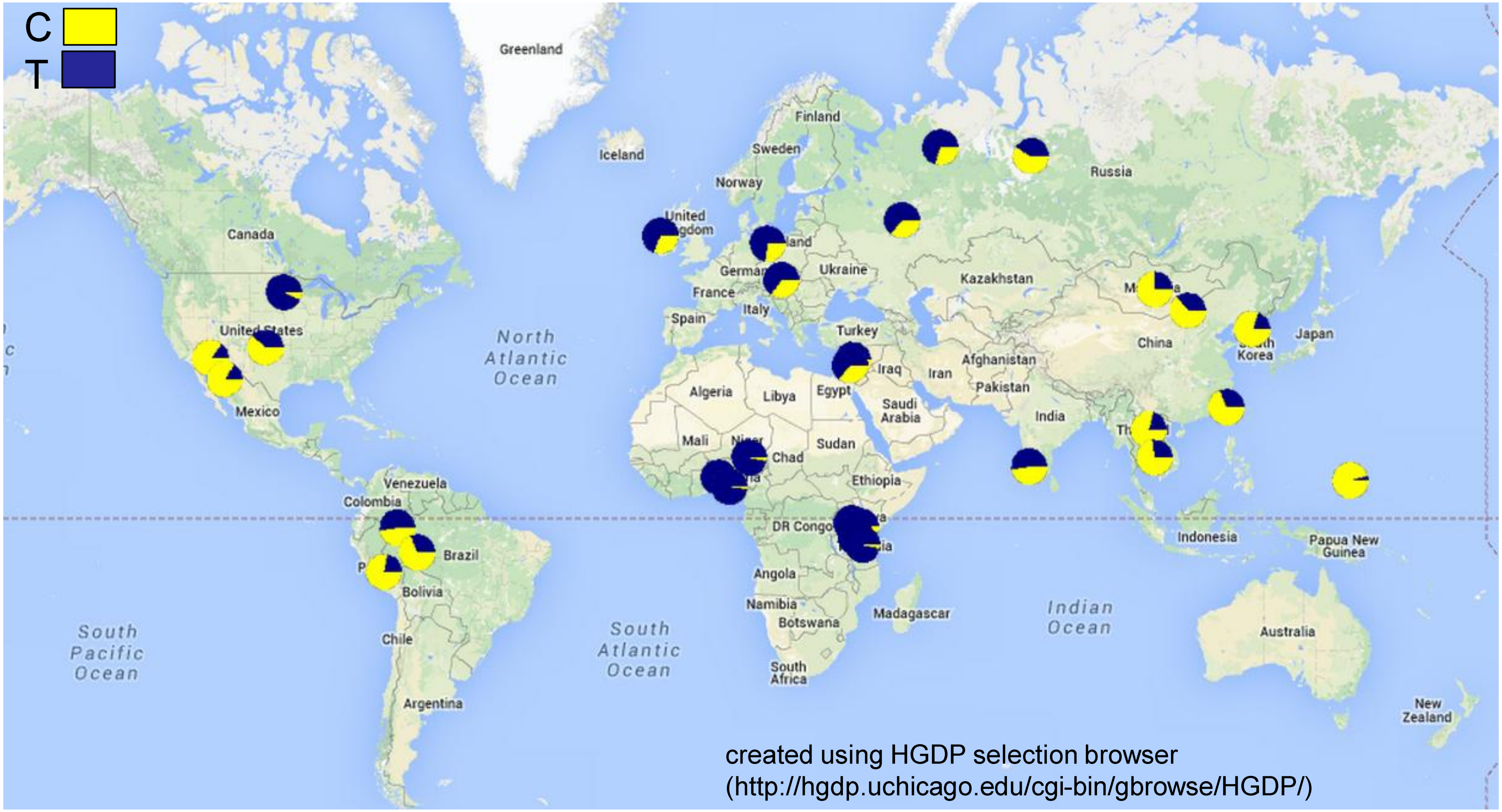
Fast individual ancestry inference from DNA sequence data leveraging allele frequencies from multiple populations*

