Population Structure

02-710 Computational Genomics

Seyoung Kim

The 23andMe Difference.

Receive an overview of your DNA – your 23 pairs of chromosomes – through detailed reports, tools and more.

Carrier Status reports*

If you are starting a family, find out if you are a carrier for an inherited condition.

Ancestry reports

Your DNA can tell you about your family history.

Wellness reports

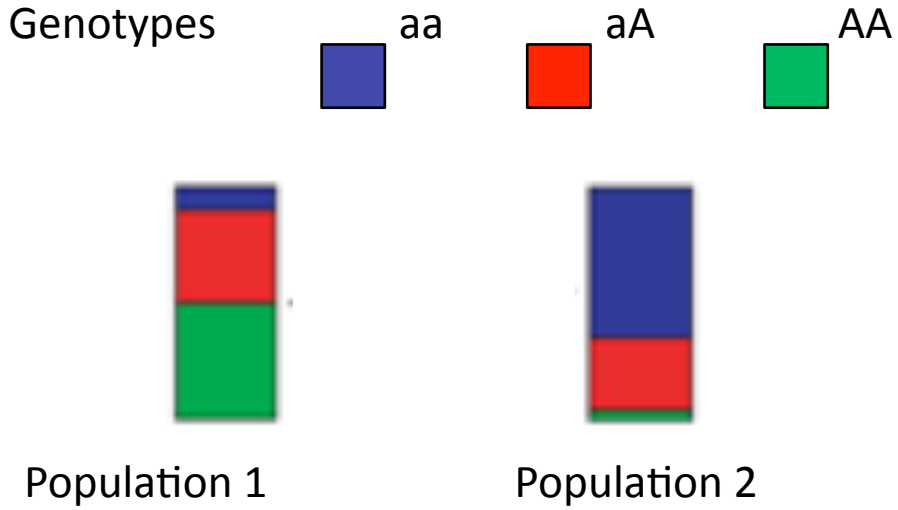
Your genetics can help you make more informed choices about your diet and exercise.

Traits reports

Explore what makes you unique, from food preferences to physical features.

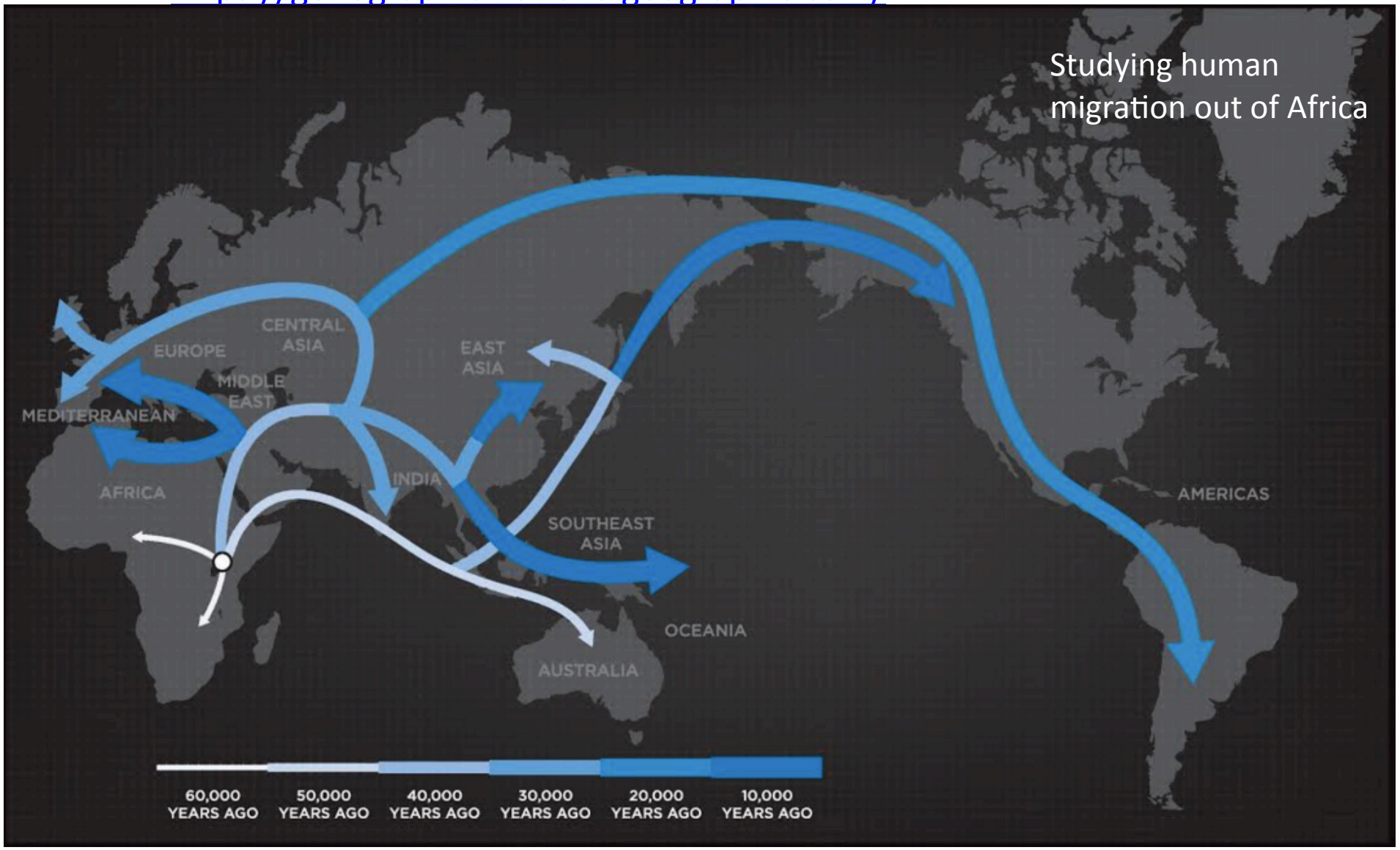
What is Population Structure?

