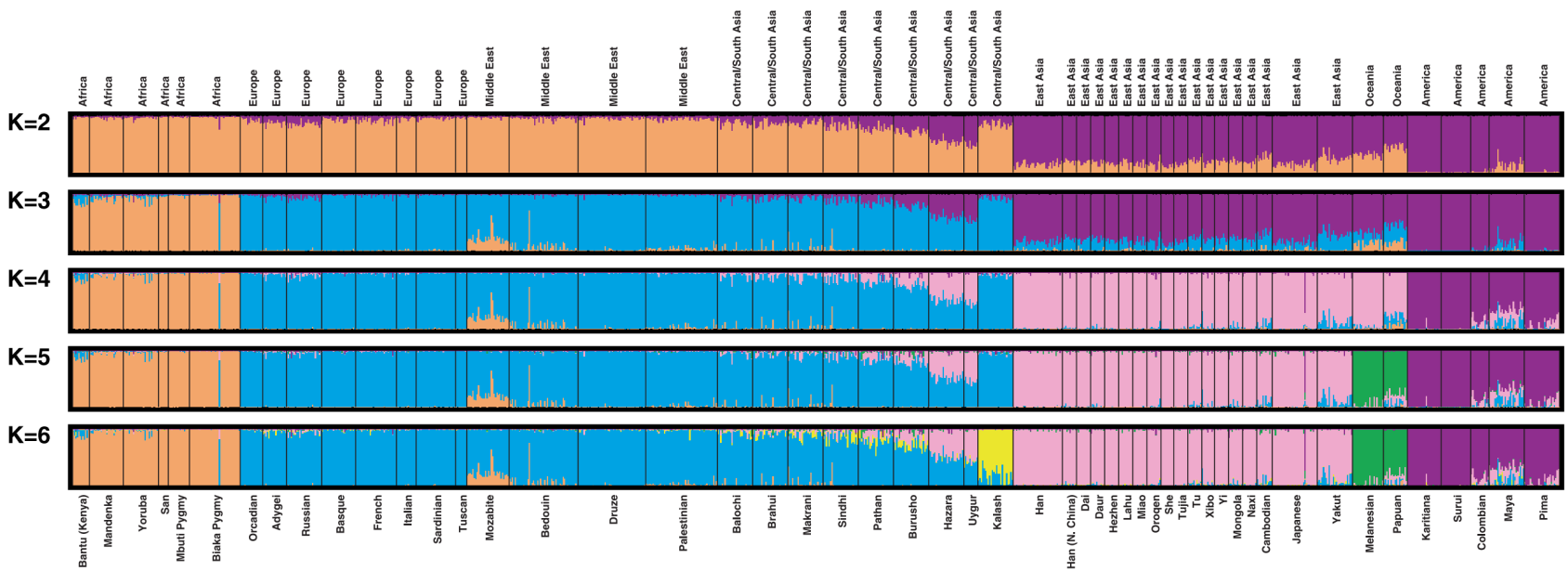
