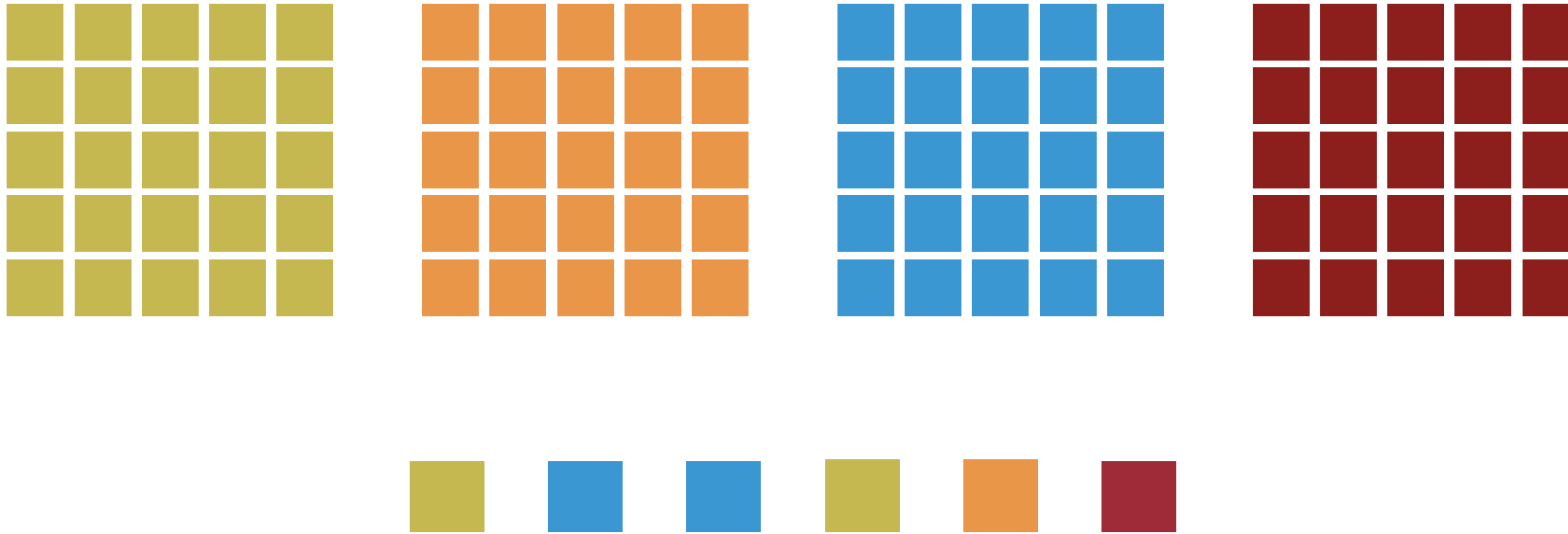# Lecture 5

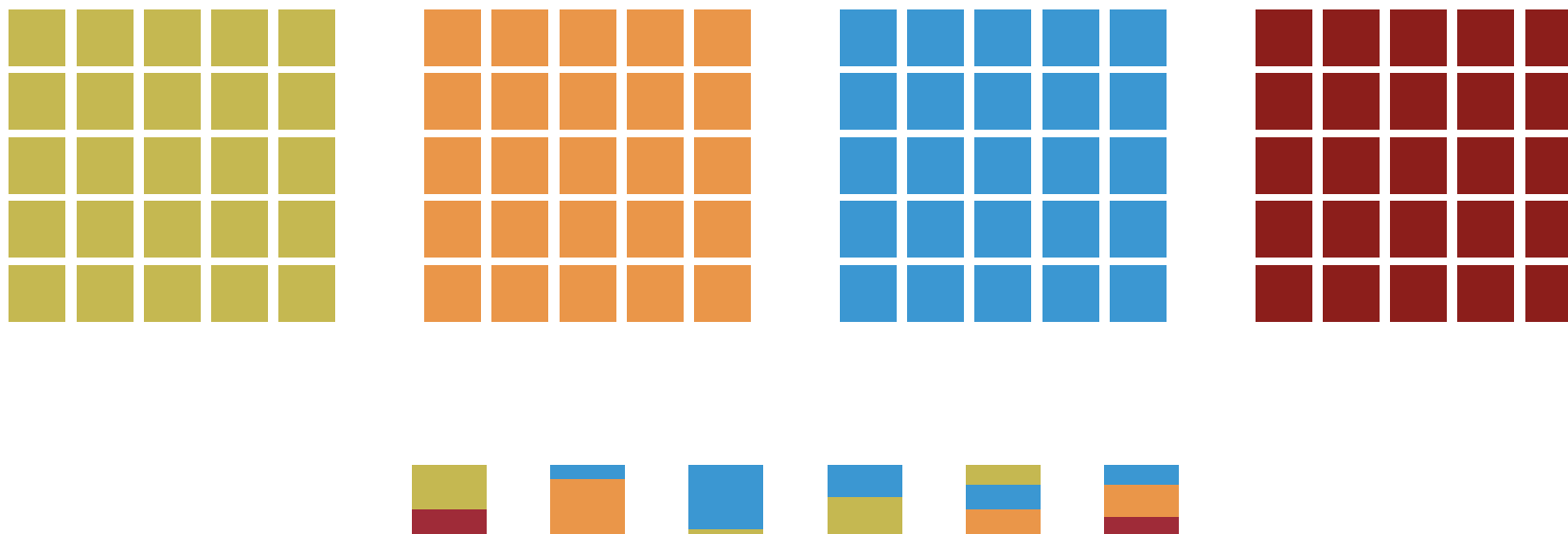# Reference Populations

Identify **origins** of individuals
each with a **single** ancestry

Approaches: Based on Hardy-Weinberg genotype proportions

Reference Populations (known or unknown)

Identify **ancestry proportions** for individuals with **admixed** ancestry