- Population Structure
  - A set of individuals characterized by some measure of genetic distinction
  - A "population" is usually characterized by a distinct distribution over genotypes
  - Example



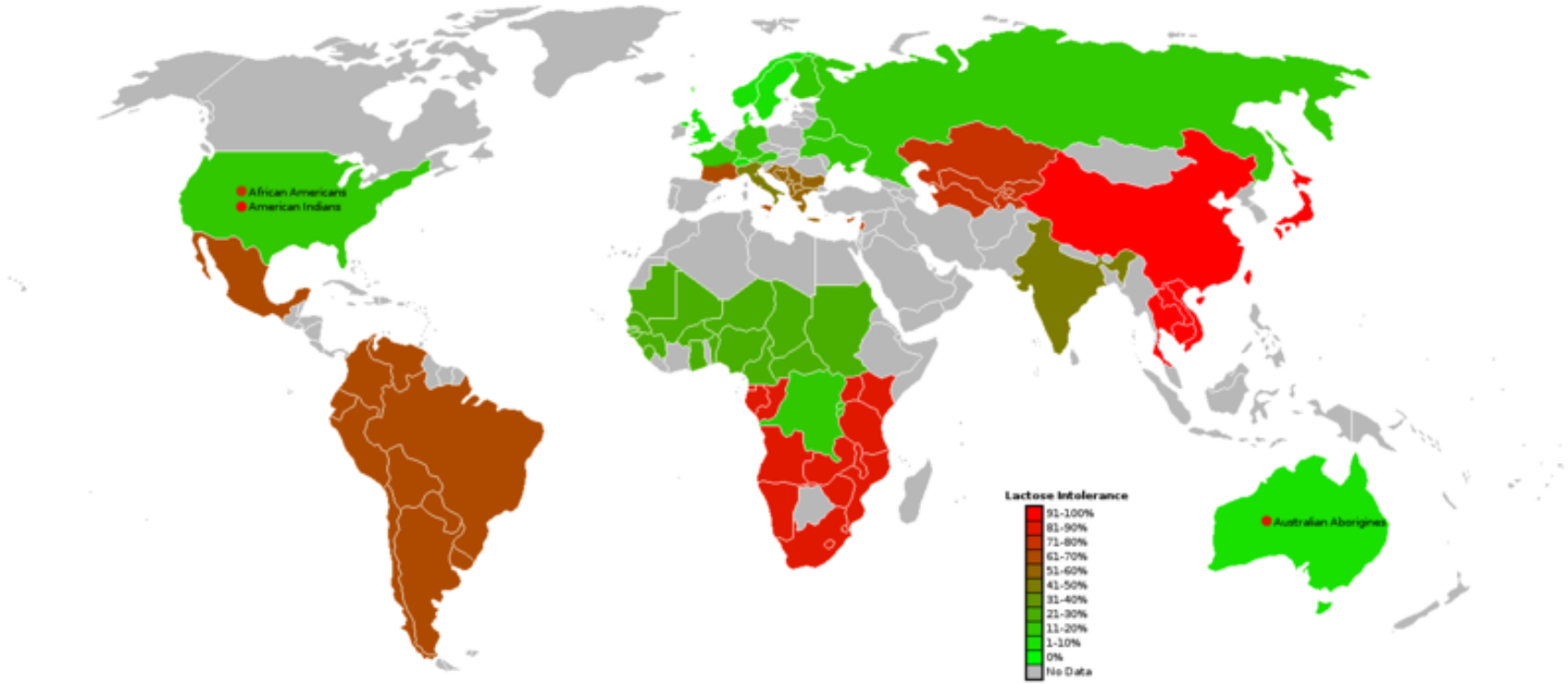
Motivation

- **Reconstructing individual ancestry:** The Genographic Project
  - https://genographic.nationalgeographic.com/



Motivation

- Study of various traits (e.g., lactose intolerance)



Overview

- Background
  - Hardy-Weinberg Equilibrium
  - Genetic drift
  - Wright's F_{ST}
- Inferring population structure from genotype data
  - Model-based method: Structure (Falush et al., 2003) for admixture model, linkage model
  - Principal component analysis (Patterson et al., PLoS Genetics 2006)

Hardy-Weinberg Equilibrium

- Hardy-Weinberg Equilibruim
  - Under random mating, both allele and genotype frequencies in a population remain constant over generations.
  - Assumptions of the standard random mating
    - Diploid organism
    - Sexual reproduction
    - Nonoverlapping generations
    - Random mating
    - Large population size
    - Equal allele frequencies in the sexes
    - No migration/mutation/selection
  - Chi-square test for Hardy-Weinberg equilibrium

Genotype/Allele Frequencies in the Current Generation

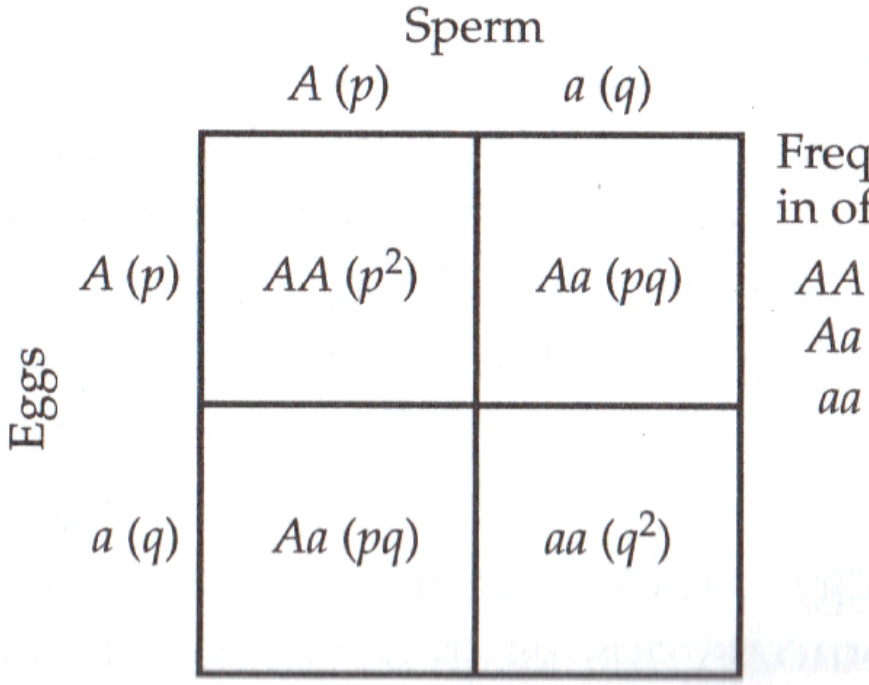
- Genotype frequencies in the current generation
  - *D*: frequency for *AA*
  - *H*: frequency for *Aa*
  - *R*: frequency for *aa*

  - $D + H + R = 1.0$

- Allele frequencies in the current generation
  - *p:* : frequency of *A*
    - $p = (2D + H) / 2 = D + H/2$
  - *q*: frequency of *a*
    - $q = (2R + H) / 2 = R + H/2$

Genotype/Allele Frequencies of the Offspring

- Genotype frequencies in the offspring
  - *D'*: frequency for *AA*
    - $D' = p^2$
  - *H'*: frequency for *Aa*
    - $H' = pq + pq = 2pq$
  - *R'*: frequency for *aa*
    - $R' = q^2$



- Allele frequencies in the offspring
  - $p' = (2D' + H')/2$
    $= (2p^2 + 2pq)/2 = p(p + q) = p$
  - $q' = (2R' + H')/2 = (2q^2 + 2pq)/2 = q(q + p) = q$

Testing Whether Hardy-Weinberg Equilibrium Holds in Data

- Given genotypes collected from a population, does HWE hold at the given locus?

