# Genotype, haplotype and copy-number variation in worldwide human populations

Priit Palta

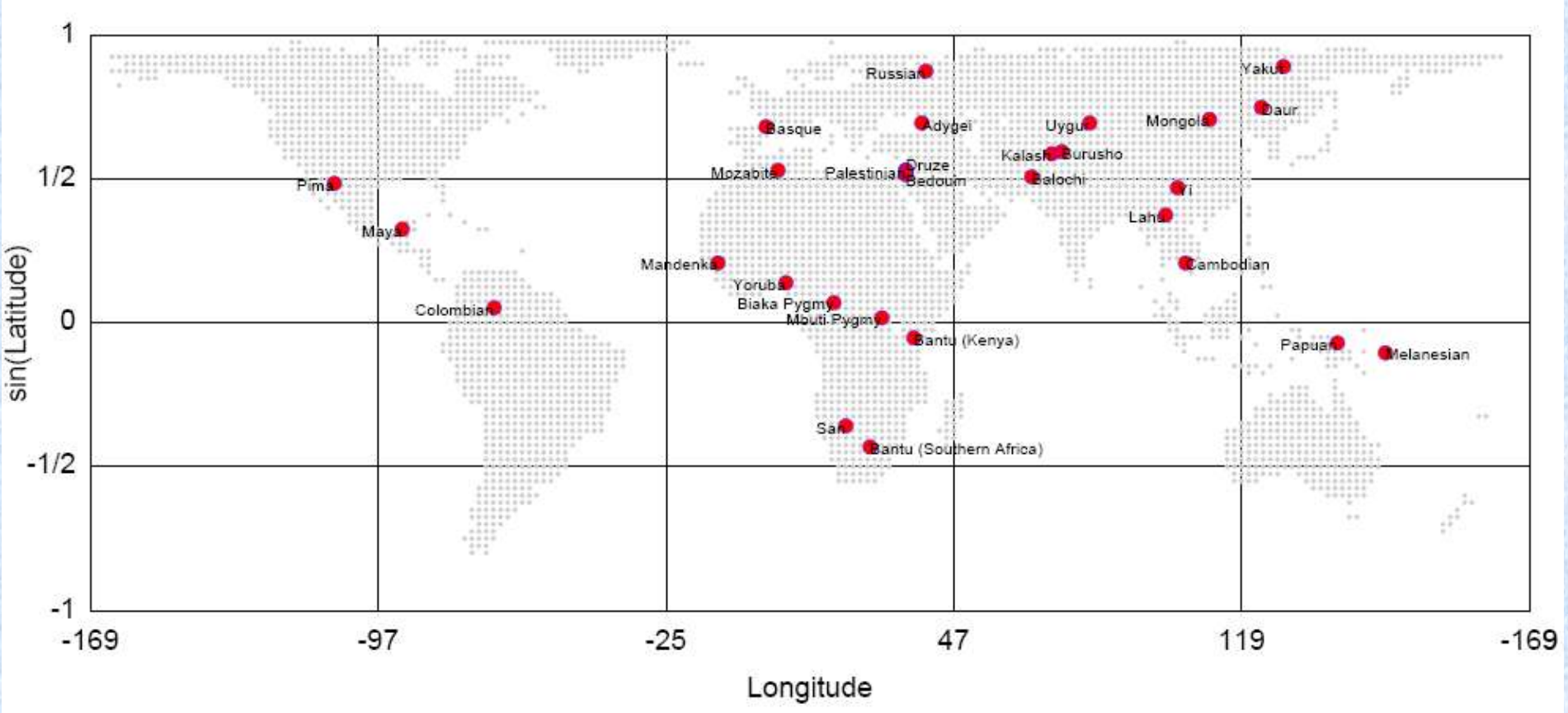
