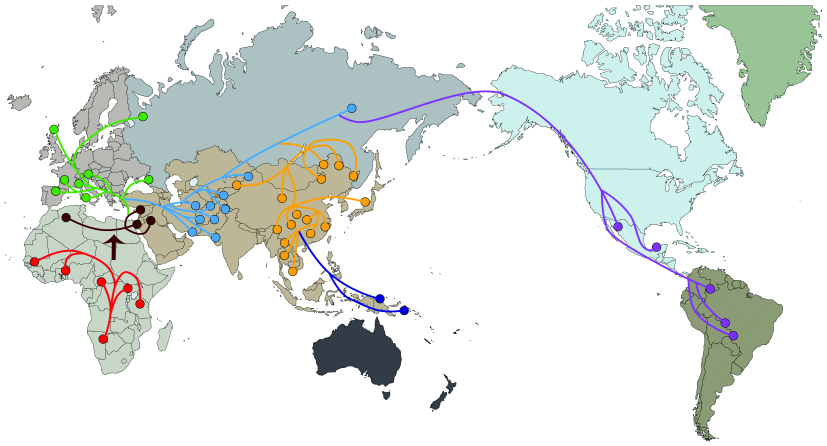
Genetic Ancestry

Hua Tang
April 13, 2011

Outline

- Population history and ancestry
- Estimating genetic ancestry
  - Source of information
  - Statistical reasoning
  - Accuracy
- Admixed populations
  - Estimation of ancestry proportions
  - Estimation of locus-specific ancestry

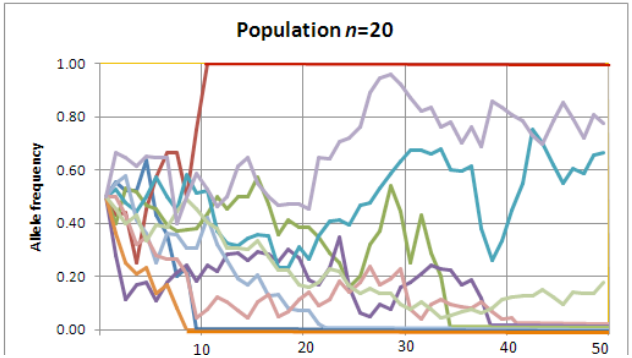
A History of the Human Population



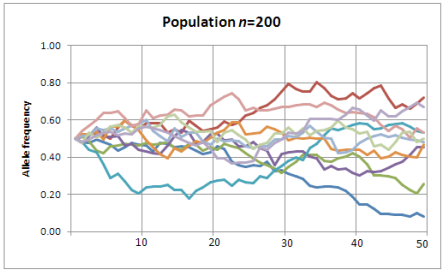
Evolutionary forces

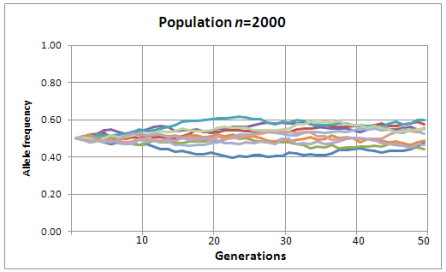
- Mutation
- Genetic drift
- Selection
- Other demographic events…

