

# Genetic variation and recent positive selection in worldwide human populations: evidence from nearly 1 million SNPs.

López Herráez D, Bauchet M, Tang K, Theunert C, Pugach I, Li J, Nandineni R, Gross A, Scholz M, Stoneking M. PLoS One. 2009 Nov 18;4(11):e7888.

Jclub in bioinformatics

08.12.2009

Tarmo Puurand

# Overview of article

- Worldwide genetic variation and structure
- Regional population structure and relationships
- Recent local selection
- GO enrichment analysis
- Novel candidates in African pygmies

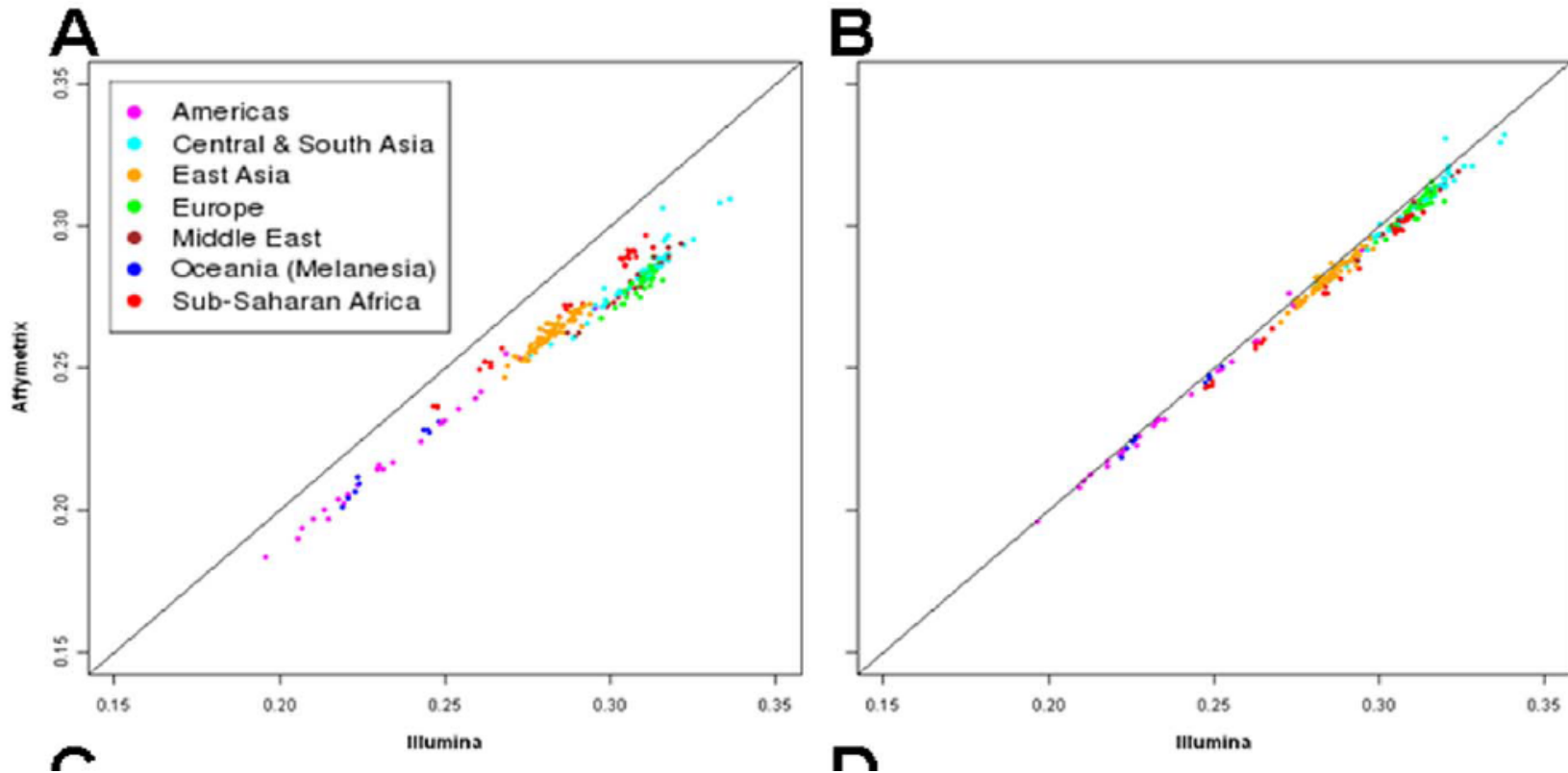
# Worldwide genetic variation and structure

- Affymetrix 500K
- 5 individuals from every population
- Illumina 650K
- HGDP most used panel
- 5 individuals from every population