Vikas Bansal, Ph.D.
Department of Pediatrics
UC San Diego

Bansal and Libiger, BMC Bioinformatics, 2015

Allele frequencies at polymorphic sites differ across human populations

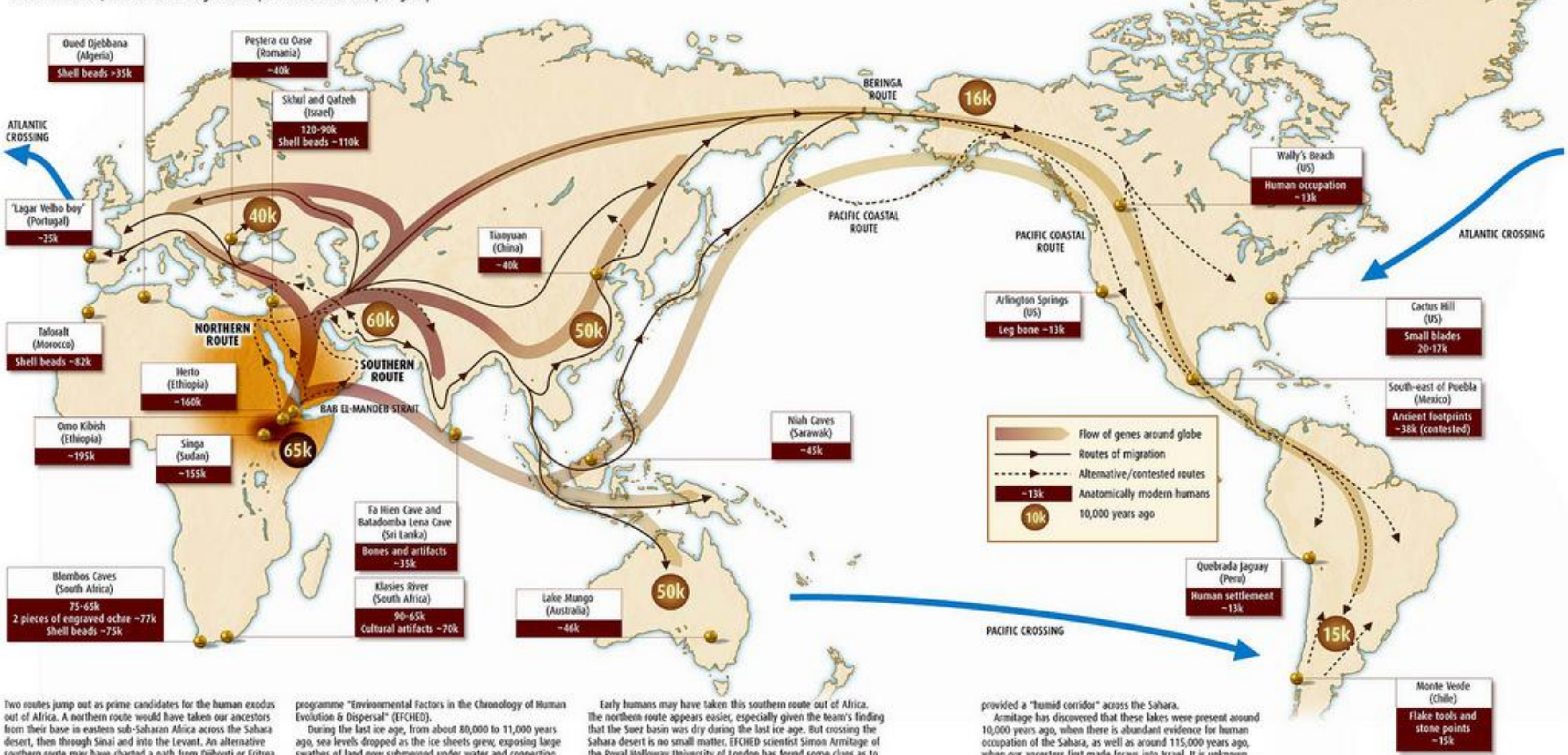
rs3098610



created using HGDP selection browser
(http://hgdp.uchicago.edu/cgi-bin/gbrowse/HGDP/)

THE MIGRATION OF ANATOMICALLY MODERN HUMANS

Evidence from fossils, ancient artefacts and genetic analyses combine to tell a compelling story



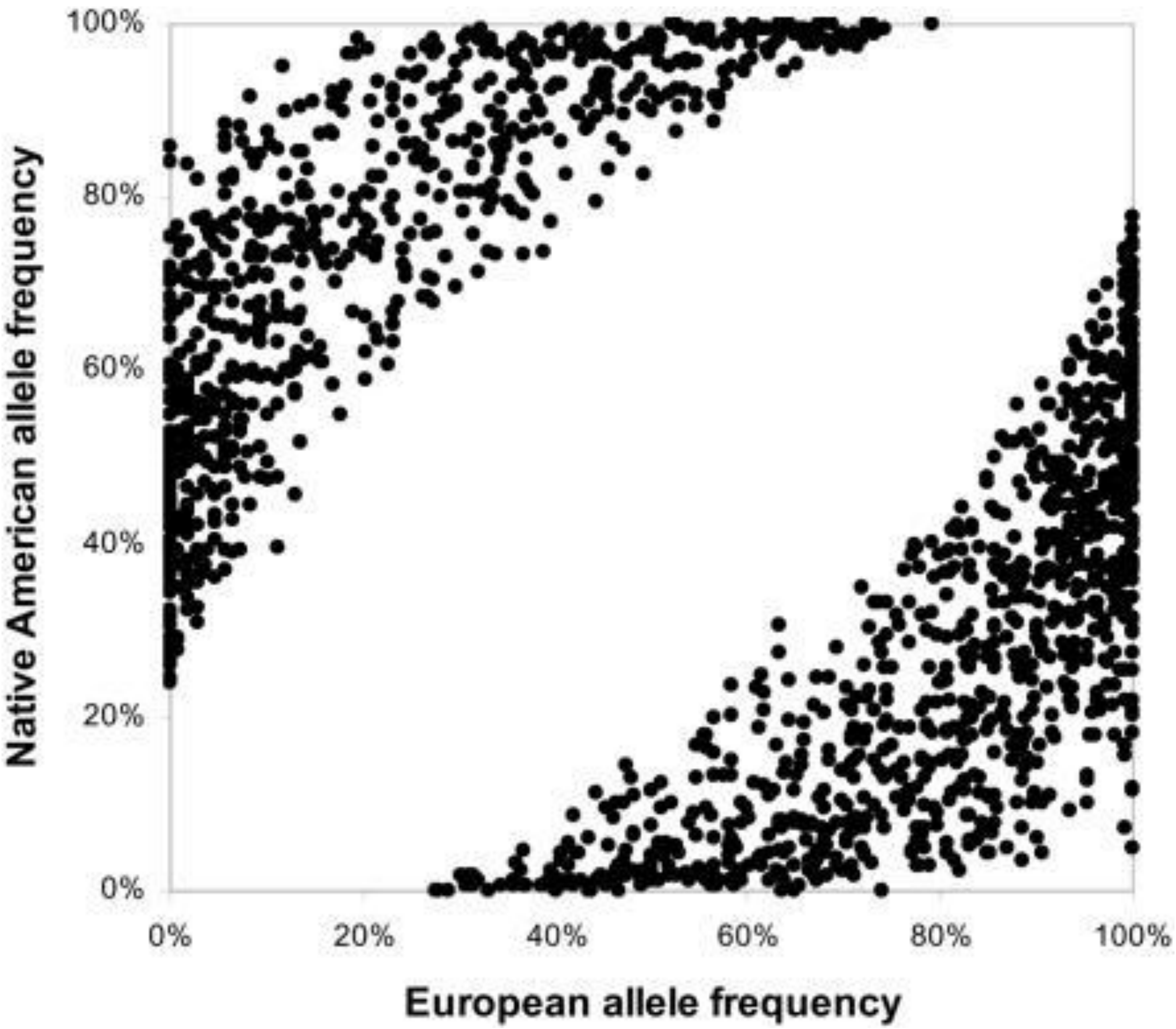
Two routes jump out as prime candidates for the human exodus out of Africa. A northern route would have taken our ancestors from their base in eastern sub-Saharan Africa across the Sahara desert, then through Sinai and into the Levant. An alternative southern route may have charted a path from Djibouti or Eritrea in the Horn of Africa across the Bab el-Mandeb strait and into Yemen and around the Arabian peninsula. The plausibility of these two routes as gateways out of Africa has been studied as part of the UK's Natural Environment Research Council's programme "Environmental Factors in the Chronology of Human Evolution & Dispersal" (EFCHED).

