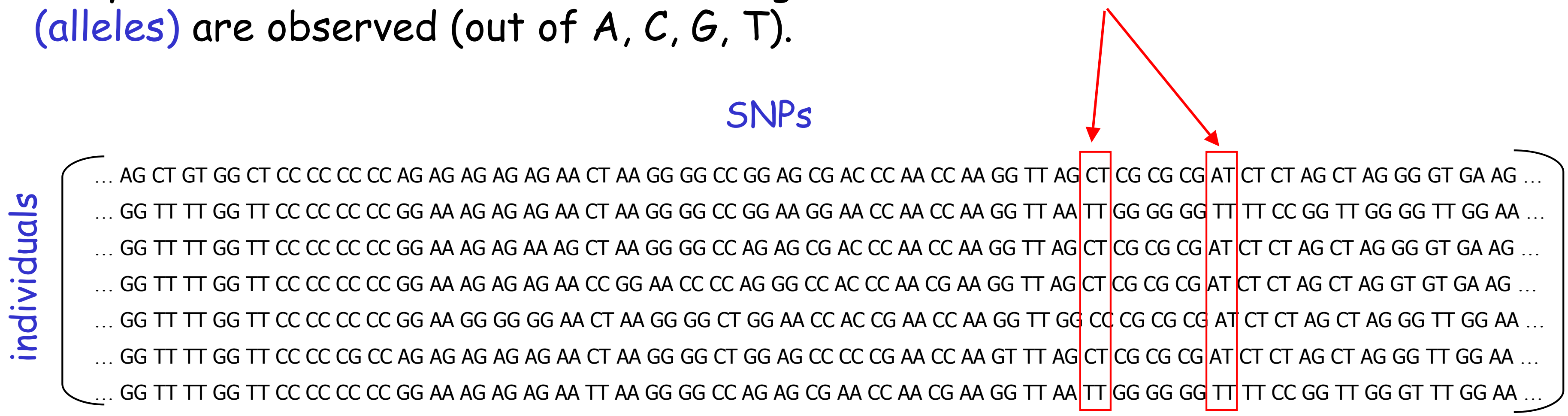
PCA: An example in human genetics

Single Nucleotide Polymorphisms: the most common type of genetic variation in the genome across different individuals.

They are known locations at the human genome where two alternate nucleotide bases (alleles) are observed (out of A, C, G, T).



… AG CT GT GG CT CC CC CC CC AG AG AG AG AG AA CT AA GG GG CC GG AG CG AC CC AA CC AA GG TT AG CT CG CG CG AT CT CT AG CT AG GG GT GA AG …

… GG TT TT GG TT CC CC CC CC GG AA AG AG AG AA CT AA GG GG CC GG AA GG AA CC AA CC AA GG TT AA TT GG GG GG TT TT CC GG TT GG GG TT GG AA …

… GG TT TT GG TT CC CC CC CC GG AA AG AG AA AG CT AA GG GG CC AG AG CG AC CC AA CC AA GG TT AG CT CG CG CG AT CT CT AG CT AG GG GT GA AG …