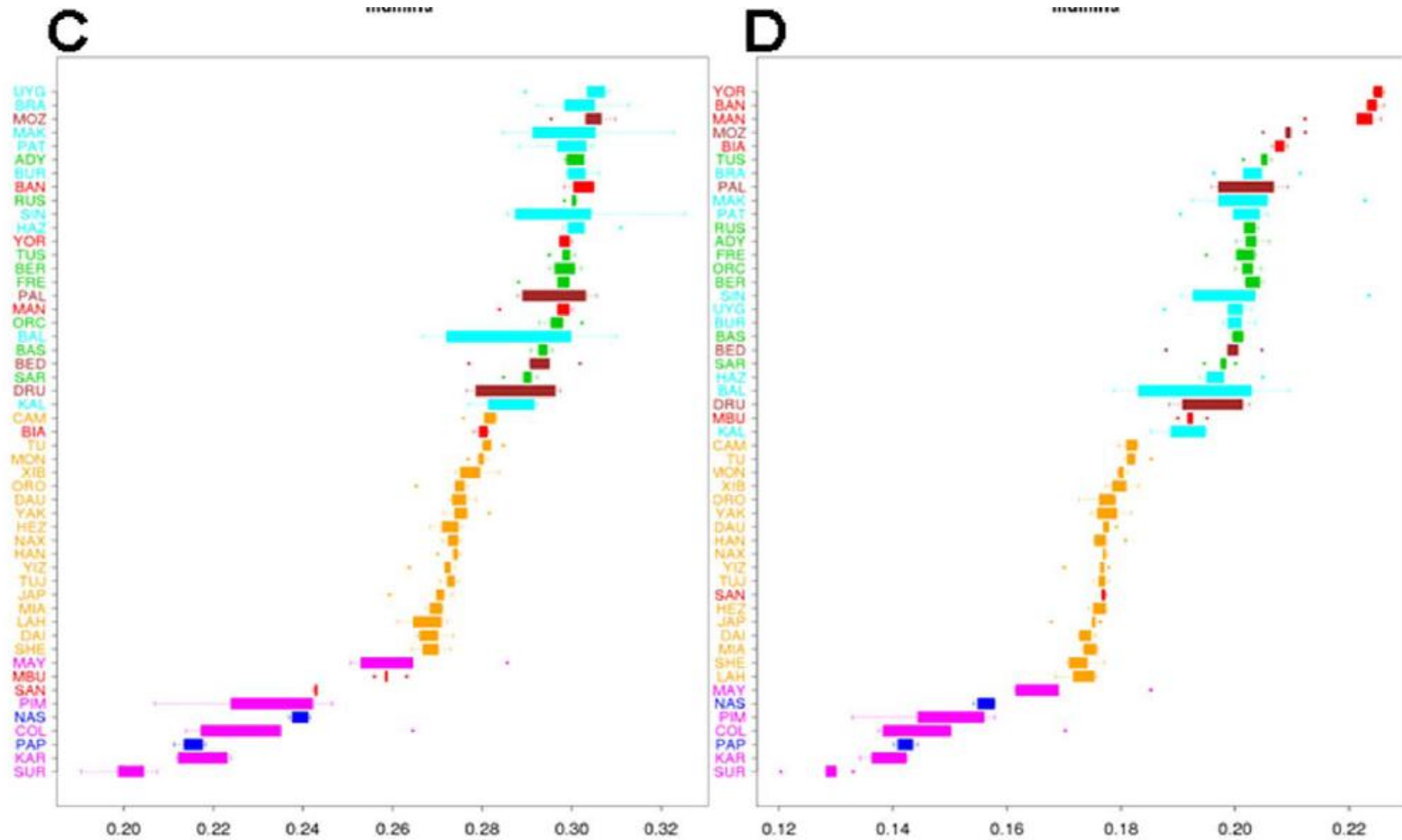
Overlap between two platforms is 96,849 SNP-s. Two platform together is about 950,000 SNP-s.

Pruned dataset is 220,247 SNP-s. (LD pruning with PLINK)