During the last ice age, from about 80,000 to 11,000 years ago, sea levels dropped as the ice sheets grew, exposing large swathes of land now submerged under water and connecting regions now separated by the sea. By reconstructing ancient shorelines, the EFCHED team found that the Bab el-Mandeb strait, now around 30 kilometres wide and one of the world's busiest shipping lanes, was then a narrow, shallow channel.

Early humans may have taken this southern route out of Africa. The northern route appears easier, especially given the team's finding that the Suez basin was dry during the last ice age. But crossing the Sahara desert is no small matter. EFCHED scientist Simon Armitage of the Royal Holloway University of London has found some clues as to how this might have been possible. During the past 150,000 years, North Africa has experienced abrupt switches between dry, arid conditions and a humid climate. During the longer wetter periods huge lakes existed in both Chad and Libya, which would have provided a "humid corridor" across the Sahara.

Armitage has discovered that these lakes were present around 10,000 years ago, when there is abundant evidence for human occupation of the Sahara, as well as around 115,000 years ago, when our ancestors first made forays into Israel. It is unknown whether another humid corridor appeared between about 65,000 and 50,000 years ago, the most likely time frame for the human exodus. Moreover, accumulating evidence is pointing to the southern route as the most likely jumping-off point.

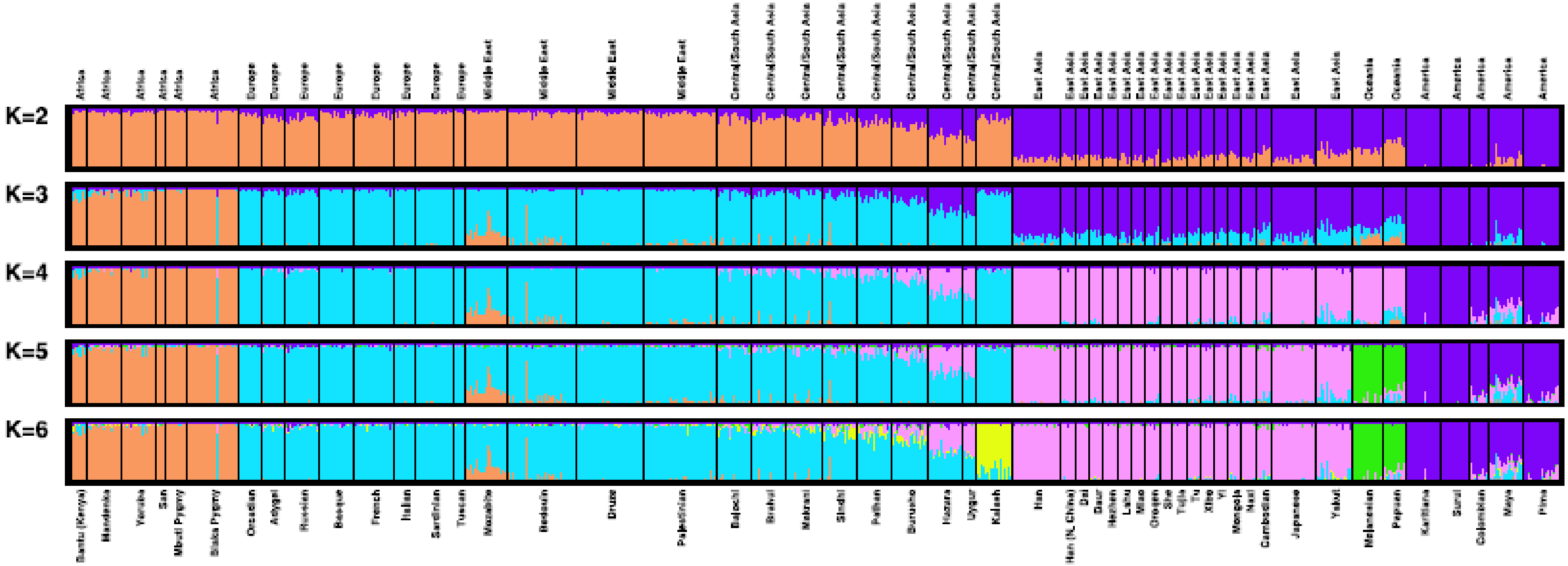
from the New Scientist



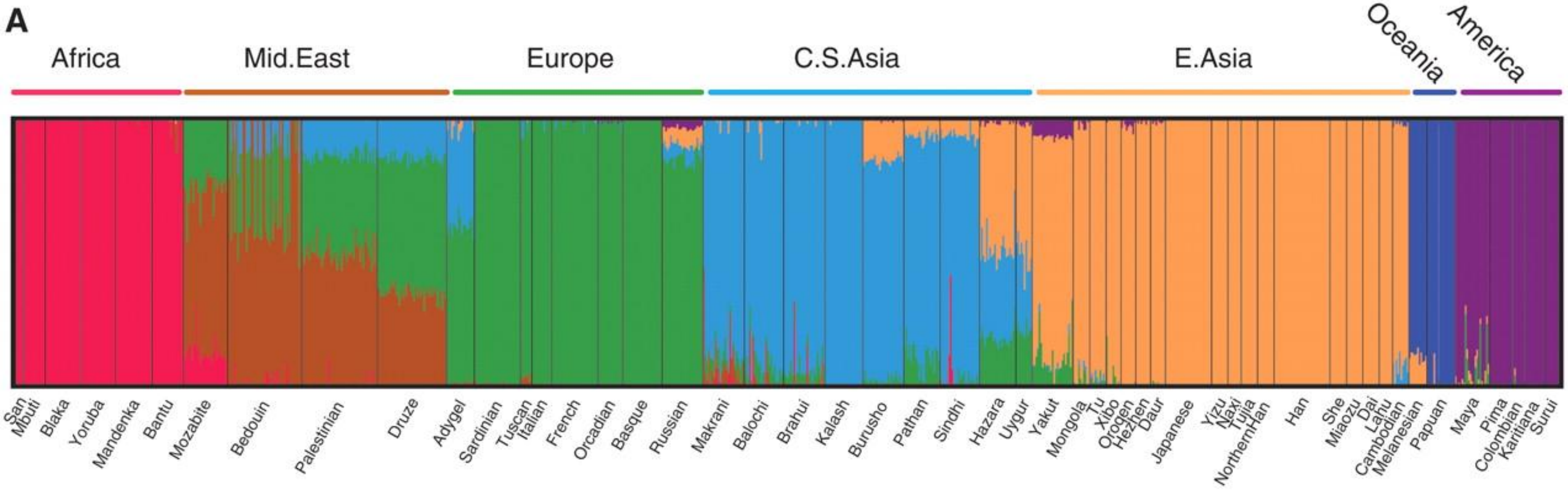
European and Native American allele frequencies for the 1,649 AIMs (Figure 4 from Price et al., AJHG 2007)

Differences in allele frequencies can be used to reconstruct human population structure using genetic data from a number of polymorphic loci

Rosenberg et al. 2002: genotyped 1052 individuals from 52 populations at 377 microsatellites