Approaches: Structure (MCMC, Bayesian)
Or ADMIXTURE (quadratic programming)

Genetic Structure
of Human Populations

Noah A. Rosenberg,[1]* Jonathan K. Pritchard,[2] James L. Weber,[3]
Howard M. Cann,[4] Kenneth K. Kidd,[5] Lev A. Zhivotovsky,[6]
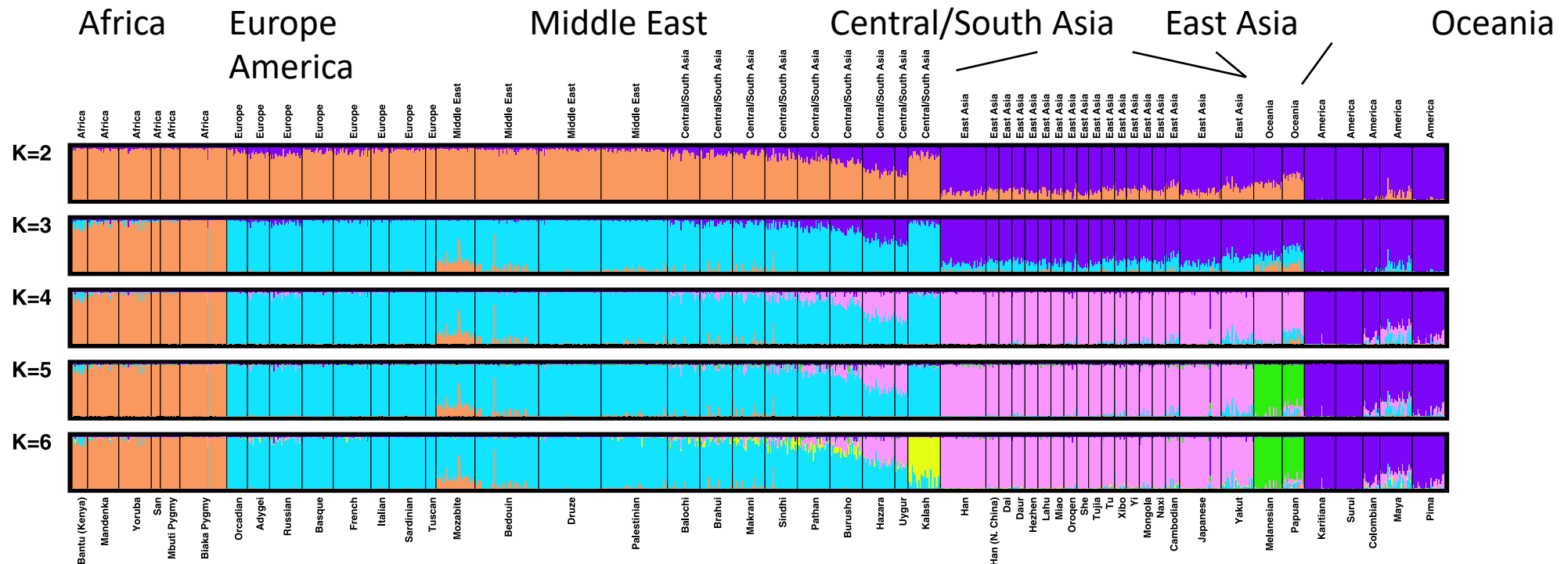Marcus W. Feldman[7]