# Worldwide genetic variation and structure



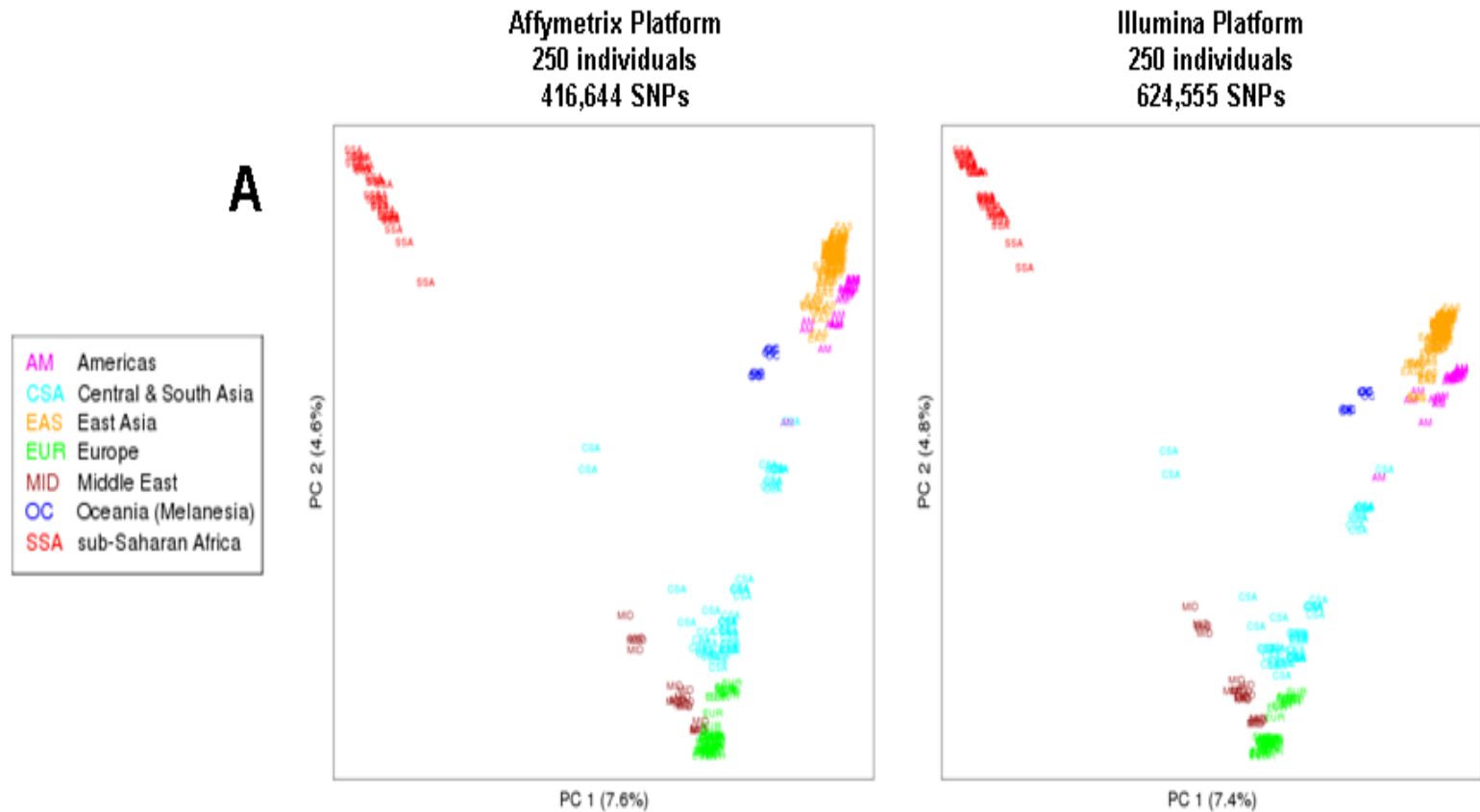
# Worldwide genetic variation and structure



# Worldwide genetic variation and structure

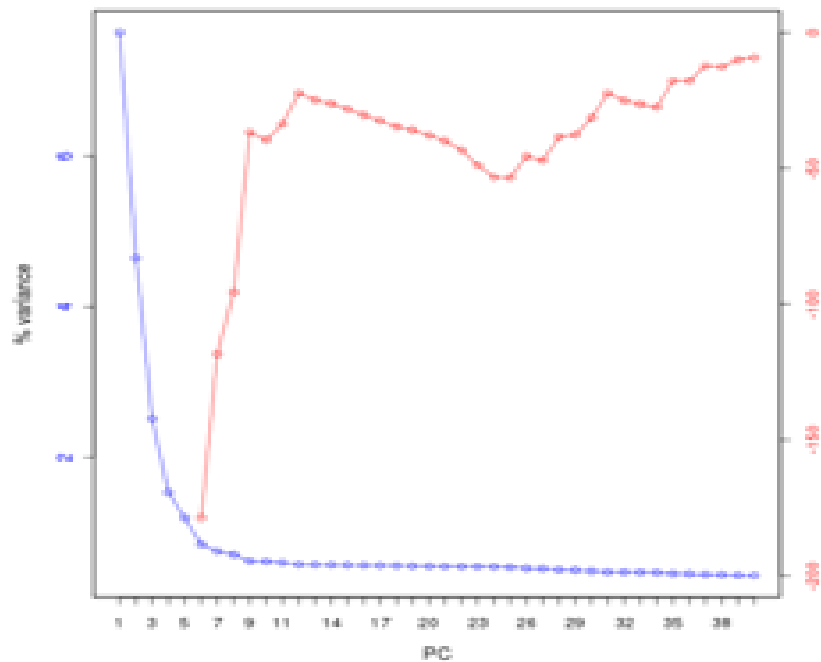
- PCA- principal component analysis
- Usually PCA – population level, but not always (like here)
- PCA outputs each individual's coordinates along axes of variation
- STRUCTURE like program *frappe*
- Central goal is to classify individuals into discrete populations