Genetic Drift

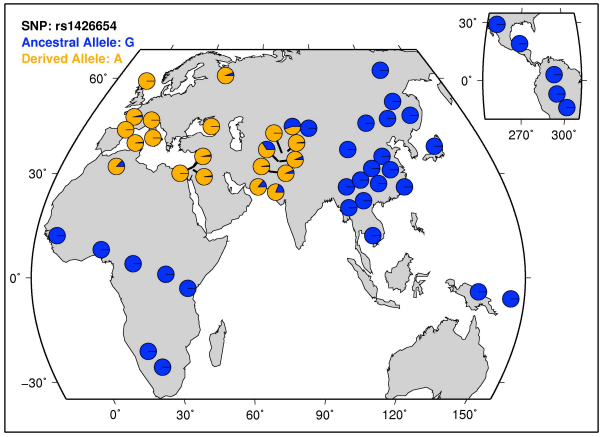


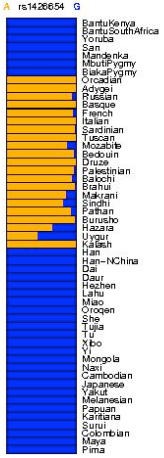
Genetic Drift





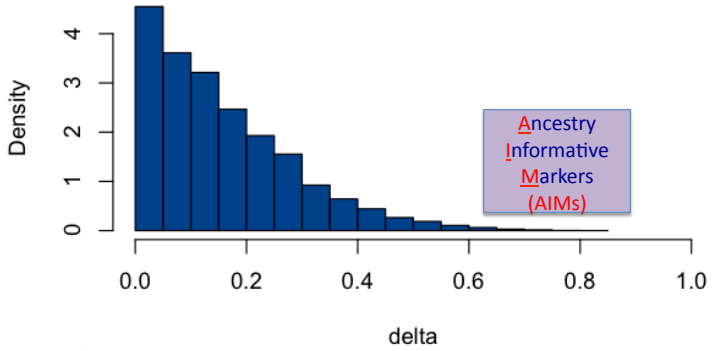
Selection





http://hgdp.uchicago.edu/

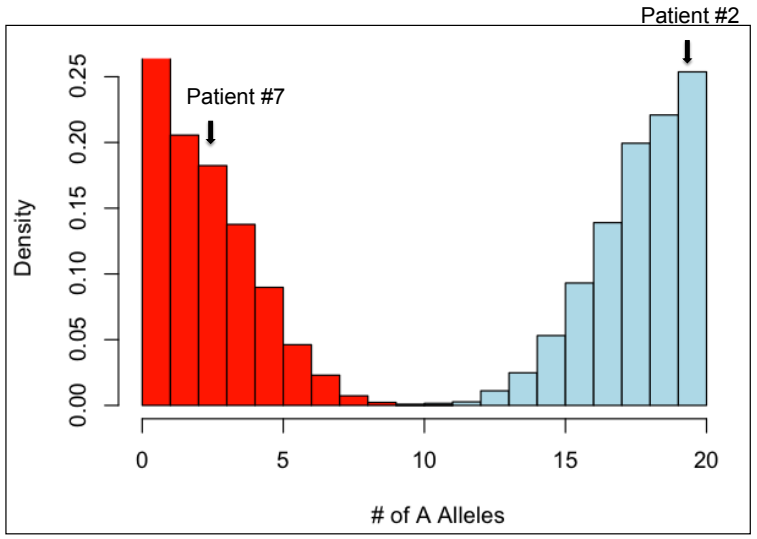
Allele frequency difference (CEU vs EA)



Examples

Rs#	Allele	Freq|CEU	Freq|EA
rs9960403	A	0.94	0.20
rs1834640	A	0.99	0.06

Inferring ancestry: intuition



Limitations

- AIMs
  - are rare, need to screen a large number of markers
  - often informative with regard to specific pairs of populations
  - biased ancestral allele frequencies

Likelihood Principle

What is the ancestry that is most likely to generate genotypes that I have?

This allows one to use all markers, and not just the AIMs.

Can distinguish multiple populations

How accurate?

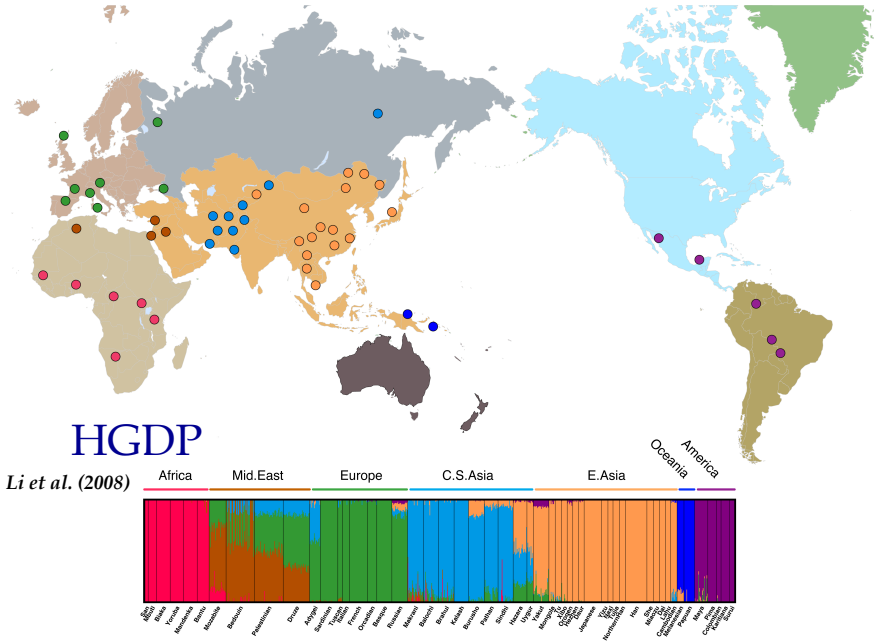
- Data
  - FBPP study of genetic and environmental determinants of hypertension in families
    - Four networks
    - 15 field centers (collection sites),
    - four major race/ethnicity groups: Caucasian, African American East Asian, Hispanic
- Genetic markers used
  - 366 STR markers