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*.
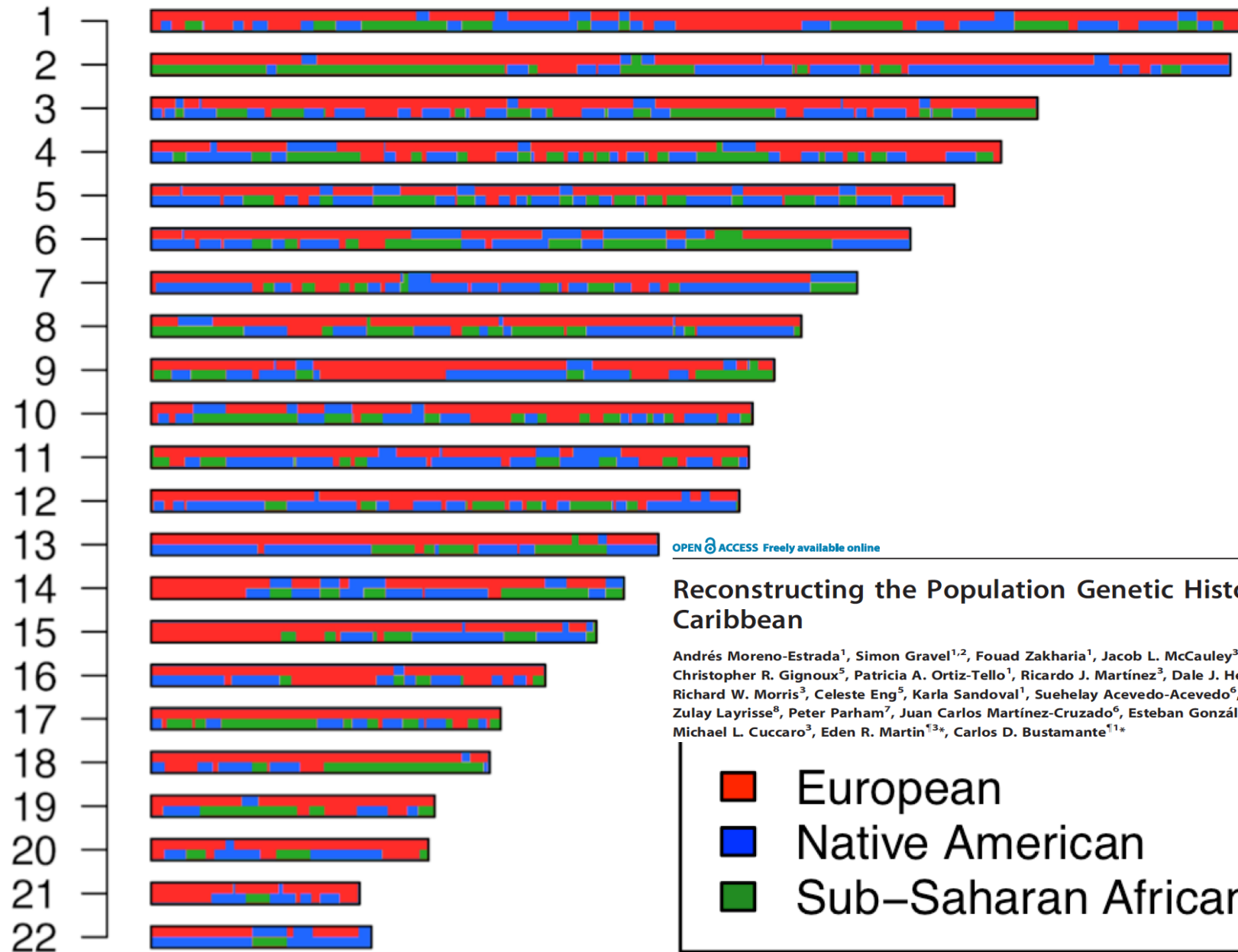
Reference Populations