# Worldwide genetic variation and structure

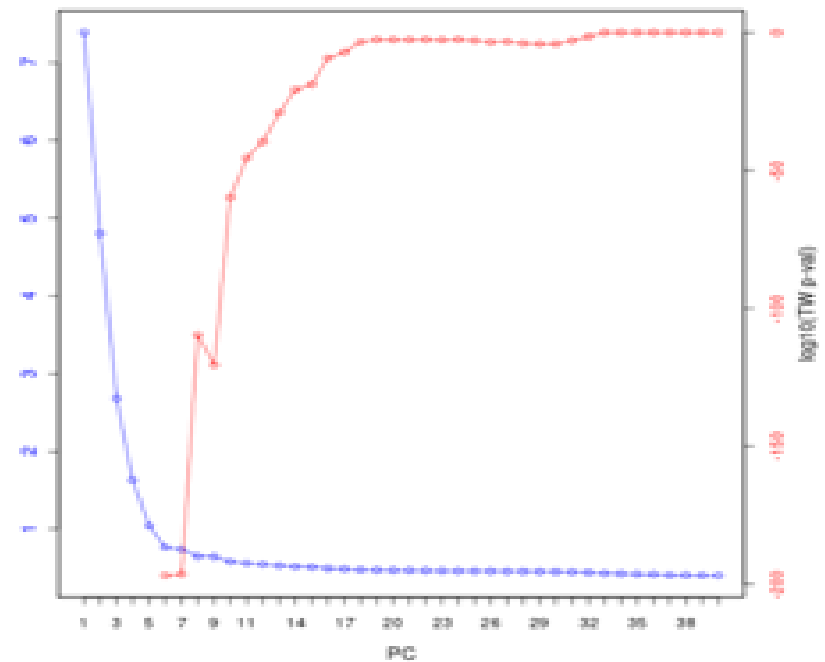


# Worldwide genetic variation and structure

**B**



**Affymetrix Platform**

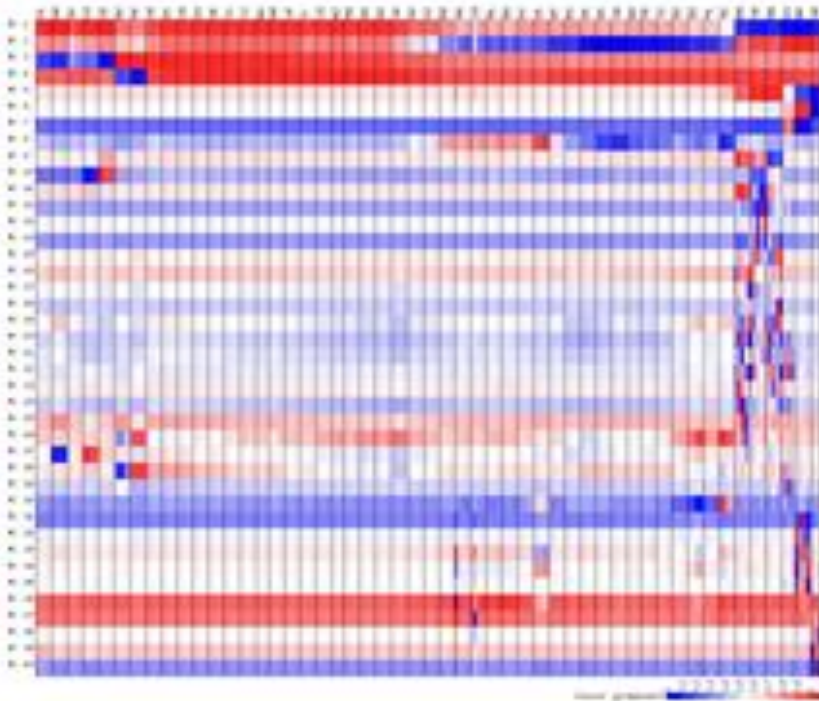