Tang et al. (2005)

Genetic Cluster Analysis
4 Clusters

	Cluster A	Cluster B	Cluster C	Cluster D
CAU	1348	0	0	1
AFR	3	0	1305	0
HIS	1	0	0	411
CHI	0	407	0	0
JAP	0	160	0	0

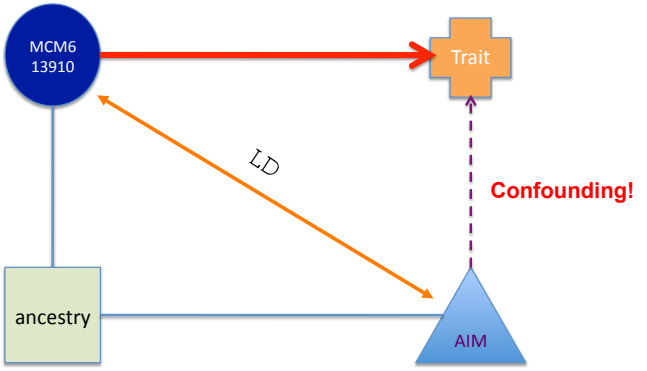
03/12/04



Why should we care about ancestry?

- Population stratification problem in genetic association study
- Ethnic specific risk factors

Population Stratification

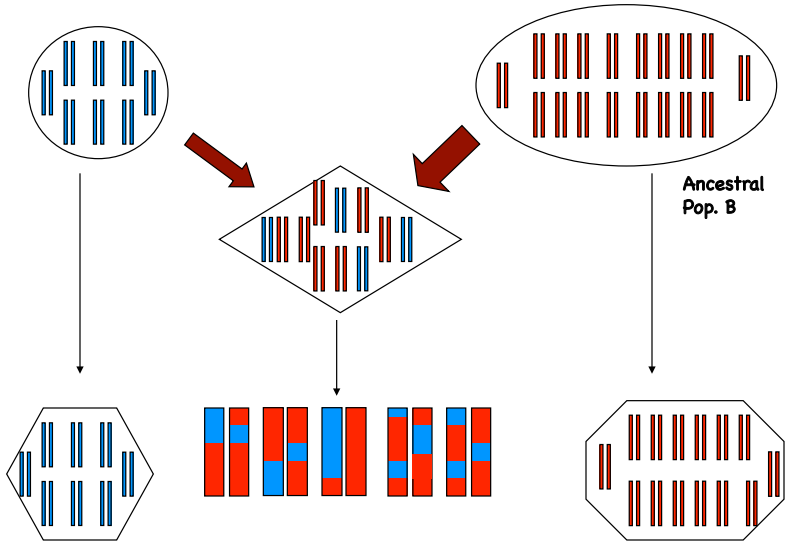


Rs#	Allele	Freq|CEU	Freq|EA
rs9960403	A	0.94	0.20
rs1834640	A	0.99	0.06

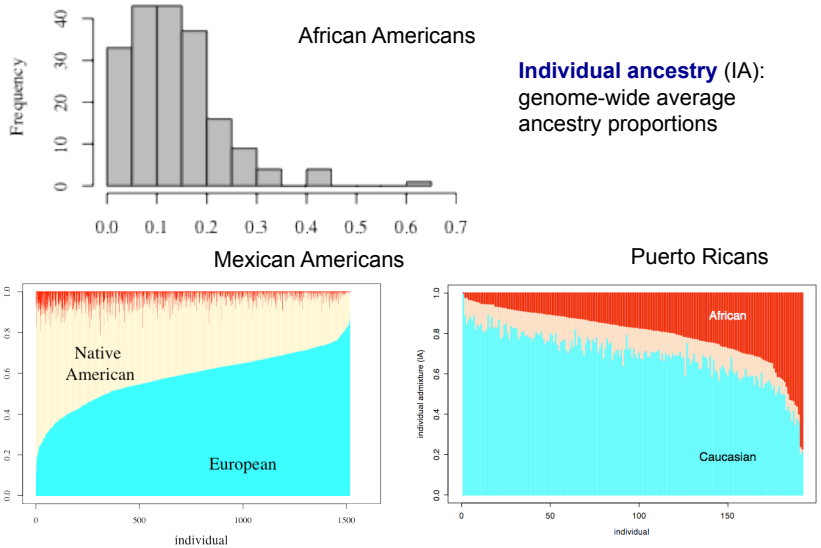
Ethnicity-Specific risk factors

- Disease risk variants can differ in occurrence across populations.
  - European populations contain only a subset of human genetic variation
  - Cardiomyopathy risk variant in MYBPC3
- Risk variants may have different effect sizes
  - HapK in LTA4H for myocardial infarction
- Implication for risk prediction

Recently admixed populations



Variation in IA



Individual ancestry (IA): genome-wide average ancestry proportions