Fine-scale genetic structure of human populations using 650,000 markers

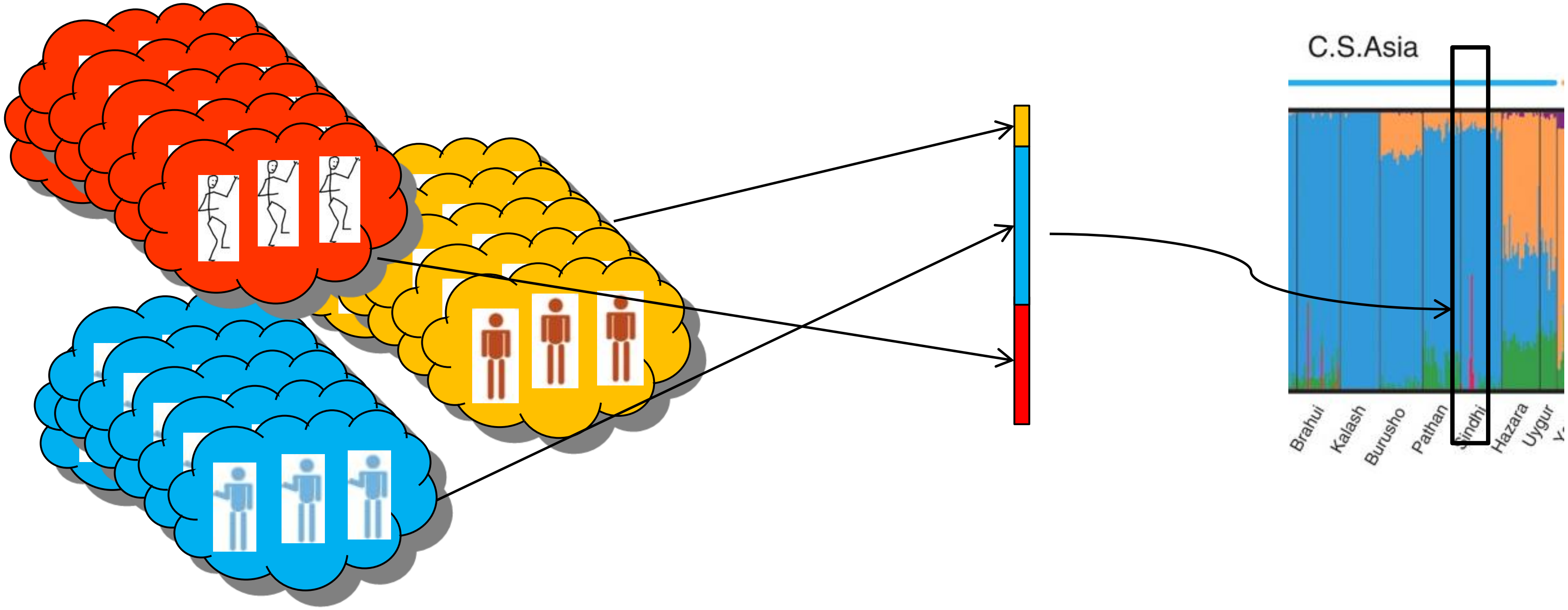


Li et al. Science 2008

The admixture model

Ancestral populations represented as allele frequency profiles

Admixture coefficients for one individual



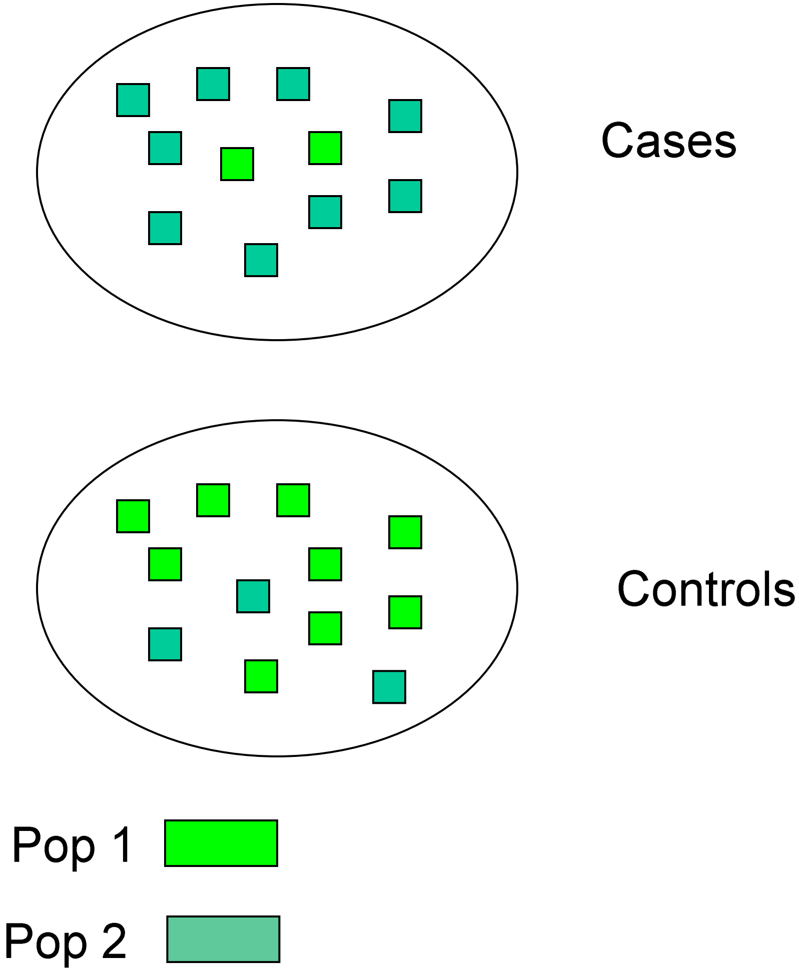
- Allele frequencies for each cluster derived using genotypes of individuals that have non-zero admixture coefficient for that cluster
- Each individual’s admixture estimated using allele frequencies

Methods for reconstructing population structure

- **STRUCTURE (Pritchard et al. 2000) :** Bayesian MCMC method for simultaneous inference of allele frequencies for ‘K’ populations and admixture coefficients for each individual
    - very popular and useful tool
    - Not scalable to genome-wide datasets

- **ADMIXTURE (Alexander et al. 2009):** Maximum likelihood approach to population structure that used fast optimization algorithms for efficiency

Important to control for population stratification in disease association studies



- Cases and controls are sampled from two populations in different proportions
- Loci that differ in allele frequency between the two ancestral populations will show association with the phenotype

Motivation for our work

- Previous methods designed for unsupervised population structure analysis but not for individual ancestry determination
    - Analysis of one new individual requires genotype data for individuals with known ancestry and analysis of all individuals
    - Cannot handle sequence data where genotypes are not known with confidence, e.g. low coverage sequence data