- Chi-square test
  - Null hypothesis: HWE holds in the observed data
  - Test if the null hypothesis is violated in the data by comparing the observed genotype frequencies with the expected frequencies

Testing Whether Hardy-Weinberg Equilibrium Holds

Step 1: Compute allele frequencies from the observed data

$$p = \frac{224 \times 2 + 64}{294 \times 2} = 0.871$$

$$q = 1 - p = 0.129$$

Contingency table for chi-square test

Genotype	AA	Aa	aa	Total
Observed	224	64	6	294
Expected	?	?	?	294

Testing Whether Hardy-Weinberg Equilibrium Holds

Step 1: Compute allele frequencies from the observed data

$$p = \frac{224 \times 2 + 64}{294 \times 2} = 0.871$$

$$q = 1 - p = 0.129$$

Contingency table for chi-square test

Genotype	AA	Aa	aa	Total
Observed	224	64	6	294
Expected	222.9	66.2	4.9	294

Step 2: Compute the expected genotype frequencies

$$\text{Expected(AA)} = p^2 n = 0.8707^2 \times 294 = 222.9$$

Step 3: Compute the test statistic (degree of freedom 1)

$$\chi^2 = \sum \frac{(\text{observed} - \text{expected})^2}{\text{expected}}$$

$$= \frac{(224 - 222.9)^2}{222.9} + \frac{(64 - 66.2)^2}{66.2} + \frac{(6 - 4.9)^2}{4.9}$$

$$= 0.32$$

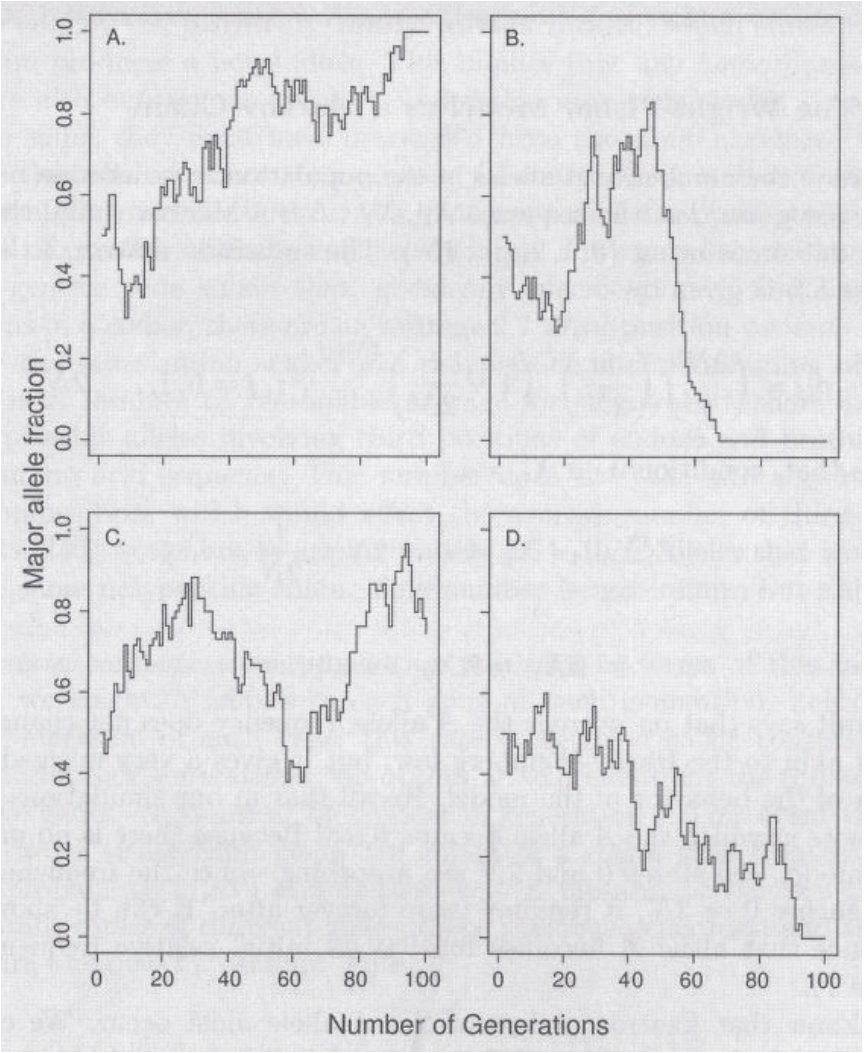
Hardy-Weinberg Equilibrium in Practice

- HWE often does not hold in reality because of the violation of the assumptions (i.e., random mating, no selection, etc.)

- Even when the assumptions for HWE hold, in reality, allele frequencies change over generations because of the random fluctuation – genetic drift!

Genetic Drift

- The change in allele frequencies in a population due to random sampling
- All mutations eventually drift to allele frequency 0 or 1 over time
- Neutral process unlike natural selection
  - But genetic drift can eliminate an allele from the given population.
- The effect of genetic drift is larger in a small population



Wright-Fisher Model

- Model for genetic drift
  - Assume population size *N*, which does not change from generation to generation. Thus, 2*N* copies of genes.
  - *p*, *q*: allele frequencies of two alleles
  - the probability that we will have *k* copies of one allele (with frequency *p* in the current generation) in the next generation is given as:

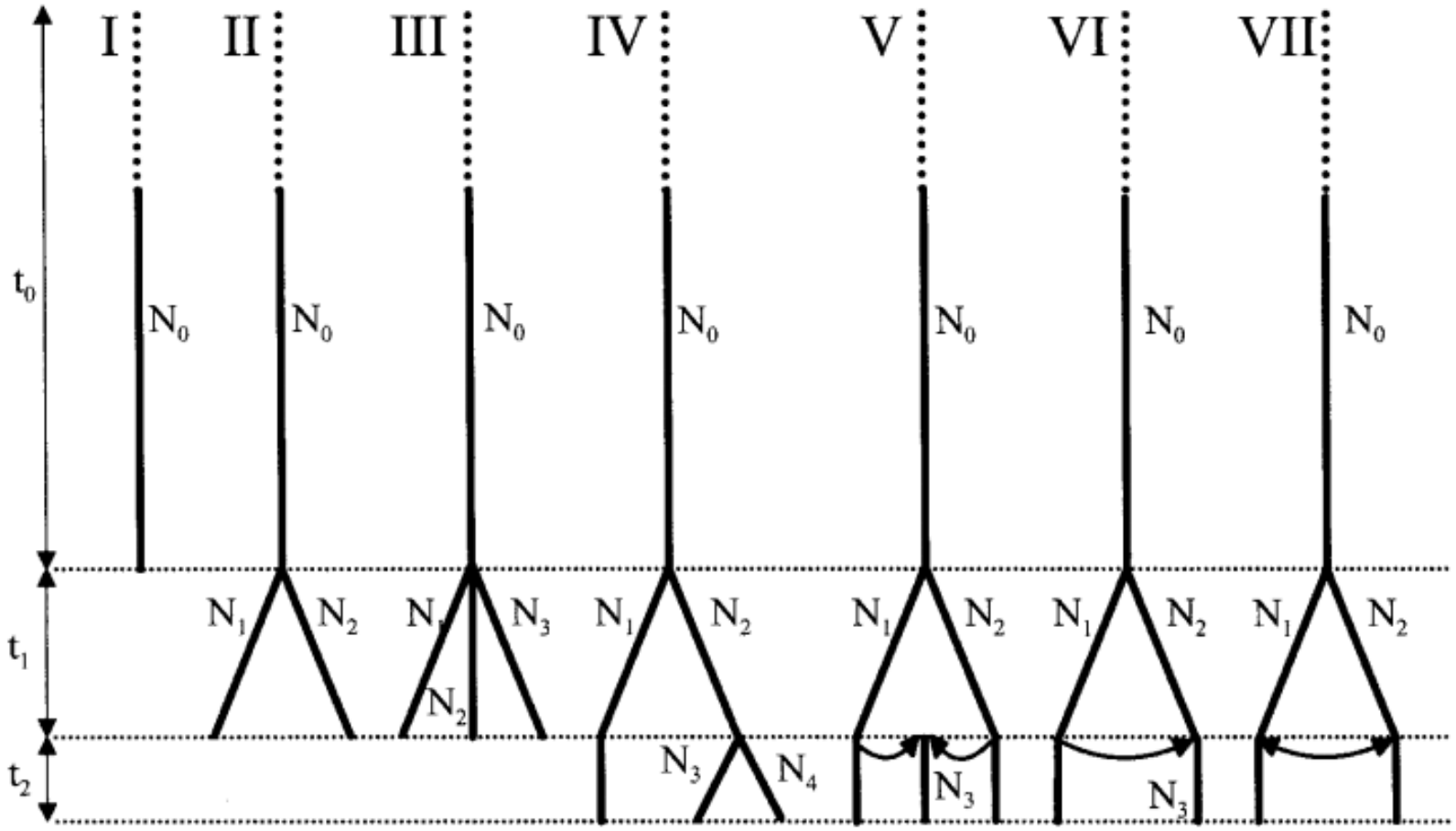
$$\binom{2N}{k} p^k q^{2N-k}$$

Population Divergence and Admixture

- Population divergence
  - Once a single population is separated into two subpopulations, each of the subpopulations will be subject to its own genetic drift and natural selection
  - Population divergence creates different allele frequencies for the same loci across different populations

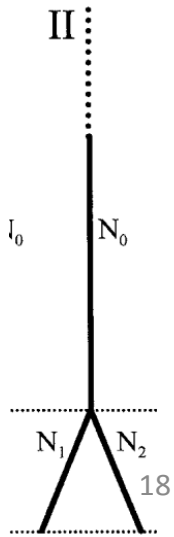
- Admixture
  - Two previously separated populations migrate/mate to form an admixed population

Scenarios of How Populations Evolve



Population Divergence

- Wright's F_{ST}
  - Statistics used to quantify the extent of divergence among multiple populations relative to the overall genetic diversity
  - Summarizes the average deviation of a collection of populations away from the mean
  - $F_{ST} = \mathrm{Var}(p_k)/p'(1-p')$
    - p': the overall frequency of an allele across all subpopulations
    - p_k : the allele frequency within population k



Overview

- Background
  - Hardy-Weinberg Equilibrium
  - Genetic drift
  - Wright's F_{ST}
- Inferring population structure from genotype data
  - Model-based method: Structure (Falush et al., 2003) for admixture model, linkage model
  - Principal component analysis (Patterson et al., PLoS Genetics 2006)

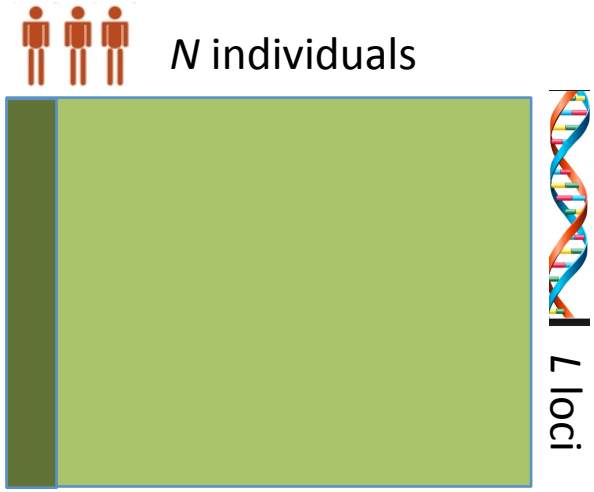
Probabilistic Models for Population Structure

- Mixture model
  - Clusters individuals into *K* populations
  - Does not model admixture
- Admixture model
  - The genotypes of each individual are an admixture of multiple ancestor populations
  - Assumes alleles are in linkage equilibrium
- Linkage model
  - Models recombination, correlation in alleles across chromosomes

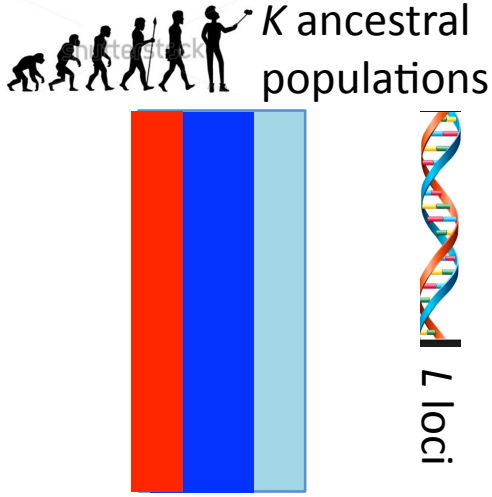
Structure Model

- Hypothesis: Modern populations are created by an intermixing of ancestral populations.
- An individual's genome contains contributions from one or more ancestral populations.
- The contributions of populations can be different for different individuals.
- Other assumptions
  - No linkage disequilbrium
  - Markers are i.i.d (independent and identically distributed)