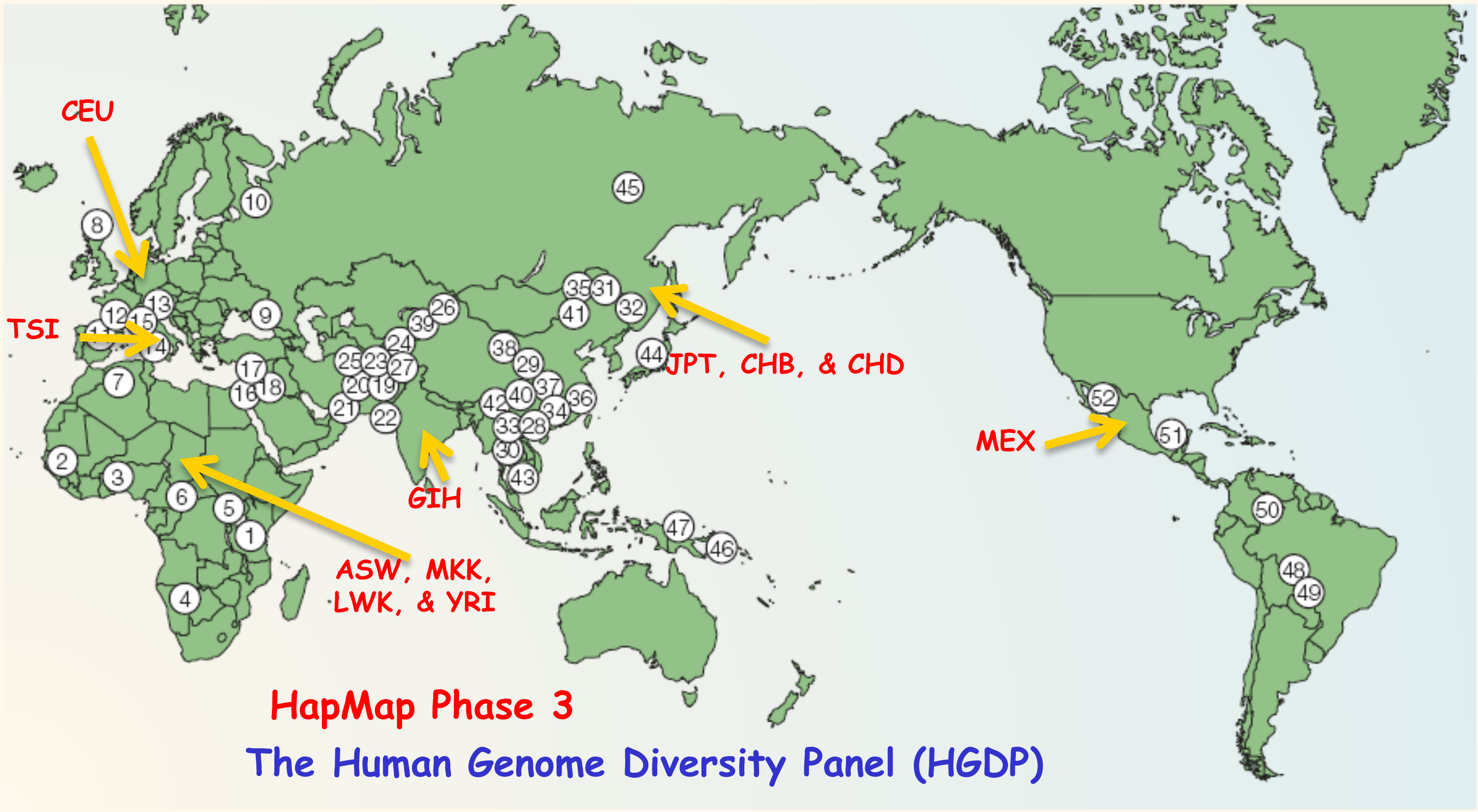
… GG TT TT GG TT CC CC CC CC GG AA AG AG AG AA CC GG AA CC CC AG GG CC AC CC AA CG AA GG TT AG CT CG CG CG AT CT CT AG CT AG GT GT GA AG …

… GG TT TT GG TT CC CC CC CC GG AA GG GG GG AA CT AA GG GG CT GG AA CC AC CG AA CC AA GG TT GG CC CG CG CG AT CT CT AG CT AG GG TT GG AA …

… GG TT TT GG TT CC CC CG CC AG AG AG AG AG AA CT AA GG GG CT GG AG CC CC CG AA CC AA GT TT AG CT CG CG CG AT CT CT AG CT AG GG TT GG AA …

… GG TT TT GG TT CC CC CC CC GG AA AG AG AG AA TT AA GG GG CC AG AG CG AA CC AA CG AA GG TT AA TT GG GG GG TT TT CC GG TT GG GT TT GG AA …

Typical sizes: tens of thousands of individuals and hundreds of thousands of SNPs.



HapMap Phase 3

The Human Genome Diversity Panel (HGDP)

Africans
- 1 Bantu
- 2 Mandenka
- 3 Yoruba
- 4 San
- 5 Mbuti pygmy
- 6 Biaka
- 7 Mozabite

Europeans
- 8 Orcadian
- 9 Adygei
- 10 Russian
- 11 Basque
- 12 French
- 13 North Italian
- 14 Sardinian
- 15 Tuscan

Western Asians
- 16 Bedouin
- 17 Druze
- 18 Palestinian

Central and Southern Asians
- 19 Balochi
- 20 Brahui
- 21 Makrani
- 22 Sindhi
- 23 Pathan
- 24 Burusho
- 25 Hazara
- 26 Uygur
- 27 Kalash

Eastern Asians
- 28 Han (S. China)
- 29 Han (N. China)
- 30 Dai
- 31 Daur
- 32 Hezhen
- 33 Lahu
- 34 Miao
- 35 Oroqen
- 36 She
- 37 Tujia
- 38 Tu
- 39 Xibo
- 40 Yi
- 41 Mongola
- 42 Naxi
- 43 Cambodian
- 44 Japanese
- 45 Yakut