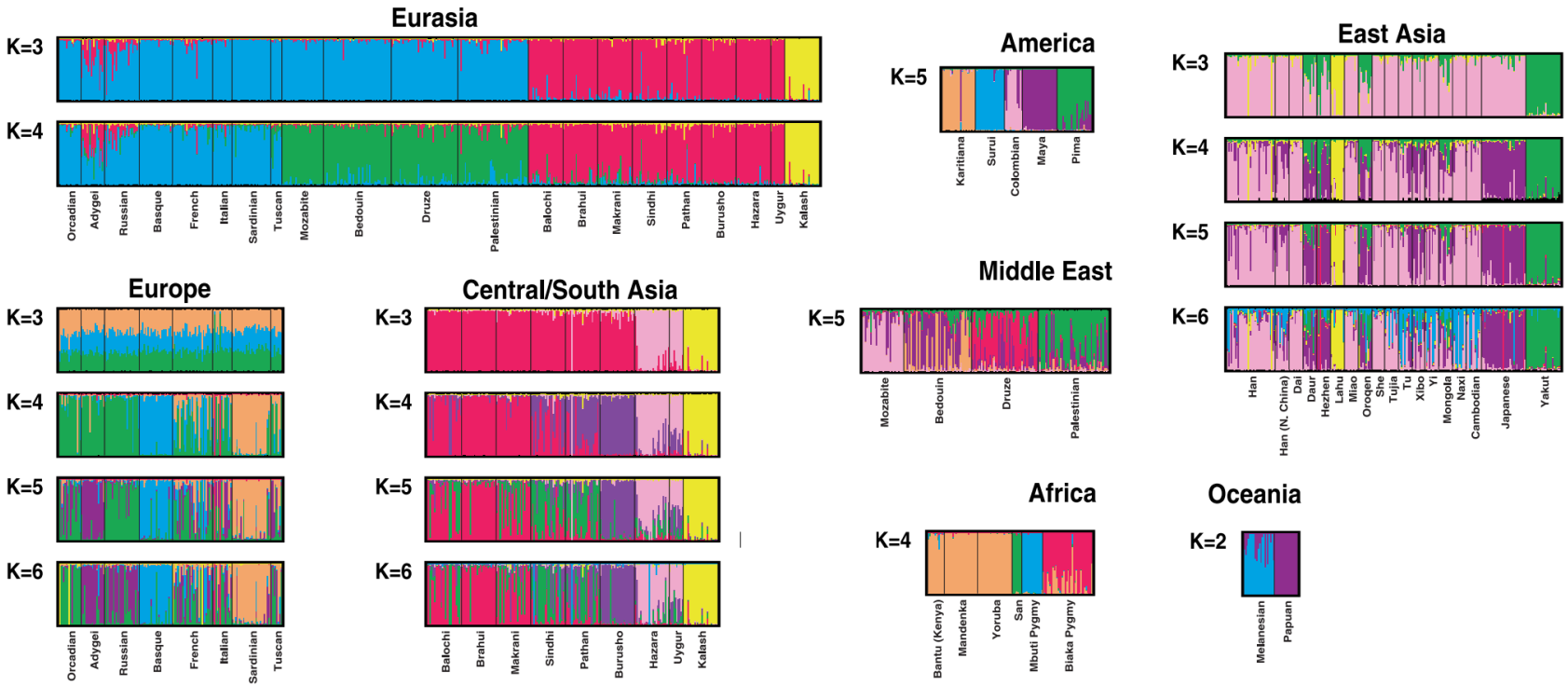