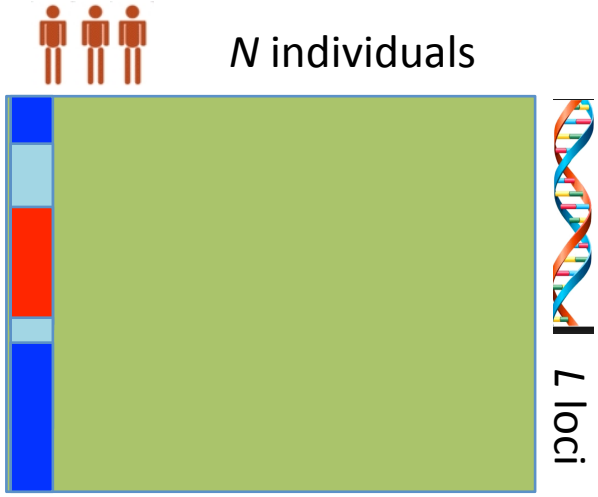
Modeling Admixture



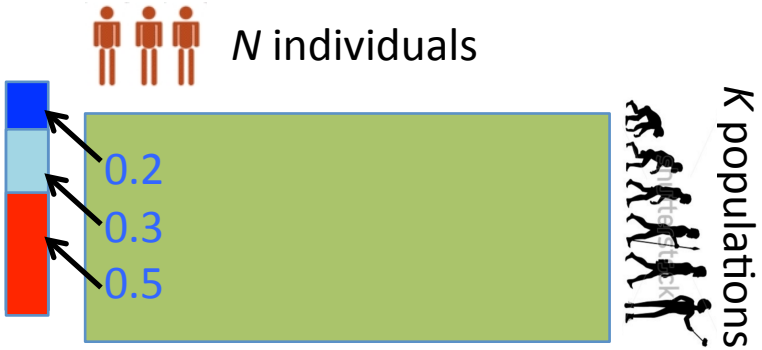
Genotype Data



Allele frequencies at each locus for each ancestral population

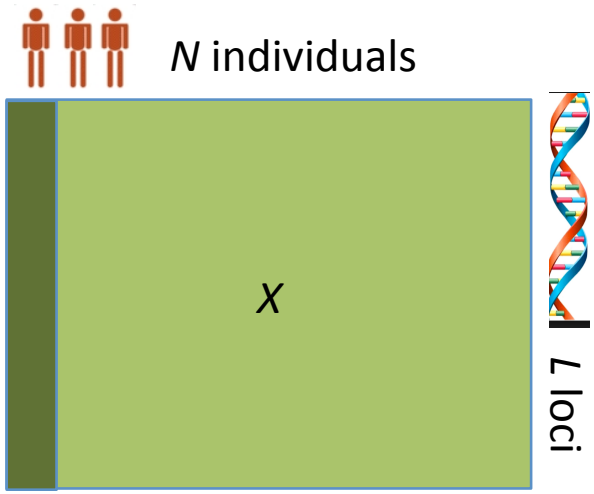


Admixed Genome

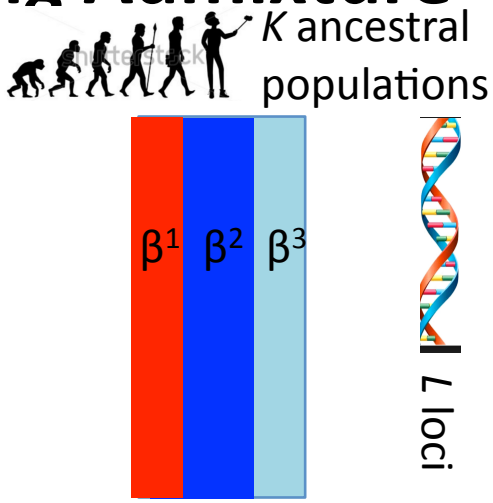


Individuals' ancestry proportions (Column sum to 1)