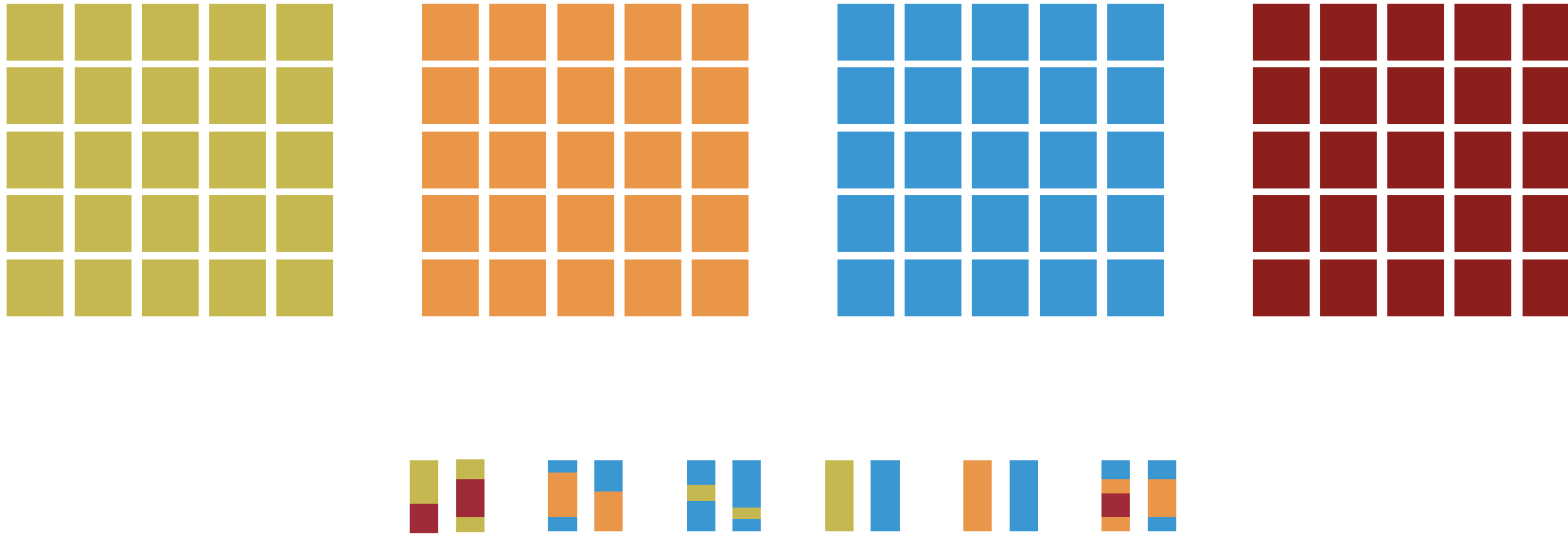
Identify origins of **chromosomal segments** in individuals of **admixed** ancestry