**Fig. 1.** Estimated population structure. Each individual is represented by a thin vertical line, which is partitioned into $K$ colored segments that represent the individual's estimated membership fractions in $K$ clusters. Black lines separate individuals of different populations. Populations are labeled below the figure, with their regional affiliations above it. Ten *structure* runs at each $K$ produced nearly identical individual membership coefficients, having pairwise similarity coefficients above 0.97, with the exceptions of comparisons involving four runs at $K = 3$ that separated East Asia instead of Eurasia, and one run at $K = 6$ that separated Karitiana instead of Kalash. The figure shown for a given $K$ is based on the highest probability run at that $K$.

(figure from Rosenberg et al., 2002)

- They identified six main genetic clusters, five of which correspond to major geographic regions.
- They also found subclusters that tend to correspond to pre-defined subgroups.

# Human genetics example: Results on each group




(figure from Rosenberg et al., 2002)

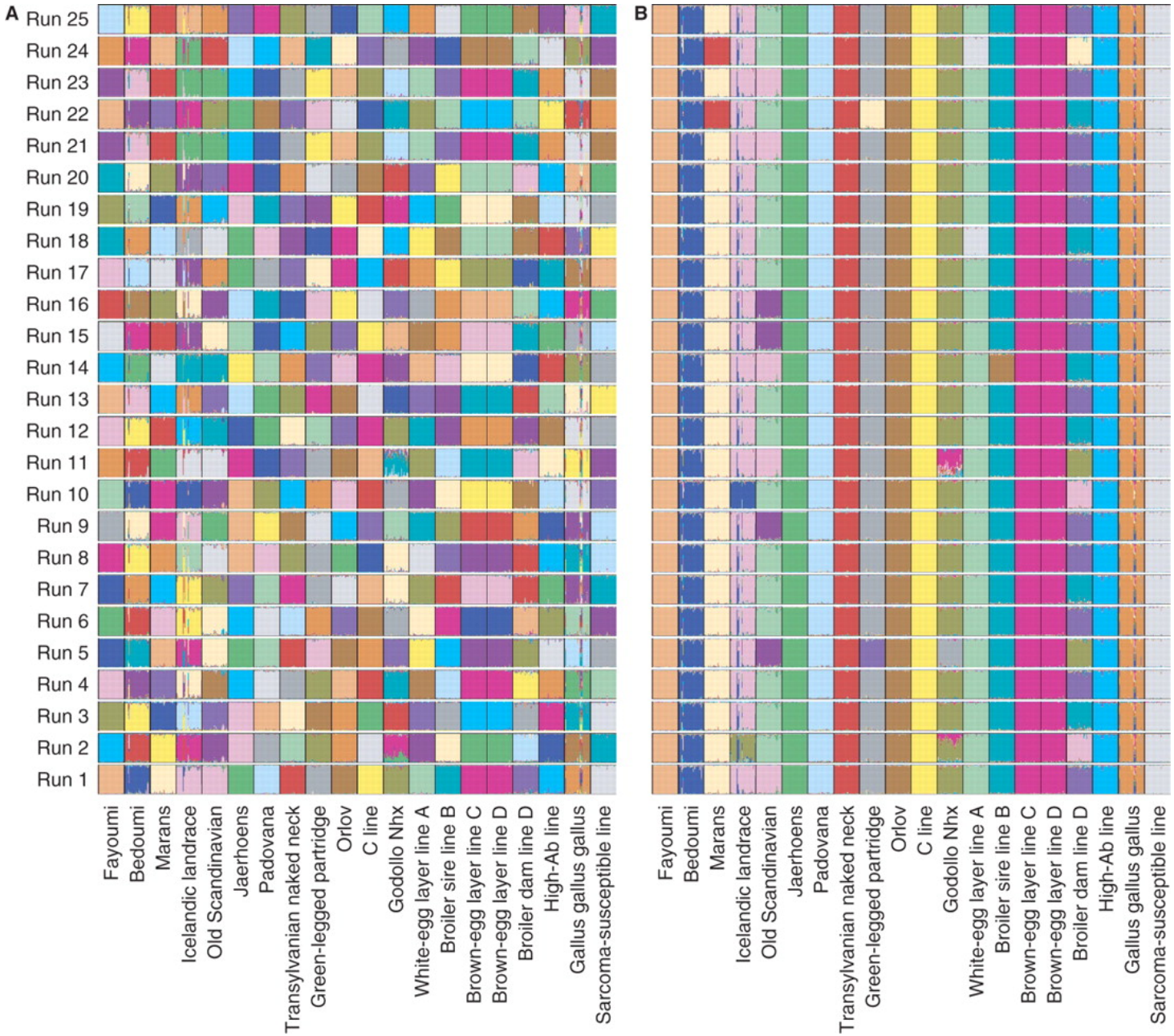
- They also ran the model separately on pre-defined regions.
- Again, the clusters tended to agree with known subgroups.

# Outline

# Issues in admixtures: Label switching

- Just like mixtures, admixtures are invariant to permutations of the component assignments.
- Thus, the label switching problem must be dealt with.
- As before, averaging label-invariant quantities is always valid.
- Meanwhile, to enable averaging over label-dependent quantites, a common approach is to relabel each posterior sample.
- CLUMPP does this by searching for labelings that minimize the distance between samples (Jakobsson et al., 2007).