Bioinfo JC

29.04.2008

# Setup

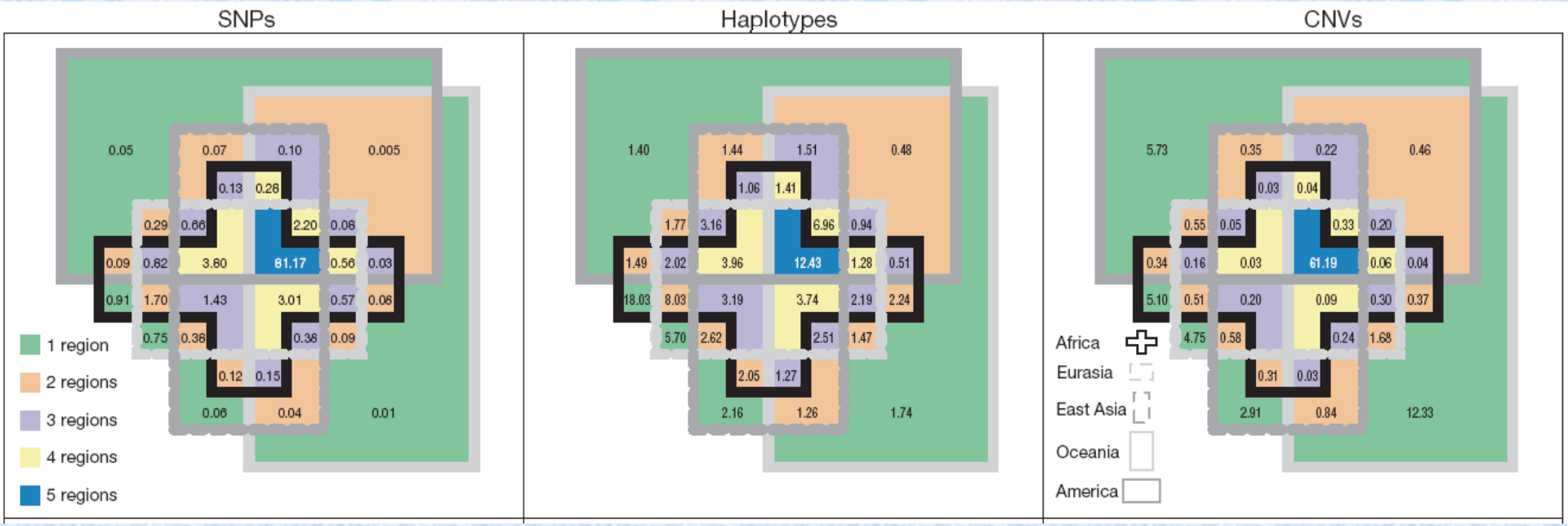
- Illumina Infinium HumanHap550 Genotyping BeadChips (~545 066 SNPs)