**Illumina Platform**

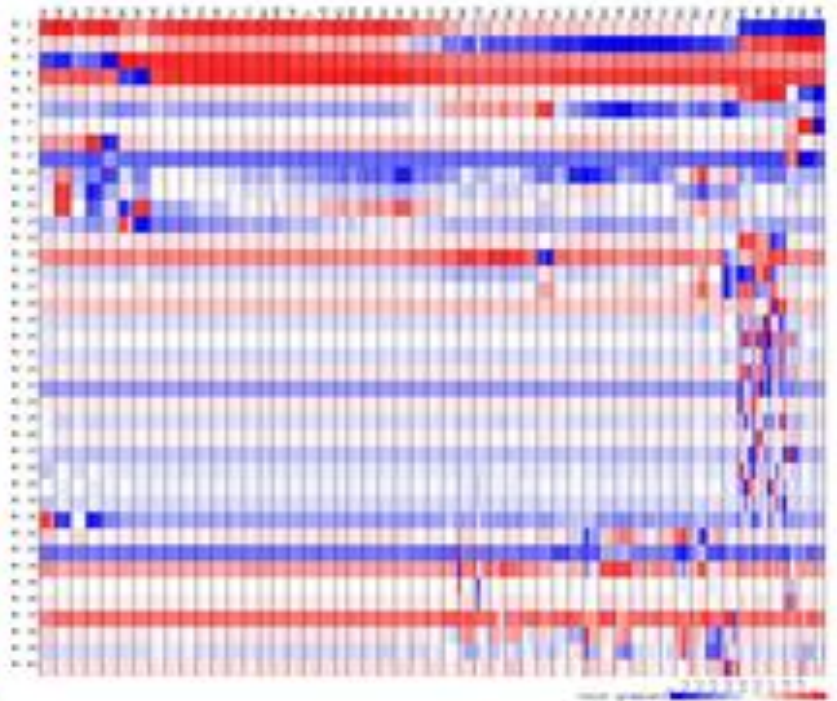


# Worldwide genetic variation and structure

C

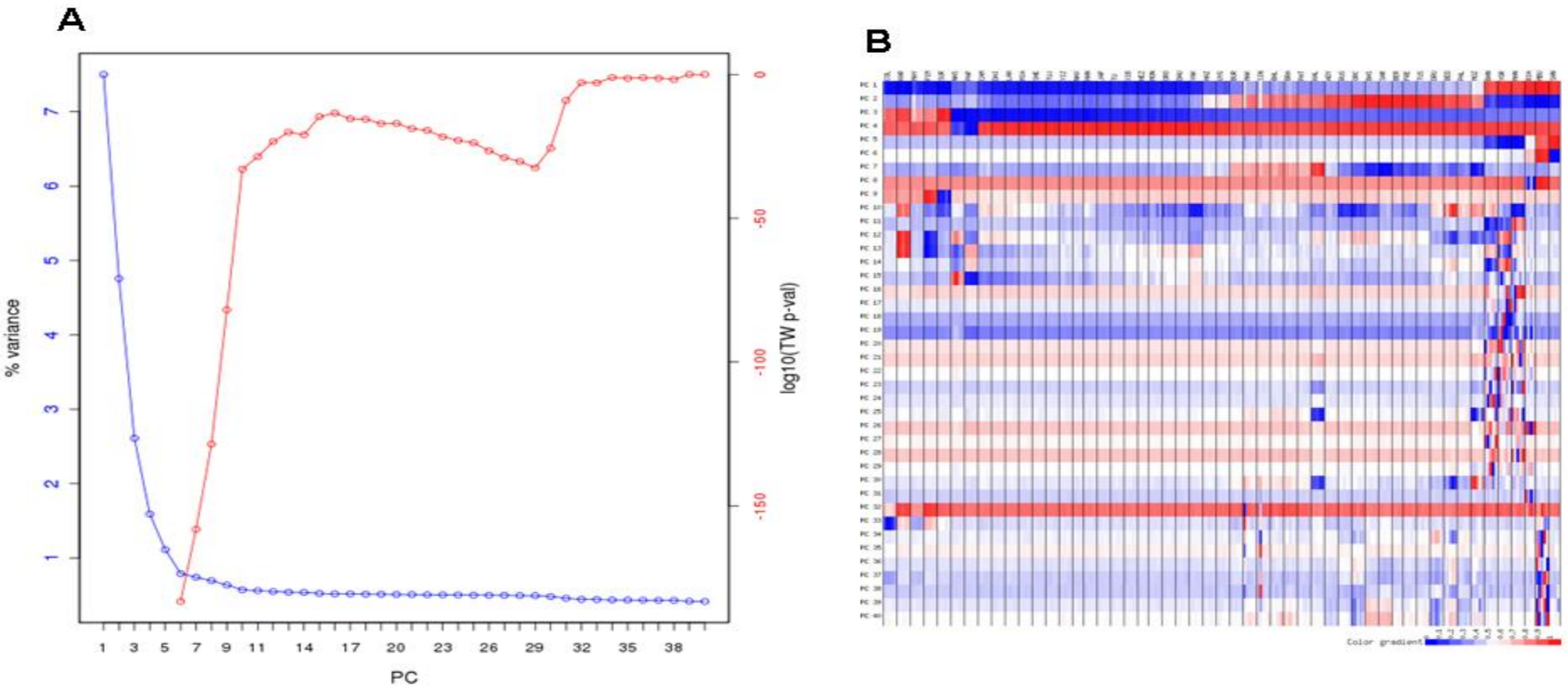