Approaches: Based on Hidden Markov Models

Local ancestry

○ **PLOS** | GENETICS

Reconstructing the Population Genetic History of the Caribbean

Andrés Moreno-Estrada[1,2], Simon Gravel[1,2], Fouad Zakharia[1], Jacob L. McCauley[3], Jake K. Byrnes[1,4], Christopher R. Gignoux[5], Patricia A. Ortiz-Tello[1], Ricardo J. Martínez[3], Dale J. Hedges[3], Richard W. Morris[3], Celeste Eng[5], Karla Sandoval[1], Suehelay Acevedo-Acevedo[6], Paul J. Norman[7], Zulay Layrisse[8], Peter Parham[7], Juan Carlos Martínez-Cruzado[6], Esteban González Burchard[5], Michael L. Cuccaro[3], Eden R. Martin[¶3]*, Carlos D. Bustamante[¶1]*

🟥 European
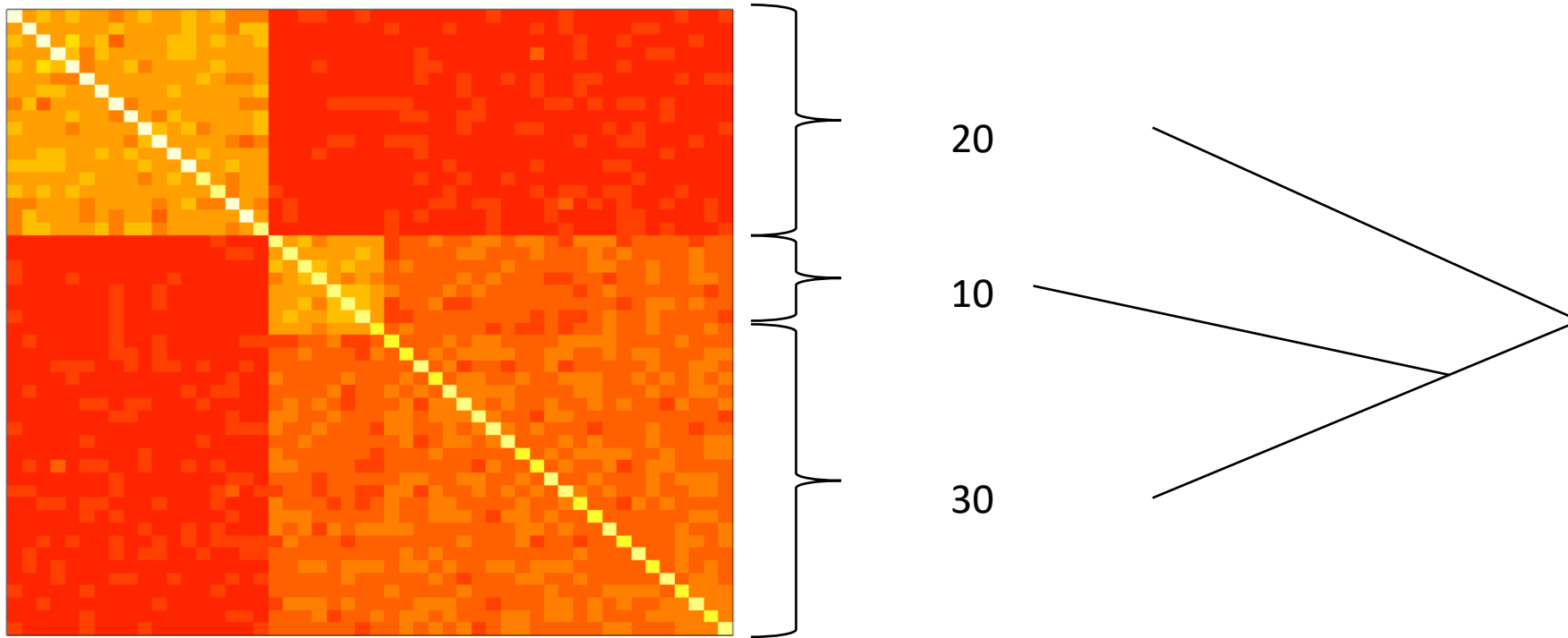🟦 Native American
🟩 Sub–Saharan African

Principal components analysis

Example

(Simulated data, N=50 individuals, L=1000 SNPs)