Likelihood approach

- Parameters
  - IA (Q) -- unknown
  - ancestral allele frequencies (P) --known
- Observations: Genotypes at unlinked markers (G)
- Log likelihood function

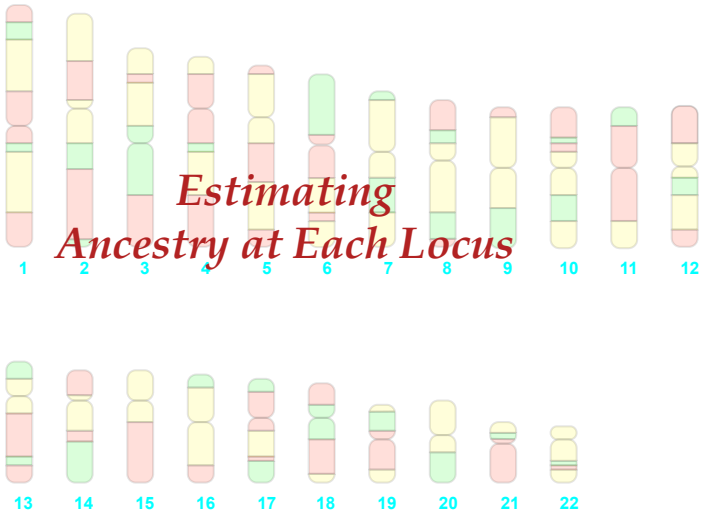
$$\text{lik}(P,Q \mid G) = \sum_i \sum_m \sum_a \sum_l 1_{(G_{i,m,a}=l)} \log r_{i,m,a}$$

where $r_{i,m,l} = P(G_{i,m,a} = l \mid q_i, p_m) = \sum_{k=1}^{K} q_{i,k} p_{m,l,k}$

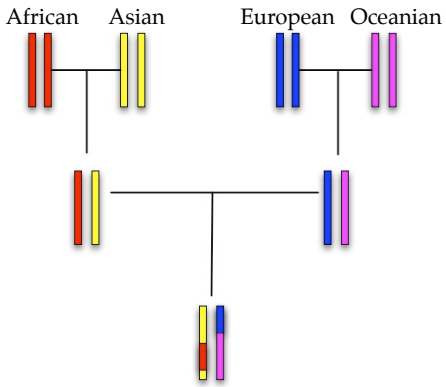
- Reference
  - Hanis et al. 1986
  - Programs from J. Long

Computational Approaches

- STRUCTURE (Pritchard et al. 2001)
  - Bayesian hierarchical clustering approach
  - Priors $p_{mk} \sim D(\lambda_1,\ldots,\lambda_{m_l})$
    $q_i \sim D(\alpha,\ldots,\alpha)$,
  - Missing data: $Z_{i,m,a}$= ancestral origin of allele $G_{i,m,a}$
  - MCMC algorithm
- EM algorithm (frappe)
  - Augmented likelihood function lik($P,Q \mid G,Z$)
  - treating Z as missing data
  - M-Step: update q and p
  - E-step: update E(Z)