**Affymetrix Platform**



**Illumina Platform**

# Worldwide genetic variation and structure



# Worldwide genetic variation and structure

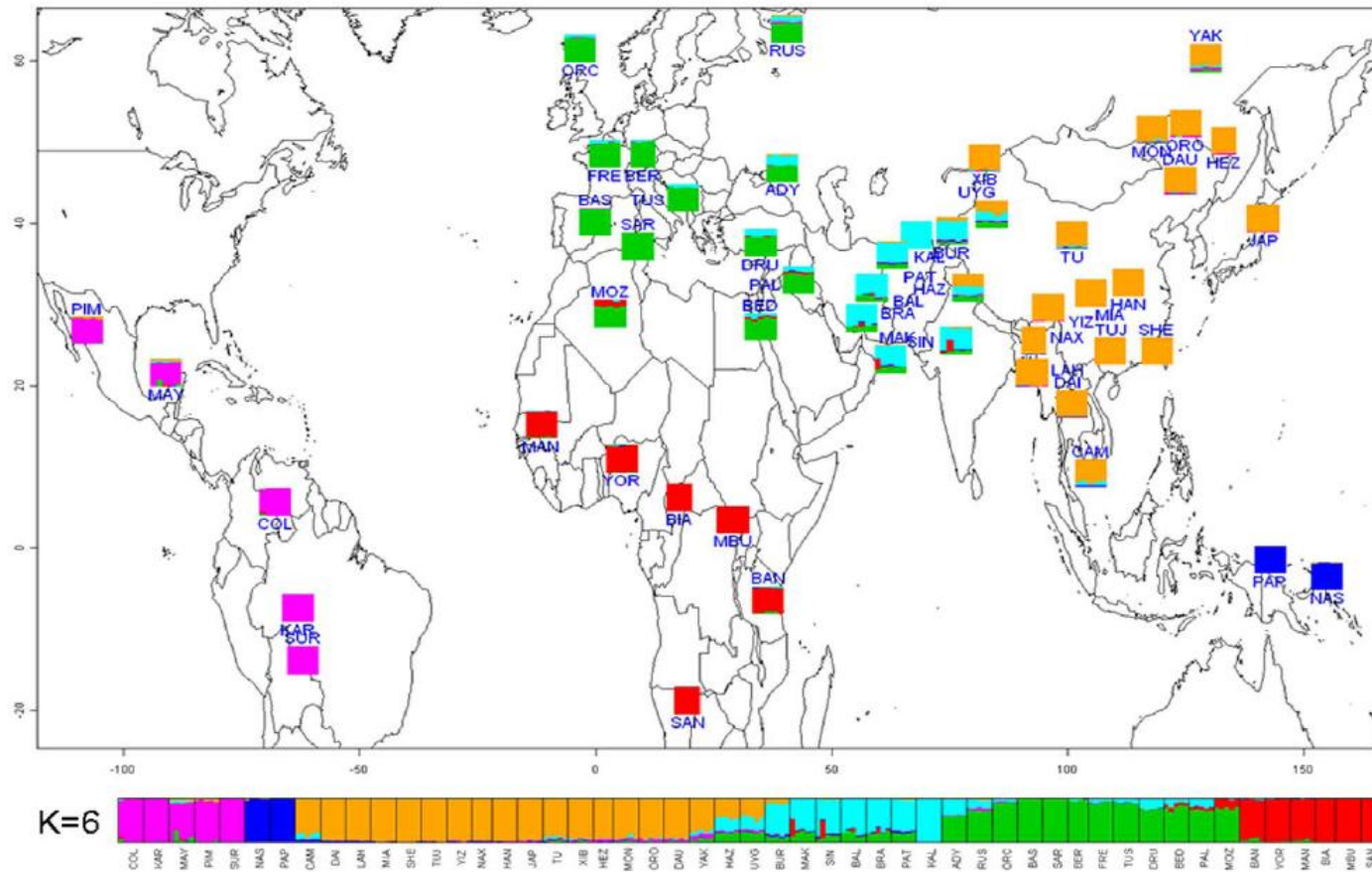
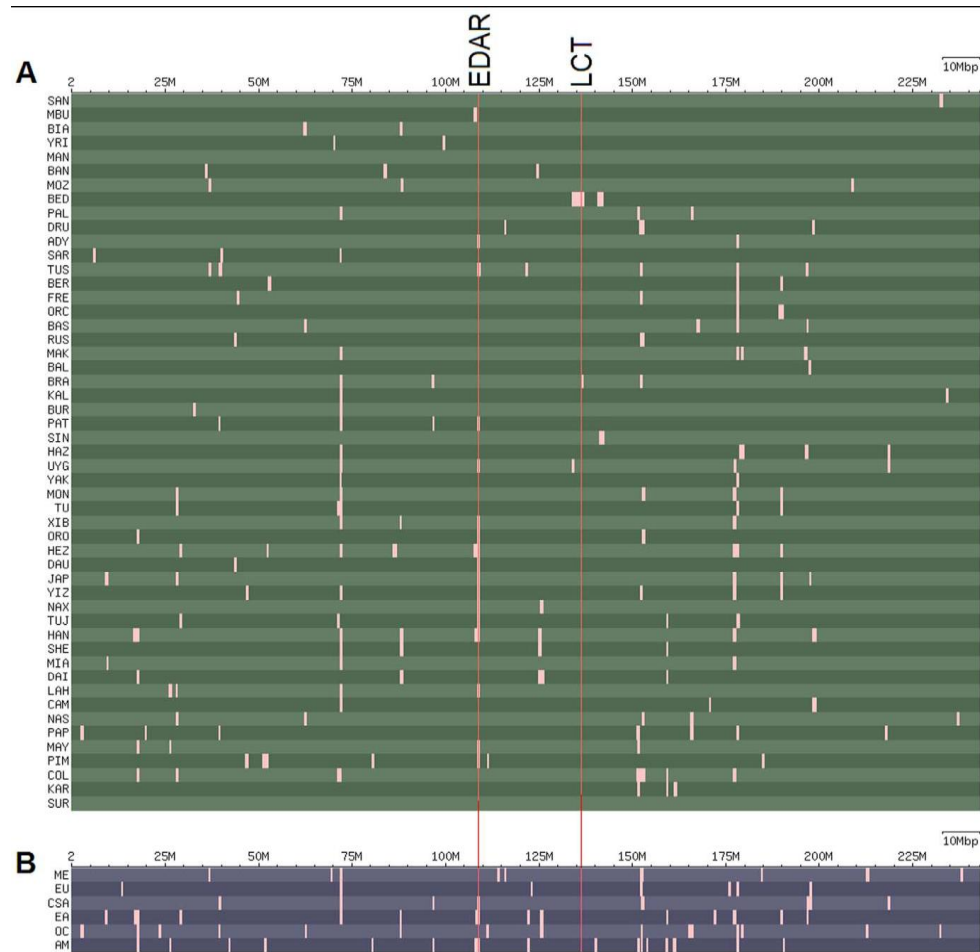