Relatedness matrix R

20

10

30

i^{th} and j^{th} entry =average over loci (l) of $(X_{li}-\overline{X_l})(X_{lj}-\overline{X_l})$
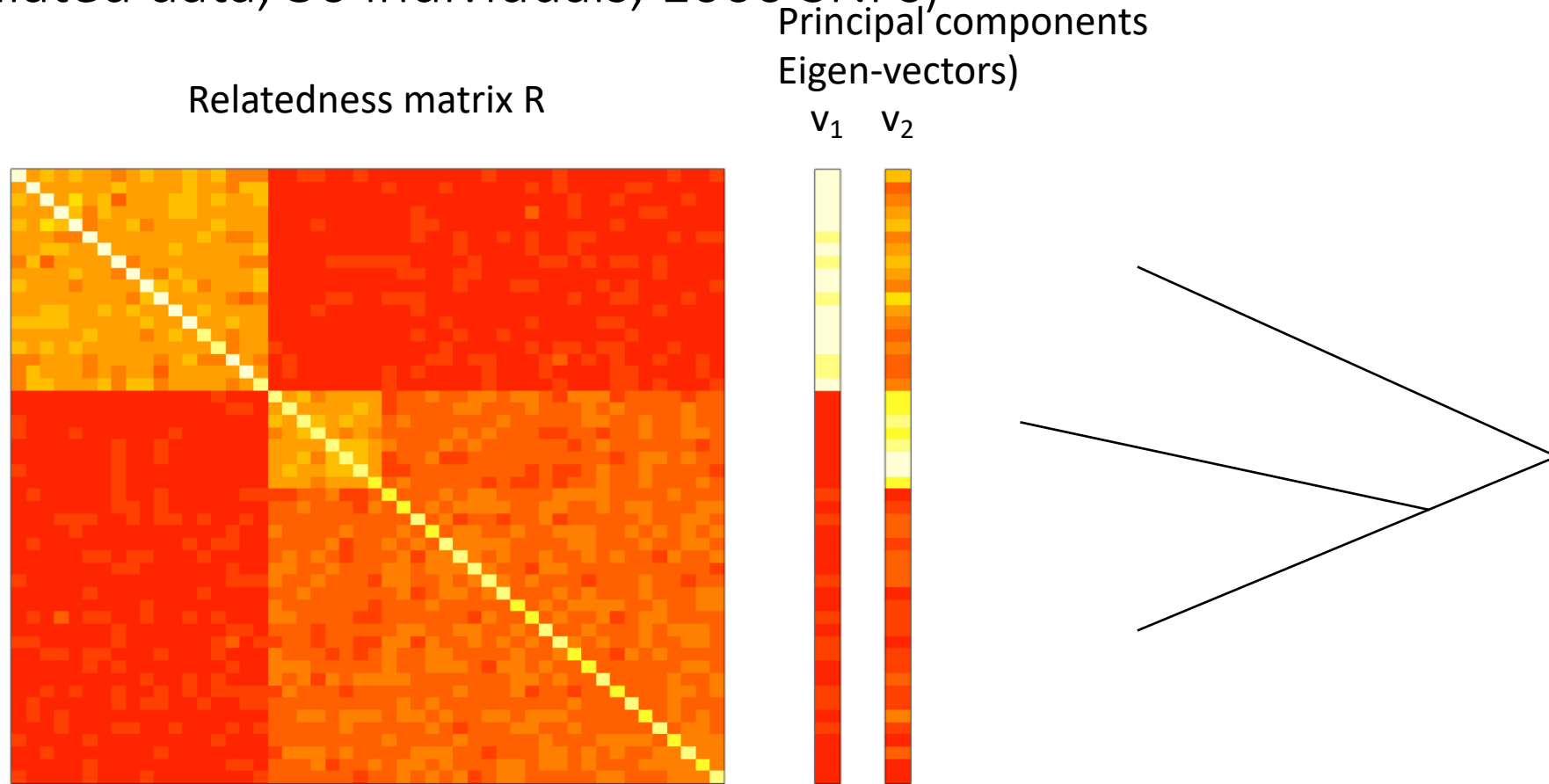
Where X_l is mean freq. of the lth locus.