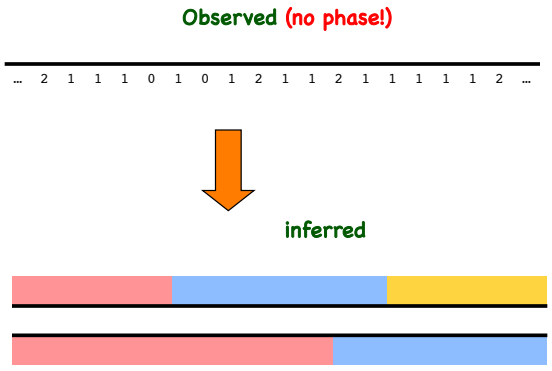
Genealogy of Chromosomes



Can we guess the ancestry origin of each chromosomal segment?

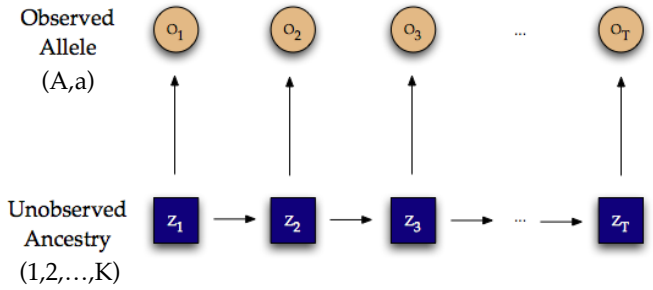
Data Structure

Full information



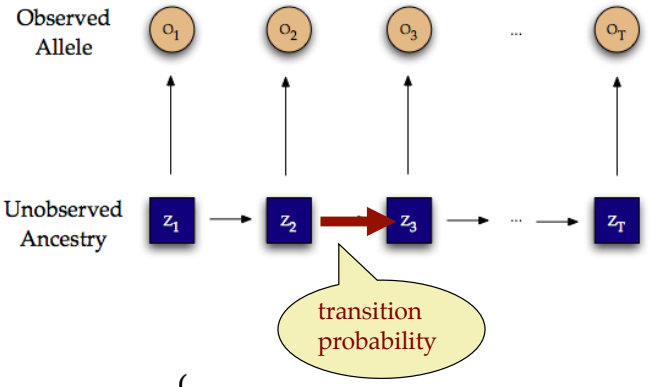
African
European
Native American

Hidden Markov Model



Borrowing strength from neighboring markers

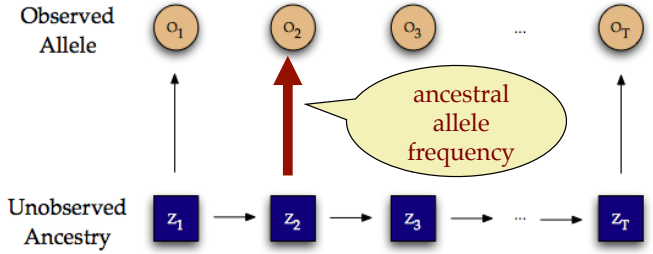
Hidden Markov Model



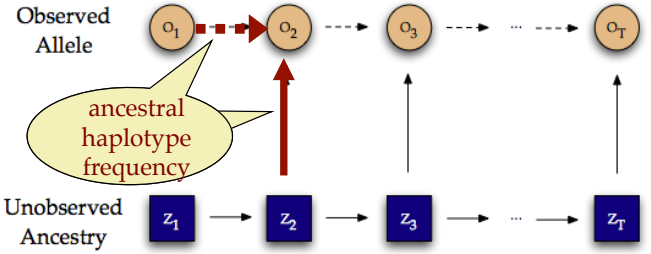
$$\mathbf{P}(Z_{t+1} = j \mid Z_t = i, \tau, \pi) = \begin{cases} \exp(-d_t\tau) + \pi_j(1 - \exp(-d_t\tau)) & i = j \\ \pi_j(1 - \exp(-d_t\tau)) & \text{otherwise} \end{cases}$$

Falush et al. (2003)