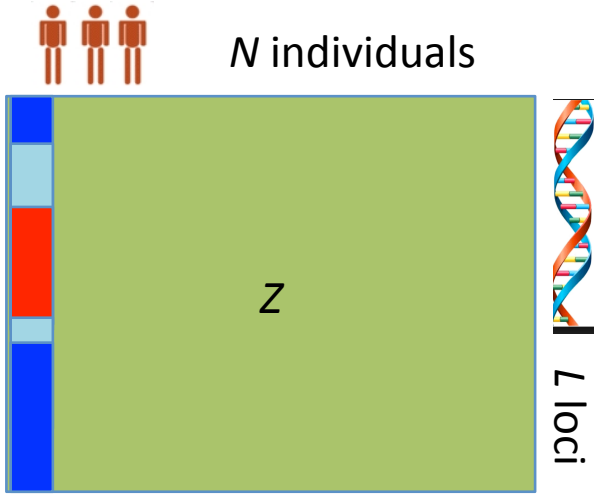
STRUCTURE for Modeling Admixture



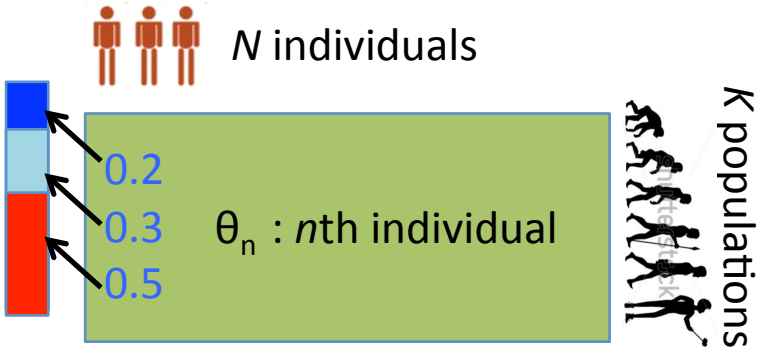
Genotype Data



Allele frequencies at each locus
for each ancestral population



Admixed Genome

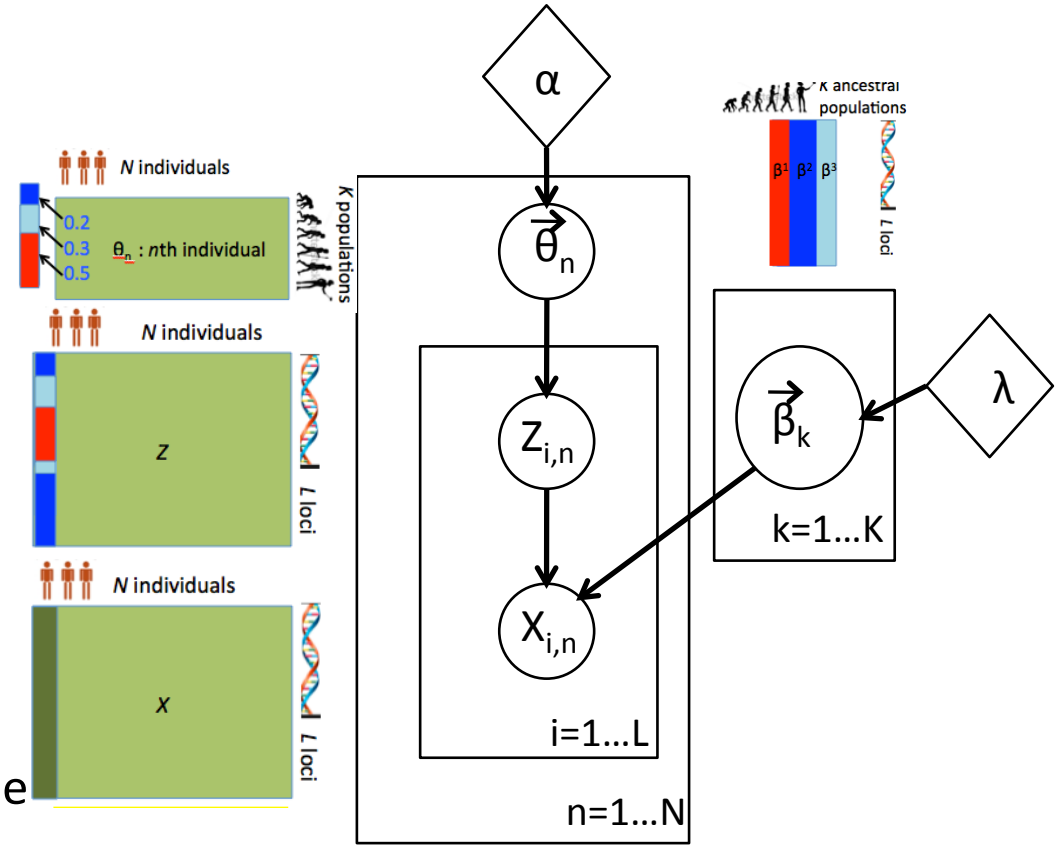


Individuals' ancestry proportions
(Column sum to 1)

Generative Model for STRUCTURE

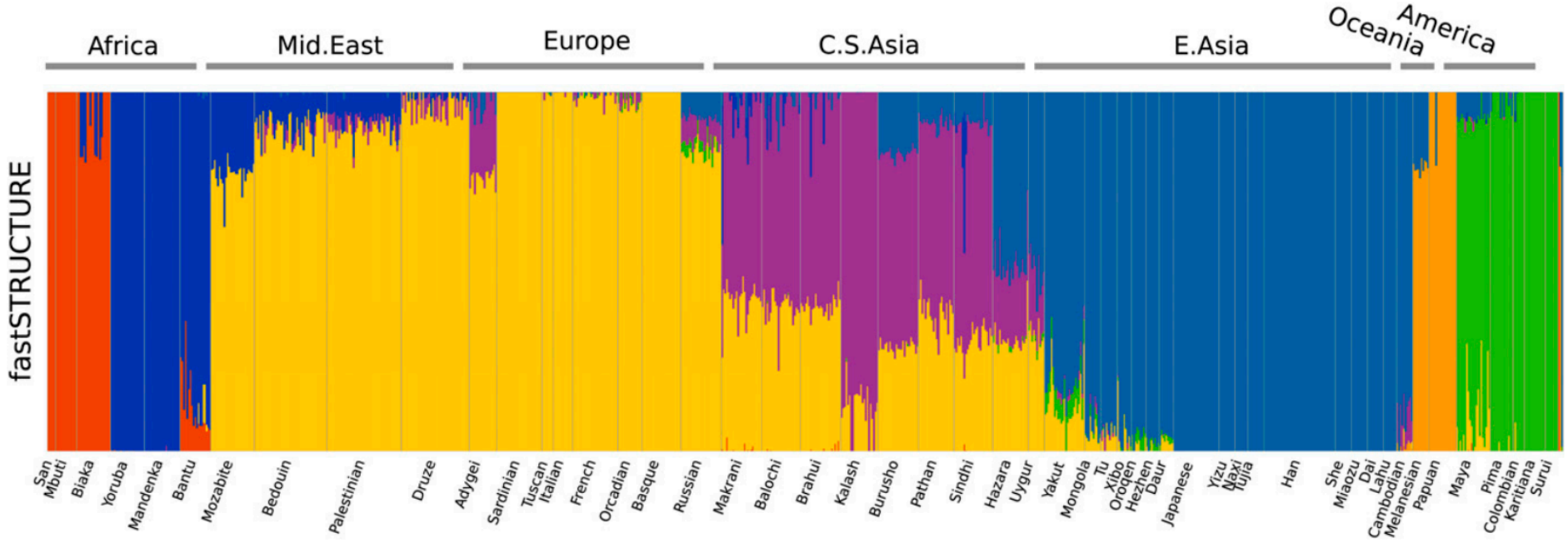
- β^k : Allele frequencies for population *k* at *L* loci

- For each individual n=1,…,N
  - Sample θ_n from Dirichlet(α)
  - For each locus i=1,…,L
    - Sample $Z_{i,n}$ from Multinomial(θ_n)
    - Sample $X_{i,n}$ from $\beta_{k,i}$ for the population chosen by $k=Z_{i,n}$



Inferring Ancestry with STRUCTURE

- Human Genome Diversity Project
  - 938 individuals from 51 populations, 657, 143 loci
  - Fit Structure model with K = 7 subpopulations
  - Infer ancestry proportions for all individuals



Each column: inferred ancestry proportion for each individual

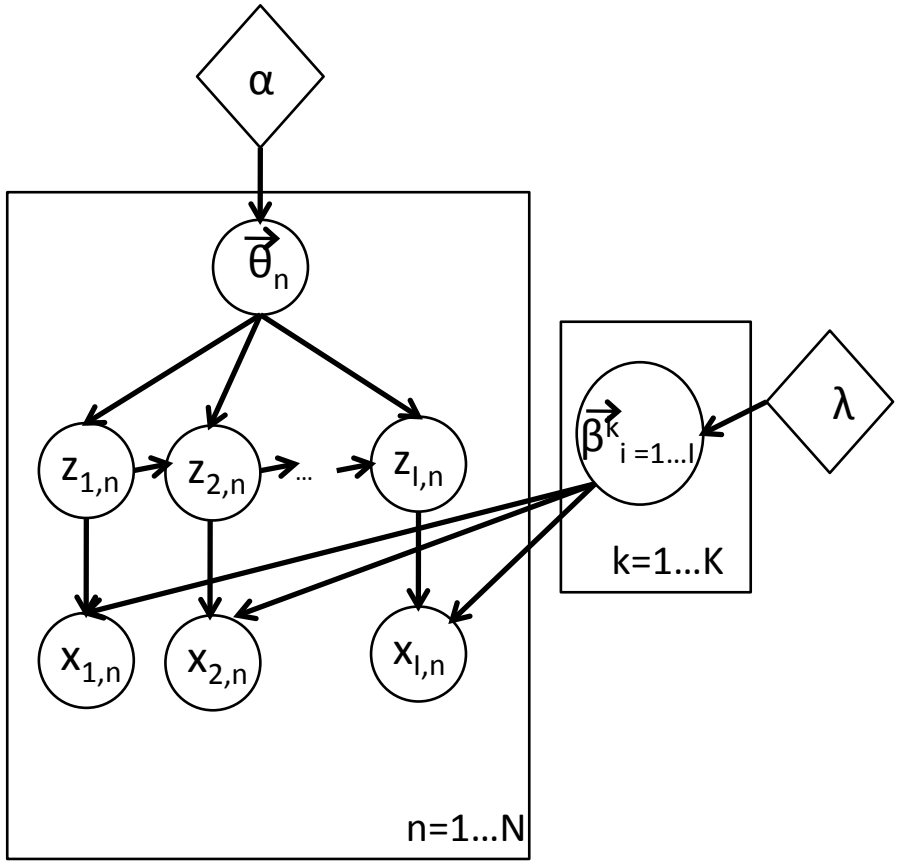
Structure Model

- Advantages
  - Generative process
    - Explicit model of admixture
  - Meaningful interpretable results
  - Clustering is probabilistic
    - Models uncertainty in clusters or population labels

- Disadvantages
  - Alleles are same in ancestral and modern populations
  - No models of mutation, recombination

Extending *Structure* to Model Linkage

- From admixture model, replace the assumption that the ancestry labels Z_{il} for individual *i*, locus *l* are independent with the assumption that adjacent Z_{il} are correlated.