Modified from slide by Gavin Band

Principal components analysis

Example
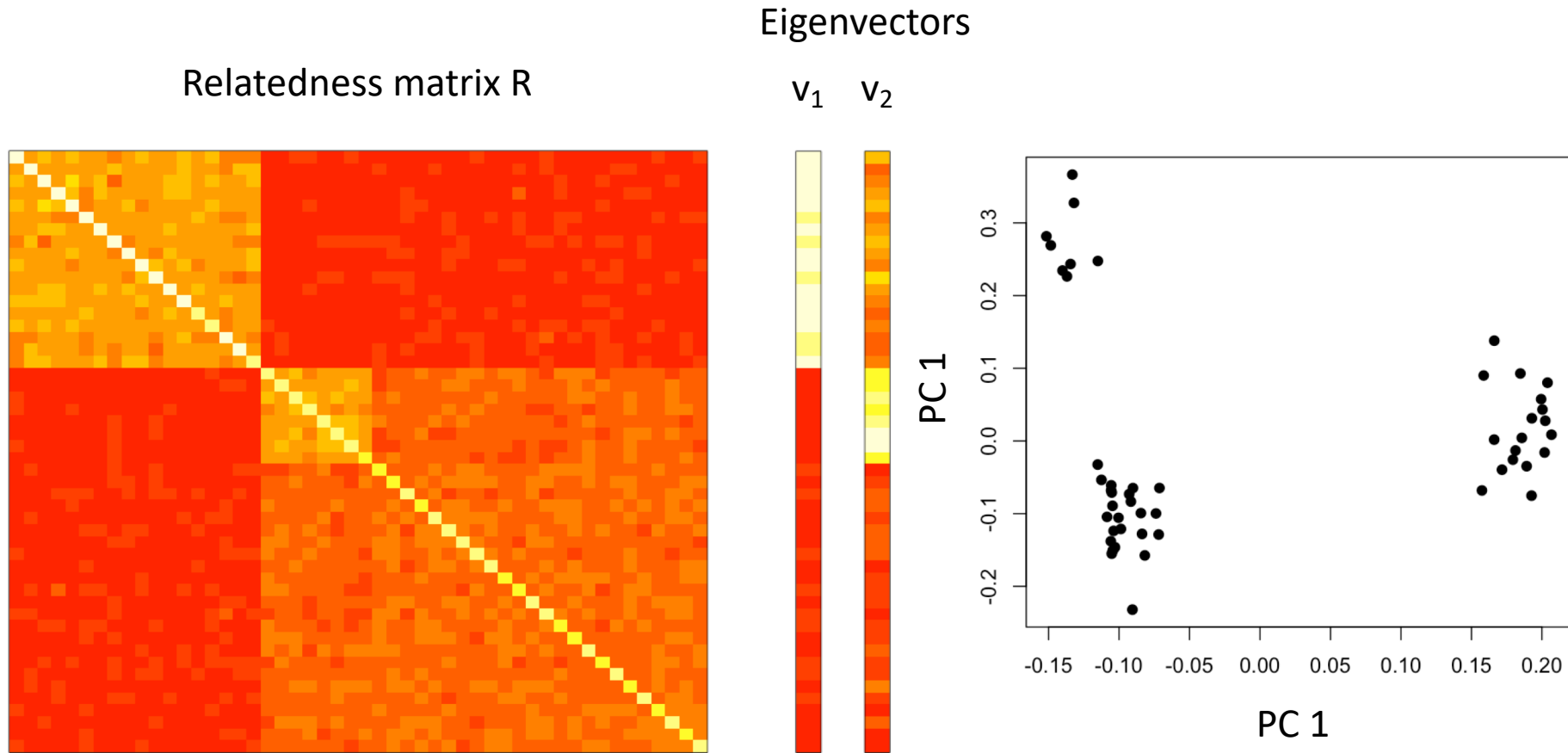(Simulated data, 50 individuals, 1000 SNPs)

Relatedness matrix R

Principal components
Eigen-vectors)

v_1 v_2

Modified from slide by Gavin Band

Principal components analysis

Example
(Simulated data, 50 individuals, 1000 SNPs)

Relatedness matrix R

Eigenvectors

v_1 v_2

Modified from slide by Gavin Band

Principal Component Analysis of Europeans

Within Europe there is very subtle population structure (F_{ST} = 0.004).

The geographic structuring of genetic variation is detectable if >100K SNPs are used.

Novembre et al (2008) *Nature*