Oceanians
- 46 Melanesian
- 47 Papuan

Native Americans
- 48 Karitiana
- 49 Surui
- 50 Colombian
- 51 Maya
- 52 Pima

Cavalli-Sforza (2005) *Nat Genet Rev*

Rosenberg et al. (2002) *Science*

Li et al. (2008) *Science*

The International HapMap Consortium (2003, 2005, 2007), *Nature*

HGDP data
- 1,033 samples
- 7 geographic regions
- 52 populations

HapMap Phase 3 data
- 1,207 samples
- 11 populations

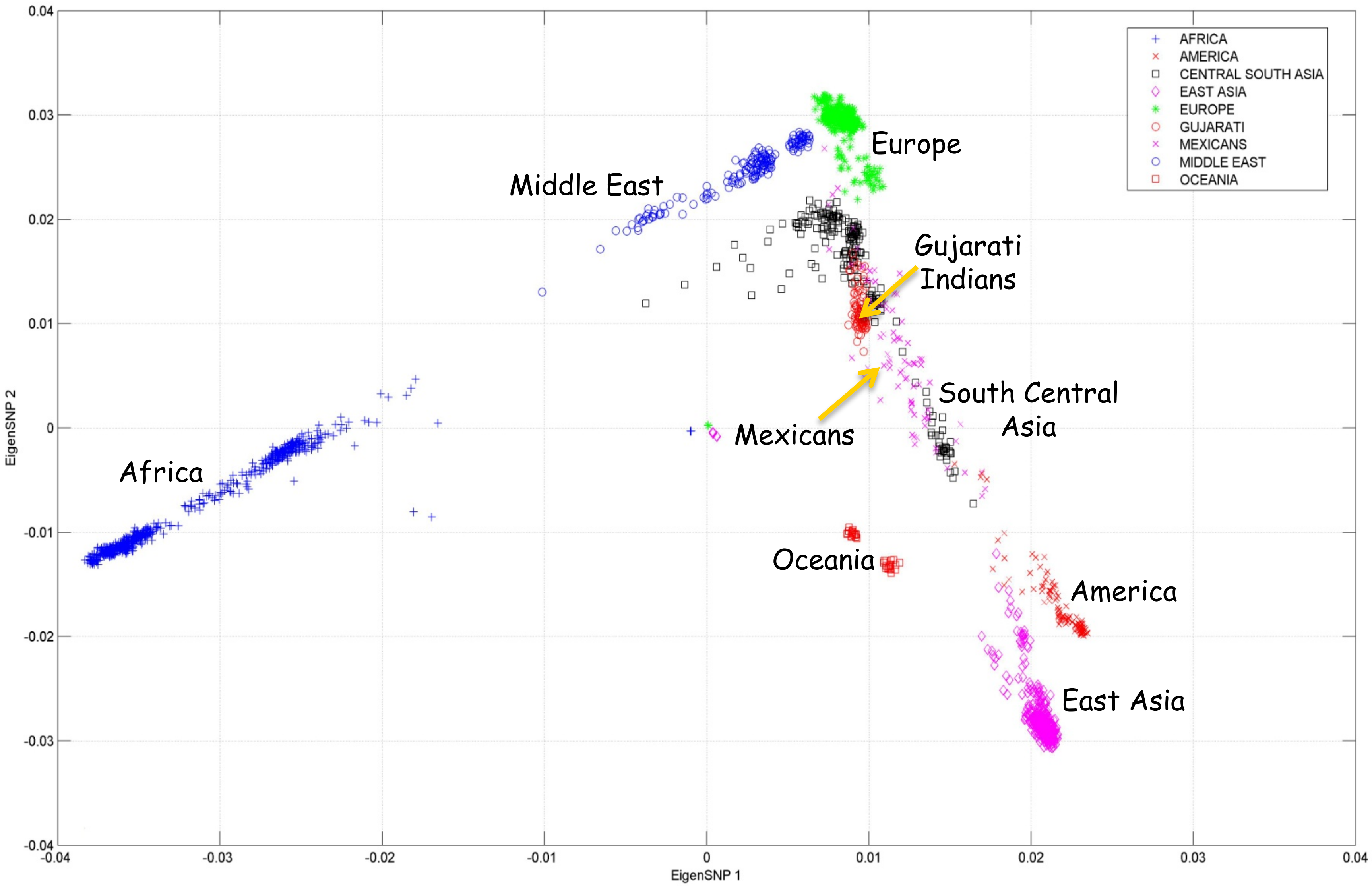
Matrix dimensions:

2,240 subjects (rows)

447,143 SNPs (columns)

We will apply PCA (i.e., SVD on a suitably rescaled covariance matrix) to visualize and/or analyze the data.