- Use Poisson process to model the correlation between neighboring alleles
  - d_l : distance between locus *l* and locus *l+1*
  - *r*: recombination rate

$$\Pr(z^{(i)}_{l+1} = k' | z^{(i)}_l = k, r, Q) = \begin{cases} \exp(-d_l r) + (1 - \exp(-d_l r)) q^{(i)}_{k'} \\ \quad \text{if } k' = k \\ (1 - \exp(-d_l r)) q^{(i)}_k \\ \quad \text{otherwise,} \end{cases}$$

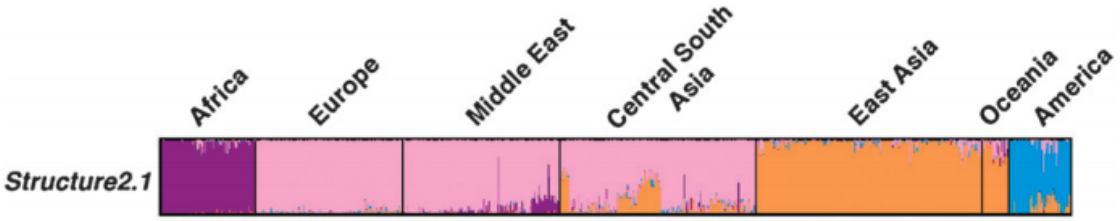
Extending *Structure* to Model Linkage

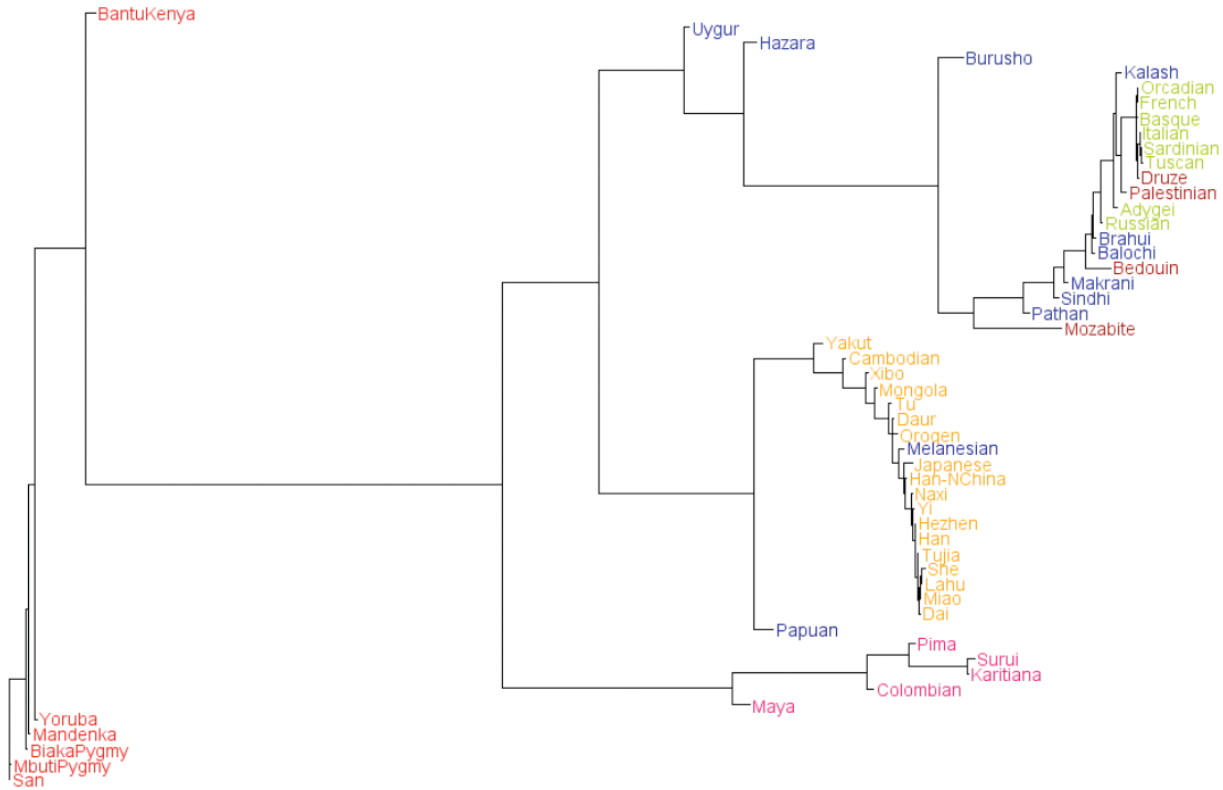
- As recombination rate *r* goes to infinity, all loci become independent and linkage model becomes admixture model.

- Recombination rate *r* can be viewed as being related to the number of generations since admixture occurred.

- Use MCMC algorithm or variational algorithms to fit the unkown parameters.