Figure 4. *frappe* results for  $K=6$ . Each color indicates a different ancestry component.

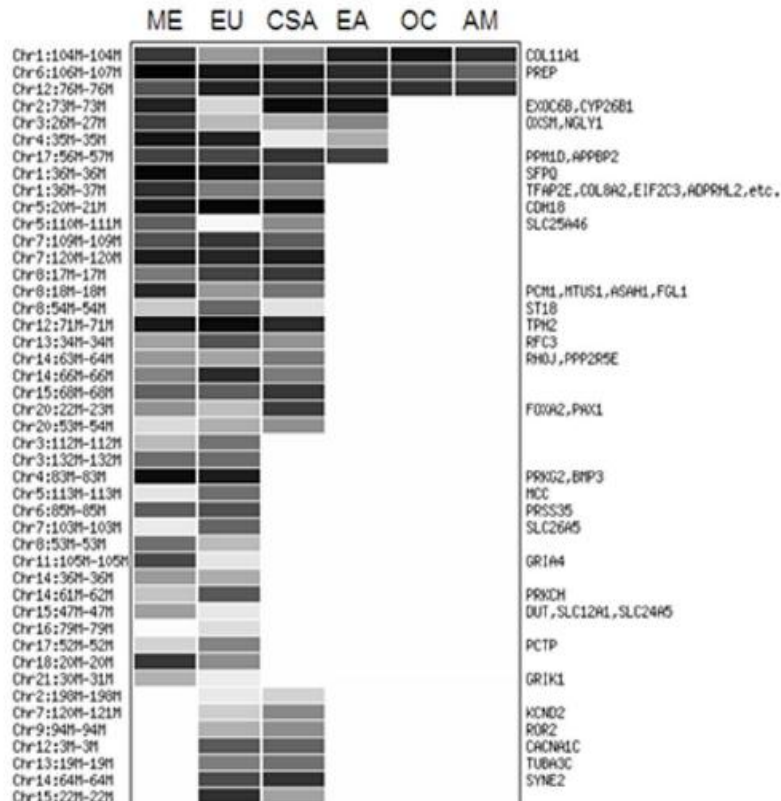
# Recent local selection

- Modified InRsb method – based on EHH. Reference population are from Africa. topSNPs (tSNP) are from upper 1% of the distribution. In candidate region should have at least 25 tSNPs in a 100kb region.
- 632 regions.