Paschou, Lewis, Javed, & Drineas (2010) J Med Genet

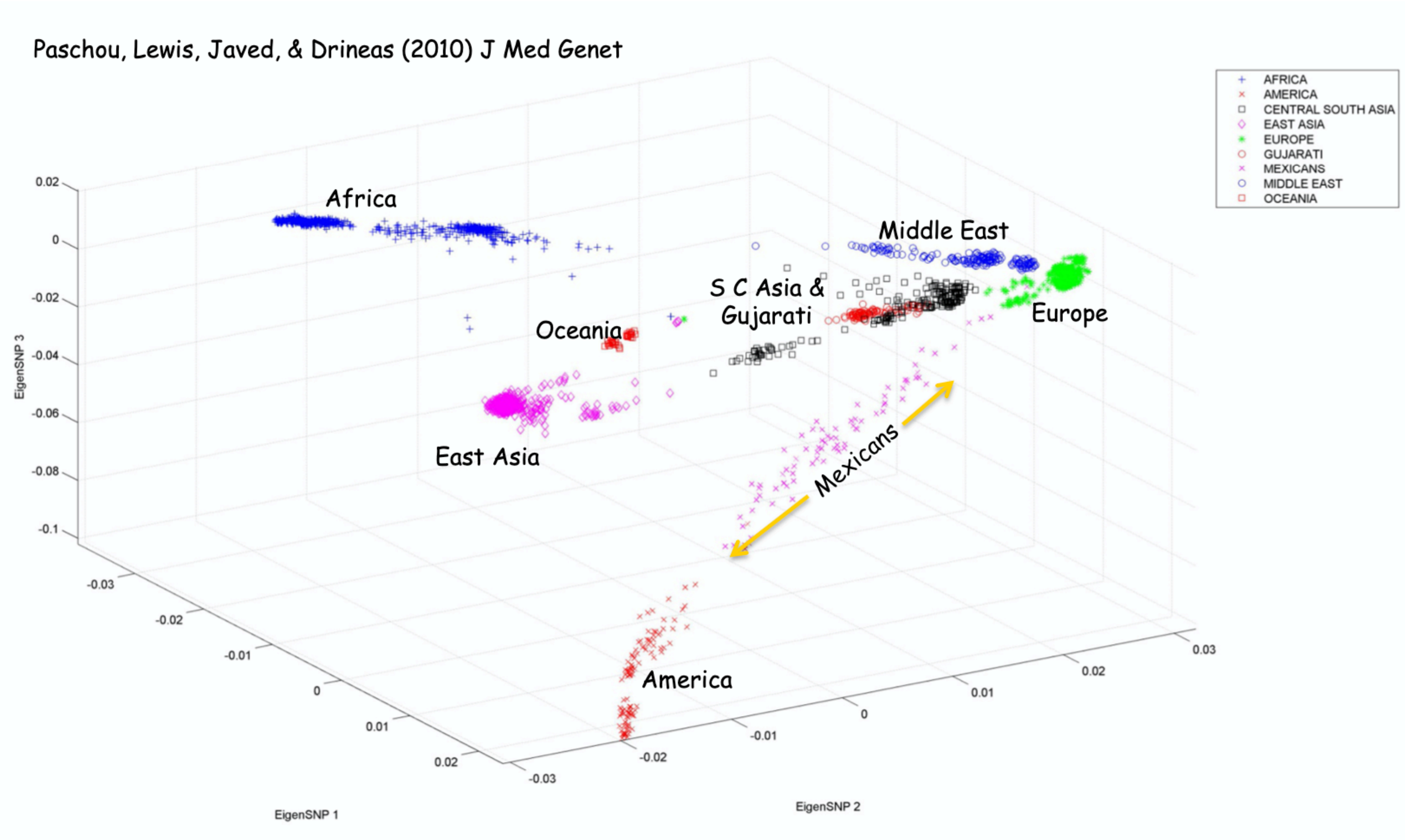


- <u>Top two Principal Components</u> (PCs or eigenSNPs)

(Lin and Altman (2005) *Am J Hum Genet*)

- Very good correlation between geography and the top two eigenSNPs.
- Mexican population seems out of place: we move to the top three PCs.

Slides courtesy of Petros Drineas, Purdue University



Paschou, Lewis, Javed, & Drineas (2010) J Med Genet