Hidden Markov Model



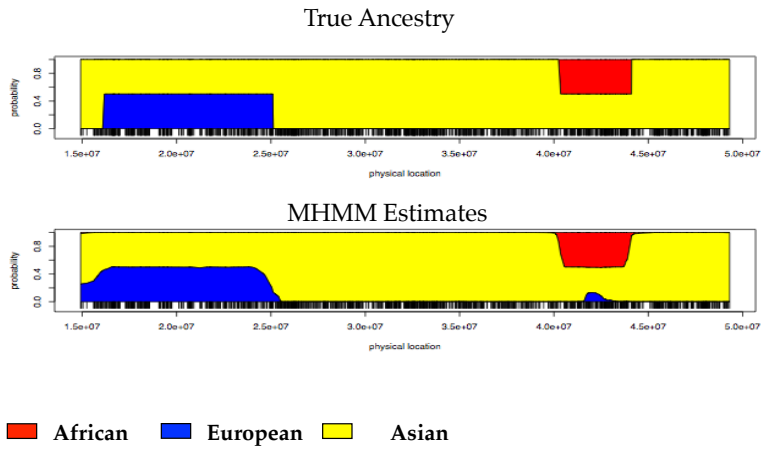
Markov-Hidden Markov Model



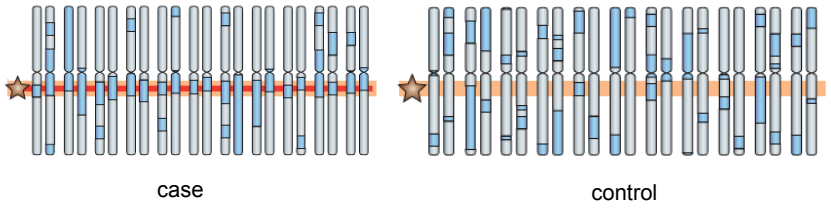
$$P(O_t \mid Z_1,\ldots,Z_t,O_1,\ldots,O_{t-1}) = \begin{cases} P(O_t \mid Z_t) & \text{if } Z_t \neq Z_{t-1} \\ P(O_t \mid Z_t,O_{t-1}) & \text{if } Z_t = Z_{t-1} \end{cases}$$

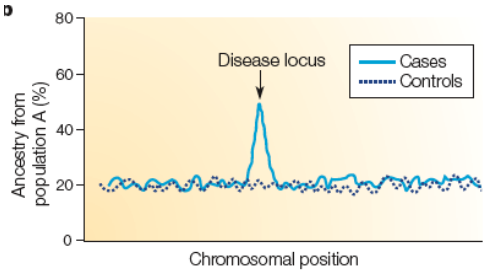
Tang et al. (2006)

Ancestry of an Admixed Individual



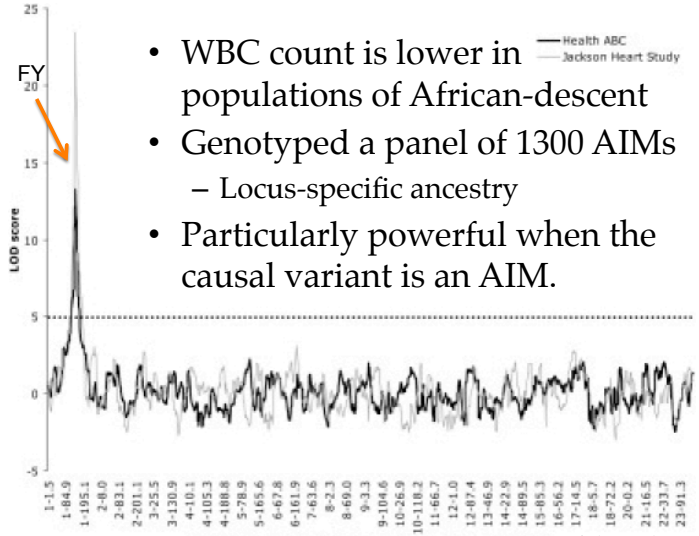
Admixture Mapping





Smith & O'Brien, NRG 2006

Admixture Mapping of WBC



- WBC count is lower in populations of African-descent
- Genotyped a panel of 1300 AIMs
  - Locus-specific ancestry
- Particularly powerful when the causal variant is an AIM.

Reich et al. 2008, AJHG

Summary

- Genetic markers provide information for
  - inferring ancestry of an individual
  - estimating ancestry proportions of admixed individuals
  - estimating ancestral origin of each chromosomal segment
- Ancestry is useful for
  - Mapping locus influencing traits / diseases
  - Assessing individual disease risk
  - Understanding ethnic difference in disease prevalence