# Issues in admixtures: CLUMPP program for label switching



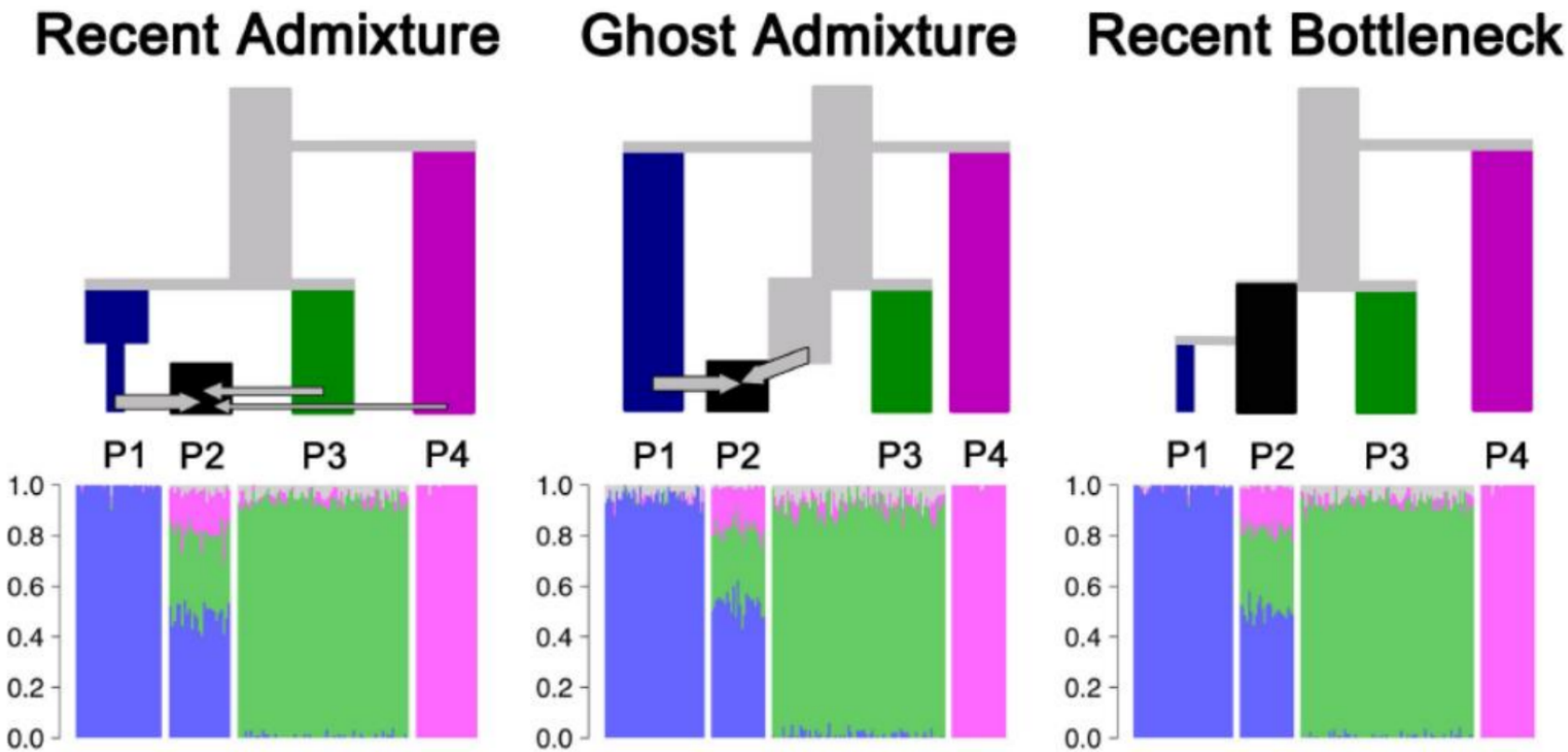

(figure from Jakobsson & Rosenberg, 2007)

# Issues in admixtures: Don't overinterpret the results

- Like mixtures, the admixture model is often misspecified. Thus, it is safest to view it as providing exploratory analysis.
- The results should not be interpreted as literally true. Overinterpreting the results can be very misleading.
- Instead, the results should be interpreted as the best fit of the model to the data.
- It is important to understand how different types of misspecification can affect the results.
- For instance, under misspecification, the model sometimes hallucinates populations.

# Issues in admixtures: Don't overinterpret the results

- Falush et al. (2016) provide instructive examples of how misspecification can lead to misleading results, if the results are taken as literally true.




Three different simulation scenarios that yield identical posteriors

(figure from Falush et al., 2016)

# Outline

# Latent Dirichlet allocation (LDA)

- LDA is a model for collections of discrete data such as documents.
- LDA is an admixture model in which each word in a document is drawn from a different topic.
- In LDA, topics are represented as distributions over words.
- The proportion of words coming from each topic varies from document to document.
- We will explore LDA in more detail when we cover variational inference.

# Group activity: Check your understanding

Go to breakout rooms and work together to answer these questions:
https://forms.gle/AZBXjVJTnZMjGrjd6

(Three people per room, randomly assigned. 10 minutes.)

# References and supplements


- Pritchard, J. K., Stephens, M., & Donnelly, P. (2000). Inference of population structure using multilocus genotype data. Genetics, 155(2), 945-959.
- Blei, D. M., Ng, A. Y., & Jordan, M. I. (2003). Latent Dirichlet allocation. Journal of Machine Learning Research, 3, 993-1022.
- Lorenzen, E. D., Arctander, P., & Siegismund, H. R. (2006). Regional genetic structuring and evolutionary history of the impala Aepyceros melampus. Journal of Heredity, 97(2), 119-132.
- Rosenberg, N. A., Pritchard, J. K., Weber, J. L., Cann, H. M., Kidd, K. K., Zhivotovsky, L. A., & Feldman, M. W. (2002). Genetic structure of human populations. science, 298(5602), 2381-2385.
- Jakobsson, M., & Rosenberg, N. A. (2007). CLUMPP: a cluster matching and permutation program for dealing with label switching and multimodality in analysis of population structure. Bioinformatics, 23(14), 1801-1806.
- Falush, D., van Dorp, L., & Lawson, D. J. (2016). A tutorial on how (not) to over-interpret STRUCTURE/ADMIXTURE bar plots. BioRxiv, 066431.