- **Efficient method designed for ancestry estimation for a single individual**
    - Utilizes allele frequency from known populations
    - Works with genotype and sequence data (BAM, VCF files)

Admixture likelihood model

- **INPUT:**
  - Allele frequencies at ‘n’ SNPs for K populations
  - Genotypes or genotype likelihoods for an individual

- **OUTPUT**: A = [a_1, a_2, …. a_K] of admixture proportions such that the sum of admixture proportions = 1

$$L(A) = \sum_{i=1}^{n} \ln(Pr(G_i = g_i|A))$$

Admixture likelihood model (contd..)

$$L(A) = \sum_{i=1}^{n} \ln(Pr(G_i = g_i|A))$$

$$Pr(G_i = 0|A) = (1 - f_i)^2$$

$$Pr(G_i = 1|A) = 2f_i(1 - f_i)$$

$$Pr(G_i = 2|A) = {f_i}^2$$

$$f_i = \sum_{j=1}^{k} q_{ij} a_j$$

- **Non-linear optimization problem with K variables with constraints on the admixture coefficients**

Optimization using the BFGS method

- Broyden-Fletcher-Goldfarb-Shanno algorithm is a quasi-Newton method for unconstrained non-linear optimization
- Uses first derivatives and approximation of Hessian matrix
- **Features**
  - Good performance even for 1000's of variables
  - BFGS-B variant handles box constraints on variables
  - Several open-source implementations are available

- Constraint on sum can be addressed by replacing a_j with $\frac{a_j}{\sum_k a_k}$
- First derivates can be easily calculated as:

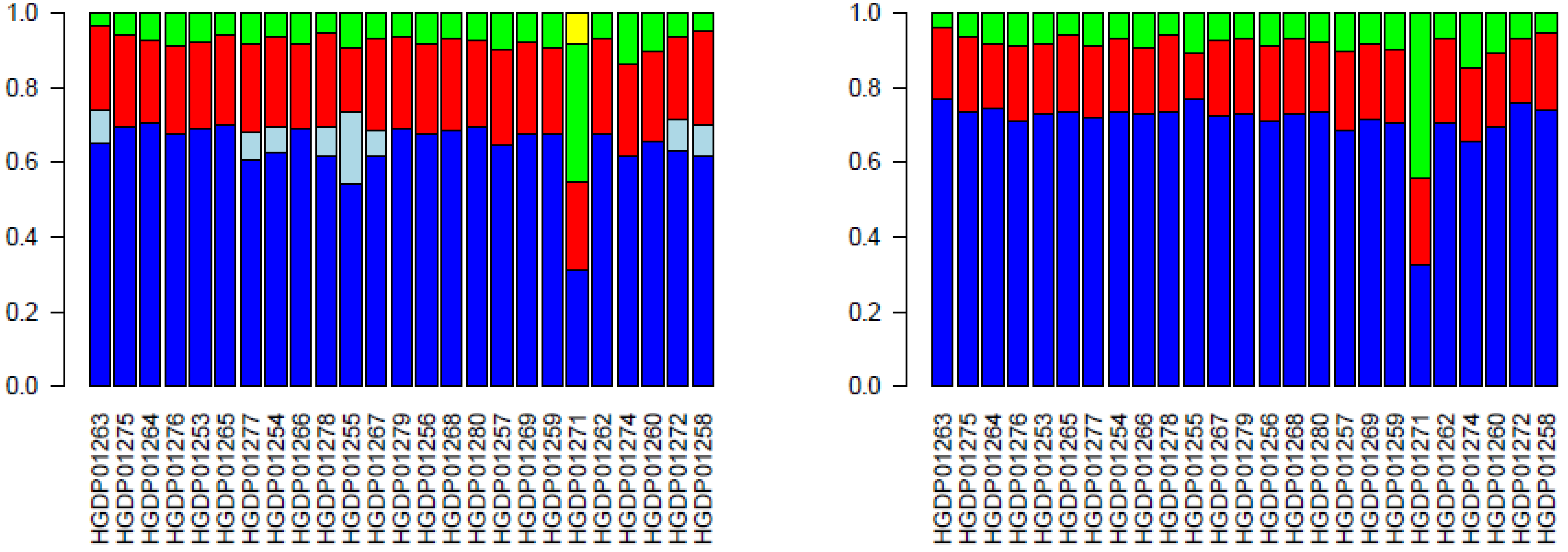
$$\frac{\partial L(A)}{\partial a_j} = \sum_{i=1}^{n} \left[\frac{g_i q_{ij}}{f_i} - \frac{g_i}{S(a)} + \frac{(2-g_i)(1-q_{ij})}{S(a)-f_i} - \frac{2-g_i}{S(a)} \right]$$