# Recent local selection



# Recent local selection



Chr1:103114611-103346640, 22 non-African populations, COL11A1

Chr1:153067871-153352772, SHC1 and KCNN3

Chr8:10917927-11894726, cluster of  $\beta$ -defensin genes

Chr5:70800637-71275894, CARTPT gene

Chr9:38623652-38761831, no genes

# GO enrichment analysis

**Table 1.** GO categories significantly over-represented among candidate gene regions for recent positive selection, as identified by the FUNC analysis.

GO name	GO ID	FDR for over-representation
sensory perception of chemical stimulus	GO:0007606	0
G-protein coupled receptor activity	GO:0004930	0
olfactory receptor activity	GO:0004984	0
integral to membrane	GO:0016021	0.00E+00
signal transducer activity	GO:0004871	1.14E-04
molecular transducer activity	GO:0060089	1.14E-04
cognition	GO:0050890	0.000779879
rhodopsin-like receptor activity	GO:0001584	0.000997632
neurological system process	GO:0050877	0.0016896
MHC protein complex	GO:0042611	0.00179347
glucuronosyltransferase activity	GO:0015020	0.00303429
extracellular region	GO:0005576	0.00539895
antigen processing and presentation	GO:0019882	0.00551127
integrin complex	GO:0008305	0.00656041
neuropeptide signaling pathway	GO:0007218	0.0208835
xenobiotic metabolic process	GO:0006805	0.0401682
metabotropic glutamate, GABA-B-like receptor activity	GO:0008067	0.0406189
sulfate transport	GO:0008272	0.0423062

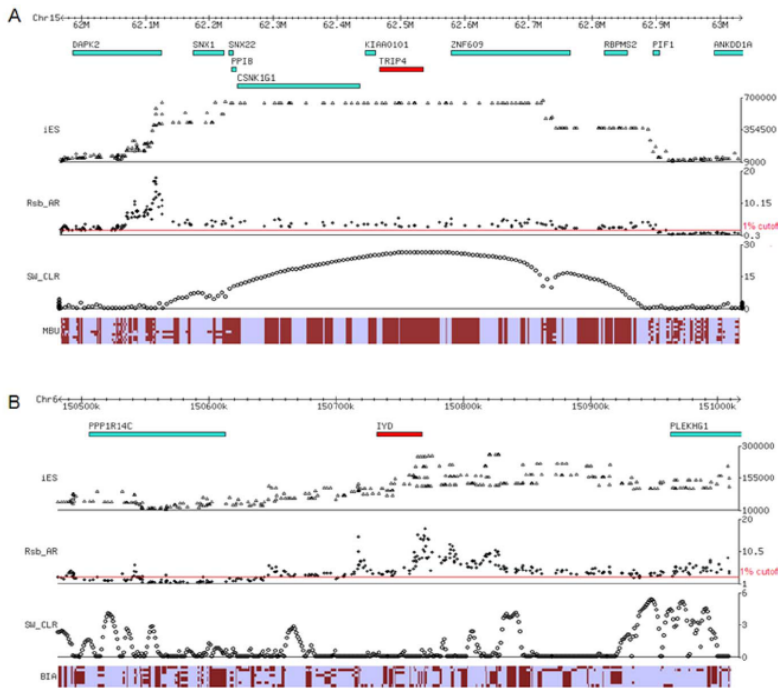
FDR, false discovery rate.

doi:10.1371/journal.pone.0007888.t001

- Gene Ontology analysis
- FUNC hypergeometric test (5000 random sets)

# Novel candidates in African pygmies

- Mbuti Pygmies – increase in EHH, absence of genetic diversity (TRIP4 gene)
- Biaka Pygmies – very strong  $\ln R_{sbAR}$  value (IYD gene)



**Figure 7. Signals of selection for thyroid hormone pathway genes in African Pygmies.** (A) 1 Mb region around TRIP4 in Mbuti Pygmies. (B) 500 Kb region around IYD in Biaka Pygmies. In each panel, the top part is the region surrounding the candidate gene with the position of genes indicated, followed by the distribution of the IES,  $R_{sbAR}$ , and Sweepfinder [7] statistics, followed by a diagram of the haplotypes observed in either Mbuti (A) or Biaka (B) Pygmies, where each row is a haplotype, each column is a SNP, and the light vs. dark shading indicates the alternative alleles for each SNP.  
doi:10.1371/journal.pone.0007888.g007