Neighbour-joining Phylogenetic Trees from the Structural Maps





Overview

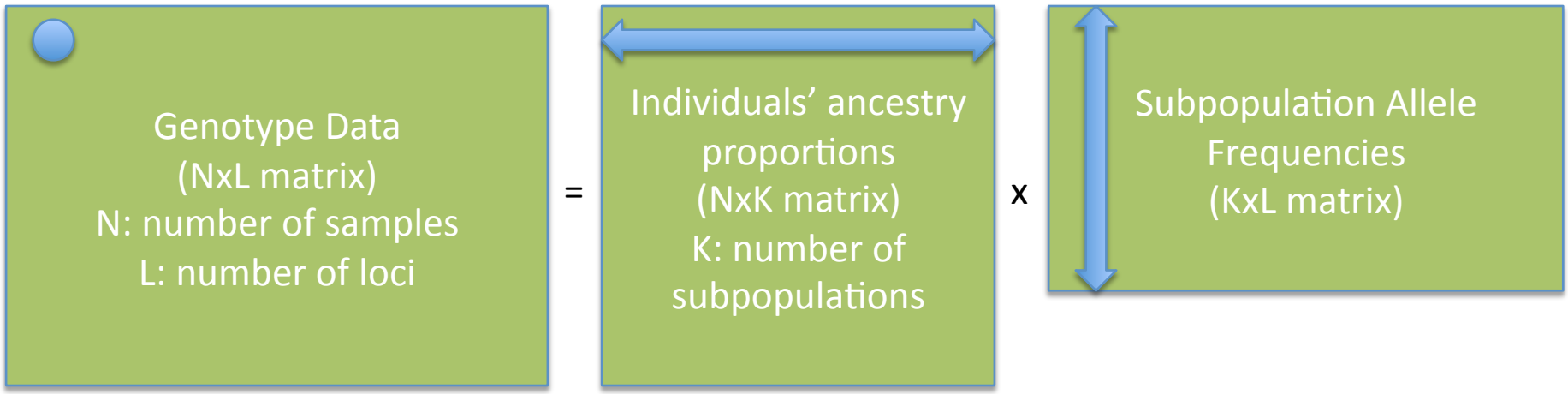
- Background
  - Hardy-Weinberg Equilibrium
  - Genetic drift
  - Wright's F_{ST}
- Inferring population structure from genotype data
  - Model-based method: Structure (Falush et al., 2003) for admixture model, linkage model
  - Principal component analysis (Patterson et al., PLoS Genetics 2006)

Low-dimensional Projections

- Genetic data is very large
  - Number of markers may range from a few hundreds to hundreds of thousands
  - Thus each individual is described by a high-dimensional vector of marker configurations
  - A low-dimensional projection of each individual allows easy visualization
- Technique used
  - Factor analysis
  - Many statistical methods exist – ICA, PCA, NMF etc.
  - Principal Components Analysis (next slide)
- Usually projected to 2 dimensions to allow visualization

Matrix Factorization and Population Structure

- Matrix factorization for learning population structure



$$E[G] = \Lambda F$$

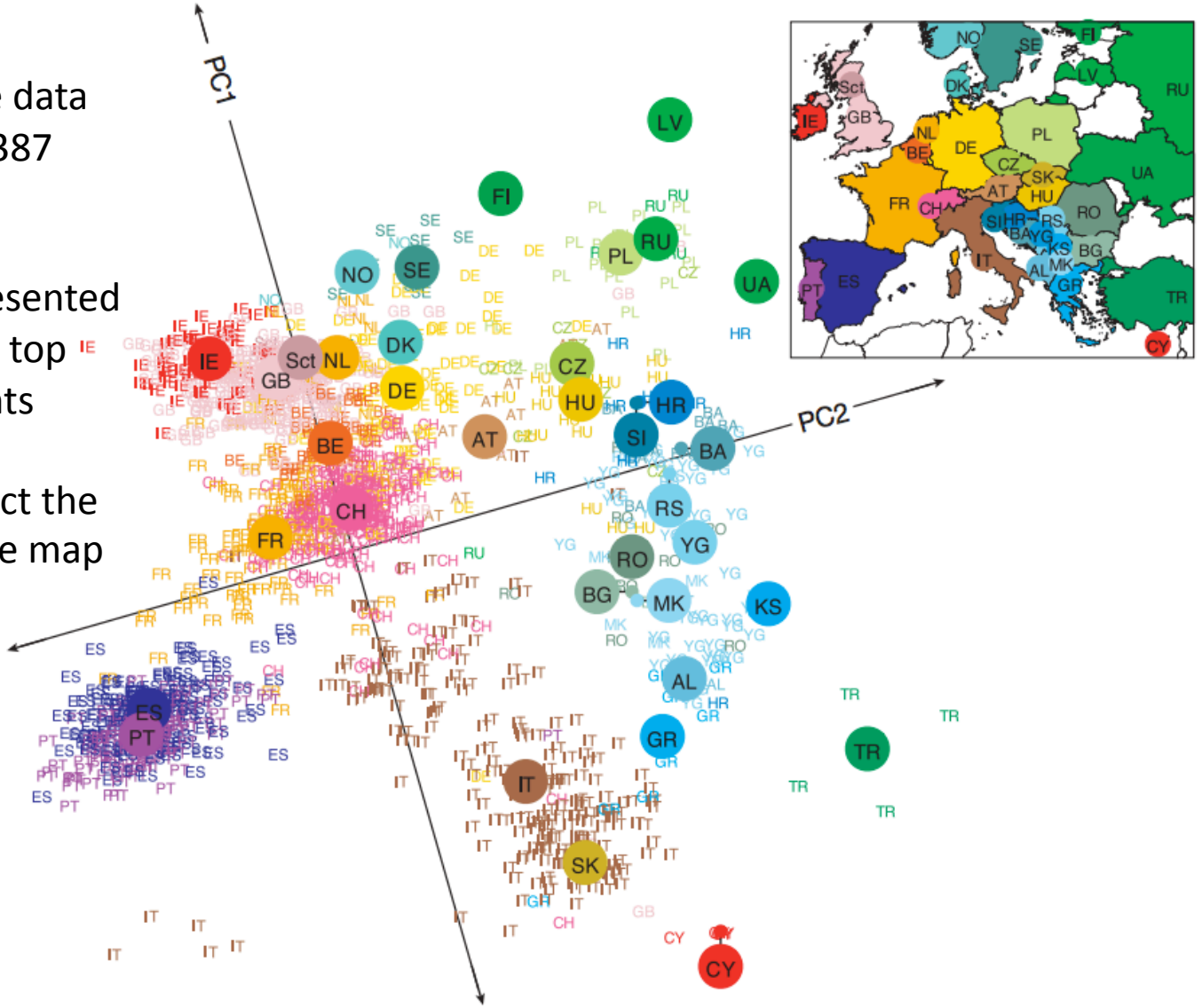
Principal Component Analysis to Reveal Population Structure

- Genotype data X
  - N x L matrix for N individuals and L loci
  - Normalize each column of the genotype data matrix

- Perform PCA on the covariance matrix (1/N)XX'
  - K principal components with top eigenvalues capture the ancestry information

Population Structure In Europe

- Apply PCA to genotype data from 197, 146 loci in 1, 387 European individuals
- Each individual is represented in two dimensions using top two principle components
- 2-dim projections reflect the geographic regions in the map quite well



Comparison of Different Methods

	PCA	Model-based Clustering
Advantages	• Easy visualization	• Generative process that explicitly models admixture • Clustering is probabilistic: it is possible to assign confidence level of clusters

Disadvantages	• No intuition about underlying processes	• Computationally more demanding
		• Based on assumptions of evolutionary models:
		  • Structure: No models of mutation, recombination
		  • Recombination added in extension linkage model by Falush et al.

Summary

- Genetic variation data can be used to infer various aspects of population history such as population divergence admixture.
- HWE describes the theoretical allele frequencies in the ideal situation.
- Genetic drift and natural selection can change allele frequencies from generation to generation.
- Model-based methods such as Structure or matrix-factorization methods can be used to infer population structure from genotype data.