Parsimonious estimation of admixture coefficients

- BFGS optimization finds maximum likelihood estimate of admixture coefficients
- Useful to determine if a non-zero admixture coefficient is statistically significant
  - Previous methods estimate confidence intervals using bootstrap

- **Backward elimination method for variable selection to obtain parsimonious vector of admixture coefficients**
  1. Find population ‘j’ for which setting a_j to 0 reduces the likelihood value the least
  2. Fix a_j = 0 and iterate until possible

Analysis of Mozabite individuals from HGDP



(a)

(b)

Figure 1: Admixture proportions for 25 Mozabite individuals estimated using the HapMap reference populations and using two methods: iAdmix (a) and ADMIXTURE(b). The population labels are as follows: TSI (blue), CEU (light blue), MKK (red), YRI (green) and LWK (yellow).

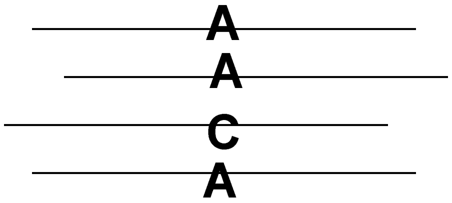
Implementation and running time

- Fortran implementation of L-BFGS-B algorithm (Zhu et al. 1997) was converted to C
- Computational complexity is linear in number of SNPs and number of reference populations
- Number of iterations for convergence (delta < 0.0001) was typically 20-30
- Our method was 10-15 times faster than ADMIXTURE (run in supervised mode) on simulated and real datasets

Estimating ancestry from sequence reads

Likelihood model for sequence reads

- Uncertainty in genotype calls derived from sequence reads
- Genotype likelihoods **Pr(reads | genotype)** capture uncertainty



e = 0.01

Pr($\mathbf{R}_\mathbf{i}$ | g = AA) = 0.0097

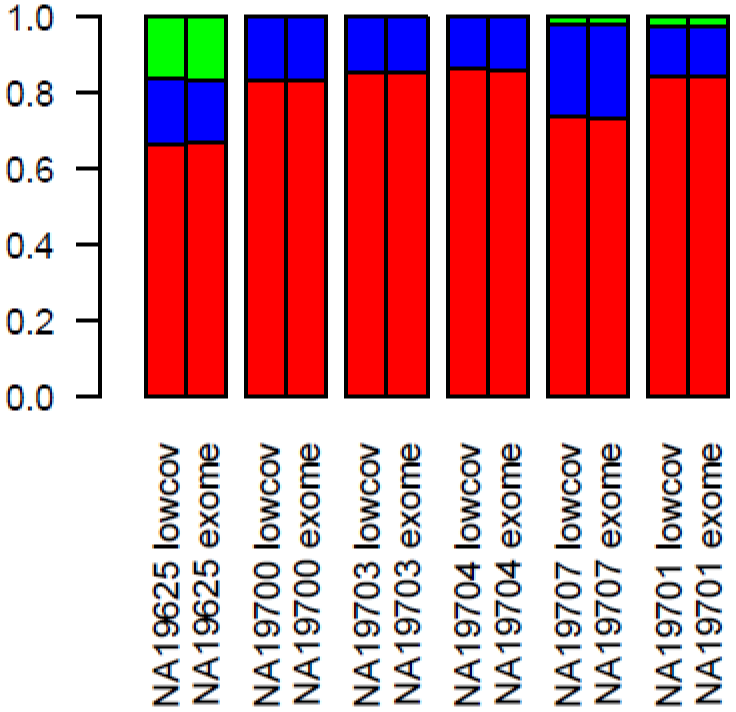
Pr($\mathbf{R}_\mathbf{i}$ | g = AC) = 0.0625

Pr($\mathbf{R}_\mathbf{i}$ | g = CC) = 0.99 x 10-6

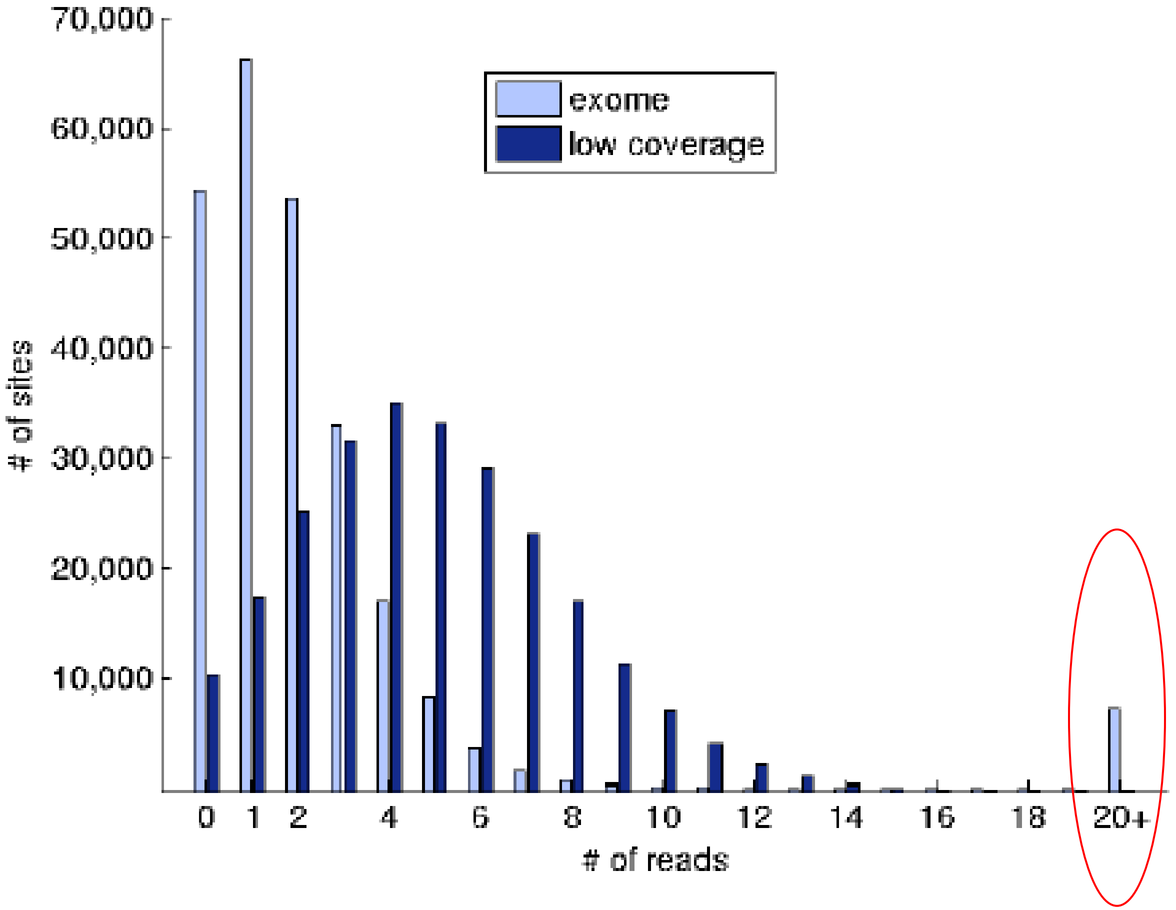
$$L(A) = \sum_{i=1}^{n} \ln \left[\sum_{g=0}^{2} Pr(\mathcal{R}_i | G_i = g) Pr(G_i = g | A) \right]$$

Analysis of 1000 Genomes data

- Analyzed sequence reads for 6 individuals from the ASW (African-Americans in USA) population
- BAM files for low-coverage whole genome sequencing and exome-sequencing available
- genotype likelihoods calculated for 249,075 SNPs that overlap the HapMap allele frequency data

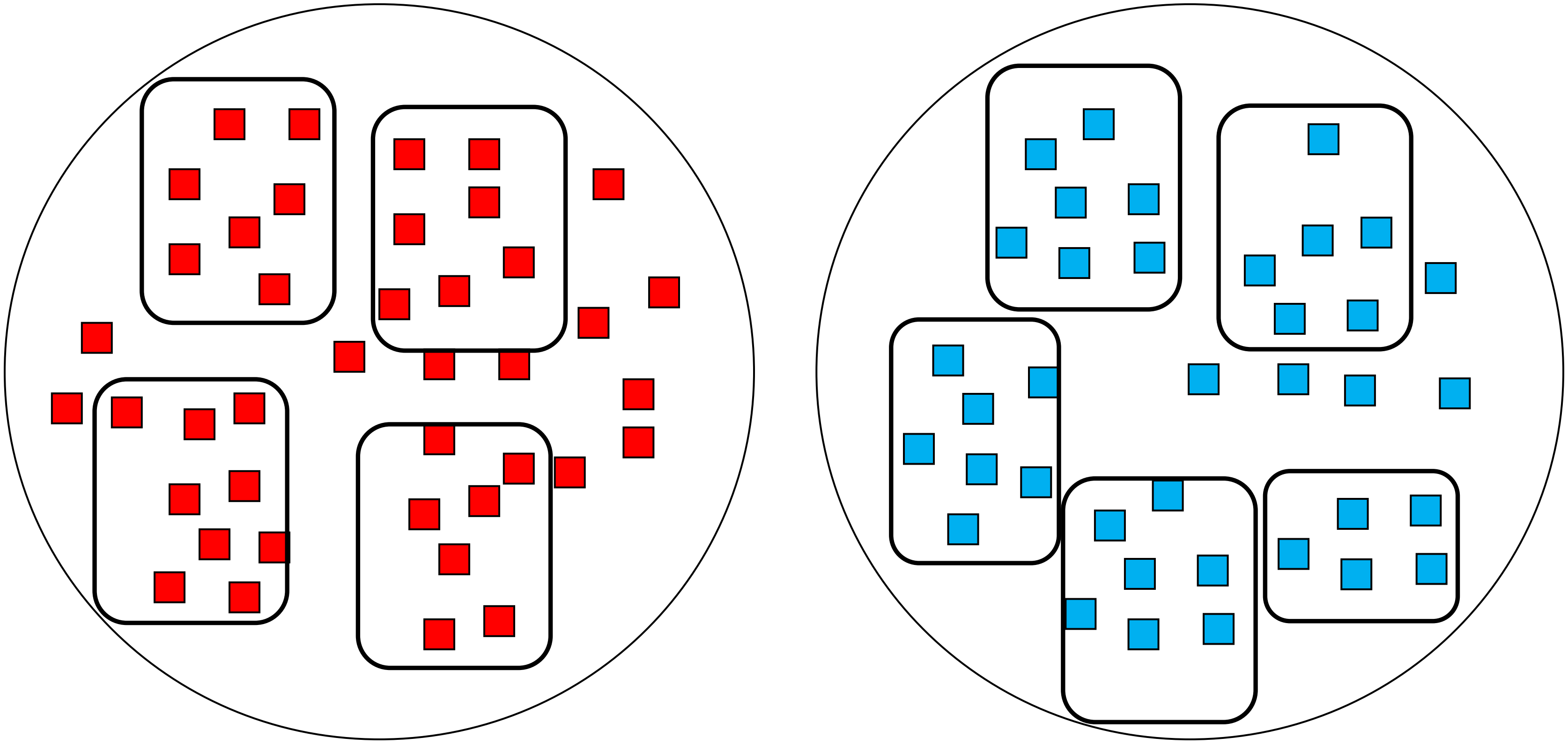


Distribution of read-depth for HapMap3 sites



High-coverage
Coding SNPs

Estimating ancestry of ‘artificial’ DNA pools

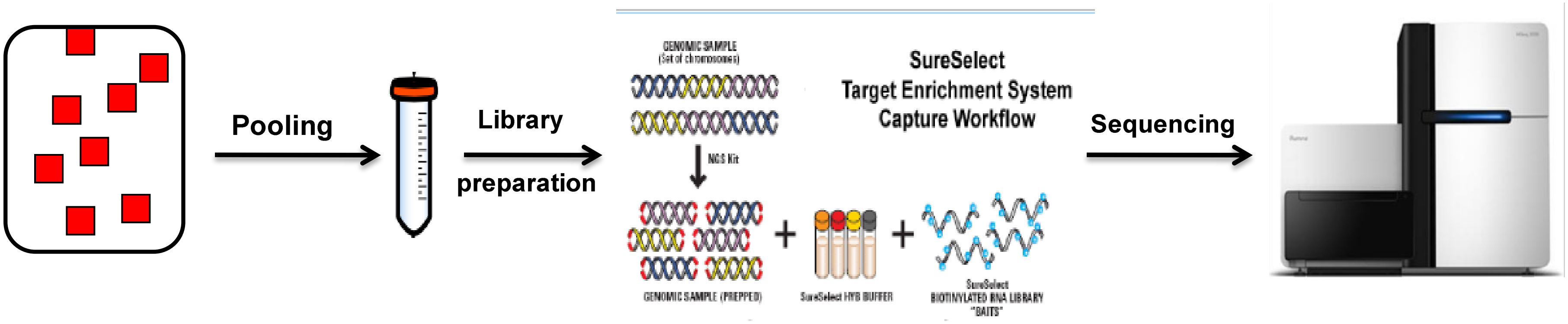


Cases

Controls

- DNA from multiple individuals (20-30) pooled to form a single sample before sequencing

Cost-effective association studies using DNA pooling



Pooling

Library preparation

Sequencing

- 2000 individuals -> 100 pools of size 20
- Targeted sequencing of coding sequence of 200-250 genes can be done for $60,000
- Individual sequencing would cost ~ $300,000

Ancestry estimation from pools: simulated data

- Extended genotype likelihood calculation for ‘pooled’ genotypes
- Evaluated ability to detect admixture in a single pool using 1000 Genomes data

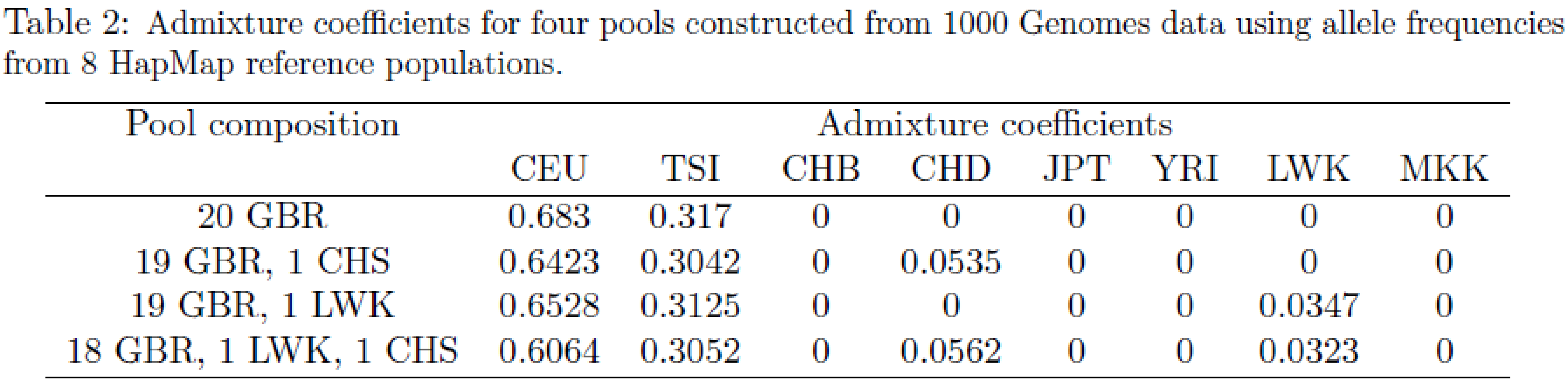


Table 2: Admixture coefficients for four pools constructed from 1000 Genomes data using allele frequencies from 8 HapMap reference populations.

Pool composition	Admixture coefficients							
	CEU	TSI	CHB	CHD	JPT	YRI	LWK	MKK
20 GBR	0.683	0.317	0	0	0	0	0	0
19 GBR, 1 CHS	0.6423	0.3042	0	0.0535	0	0	0	0
19 GBR, 1 LWK	0.6528	0.3125	0	0	0	0	0.0347	0
18 GBR, 1 LWK, 1 CHS	0.6064	0.3052	0	0.0562	0	0	0.0323	0

Summary

- Fast method for estimation of admixture coefficients from genotype or sequence data using allele frequencies
  - 10-15 times faster than previous methods
  - Ancestry can be analyzed using even targeted sequence data
  - Valuable tool for sequencing based studies

- Admixture likelihood model ignores linkage disequilibrium (LD) between markers
  - Haplotype-based likelihood model using haplotype frequencies and dynamic programming
  - BFGS algorithm can be used for optimization

- BFGS algorithm is a fast method for constrained high-dimensional non-linear optimization
  - Useful for many problems, e.g. genome scaffolding, logistic regression for low-coverage sequence data, variant calling