# Data

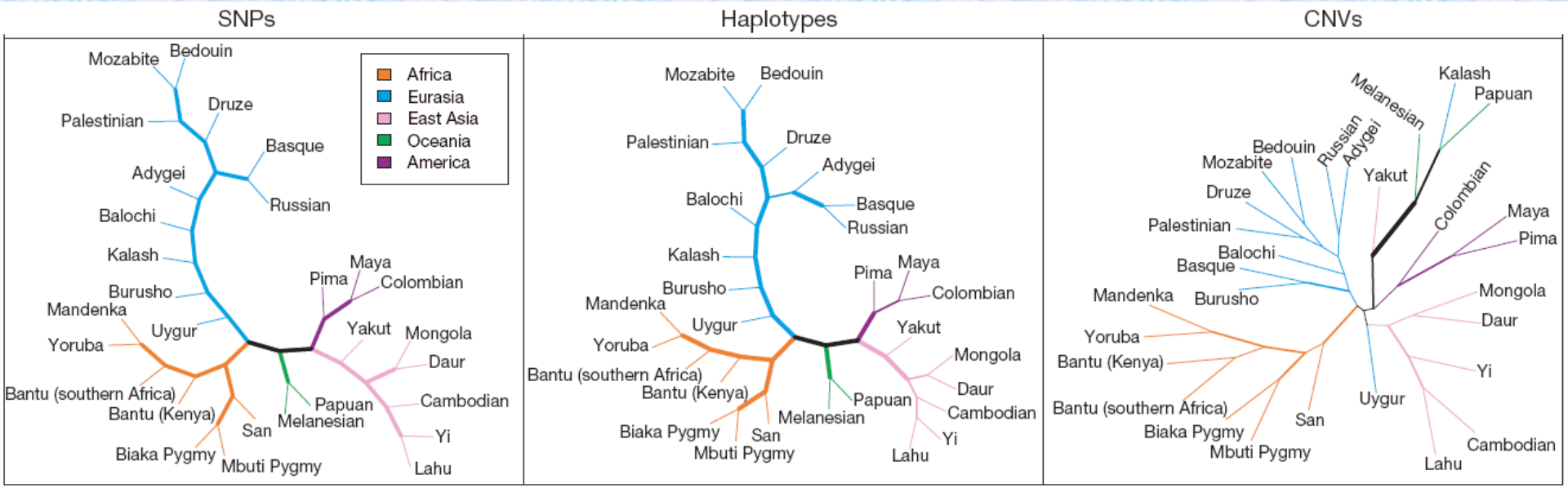



- 29 populations, 485 (513 before QC) individuals from the HGDP-CEPH panel

# Results I:
# Venn diagram of the precentages of alleles with particular geographic distributions

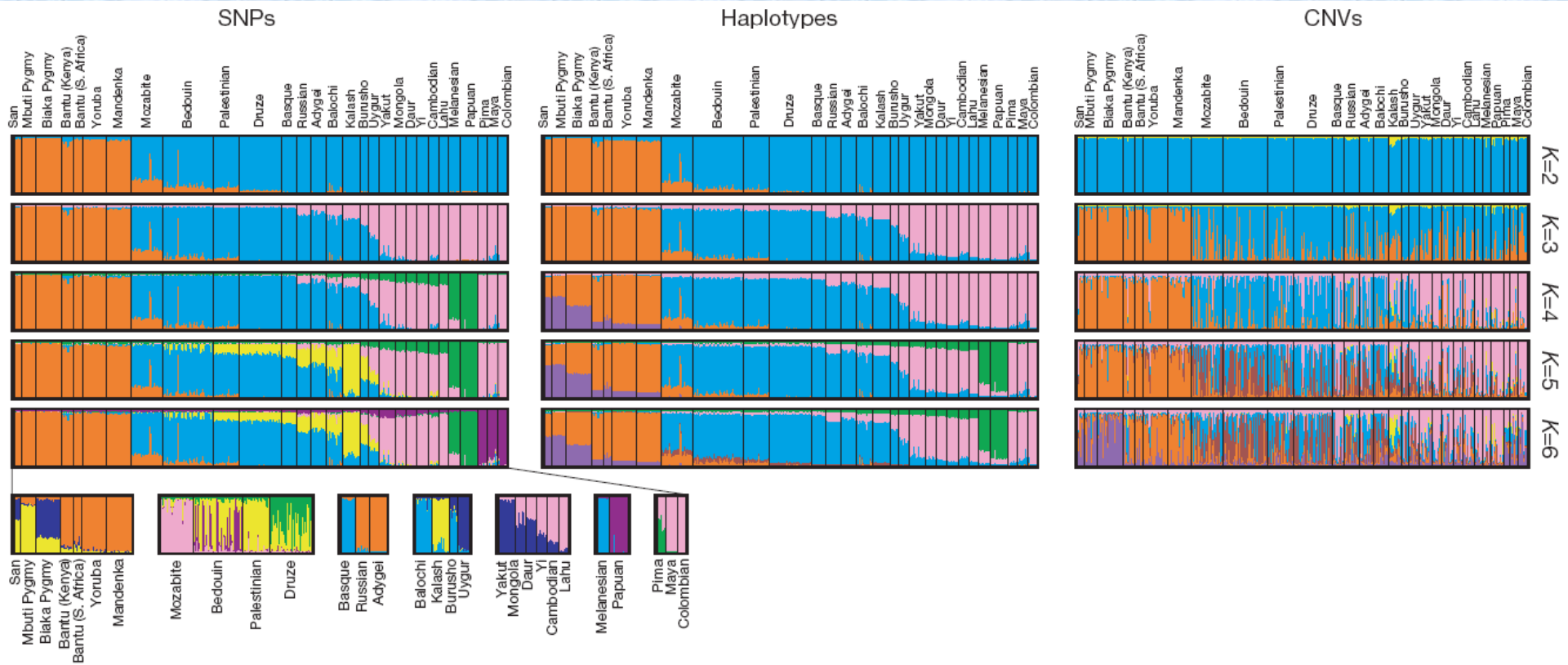


# Results II:
# Neighbour-joining trees of population relationships



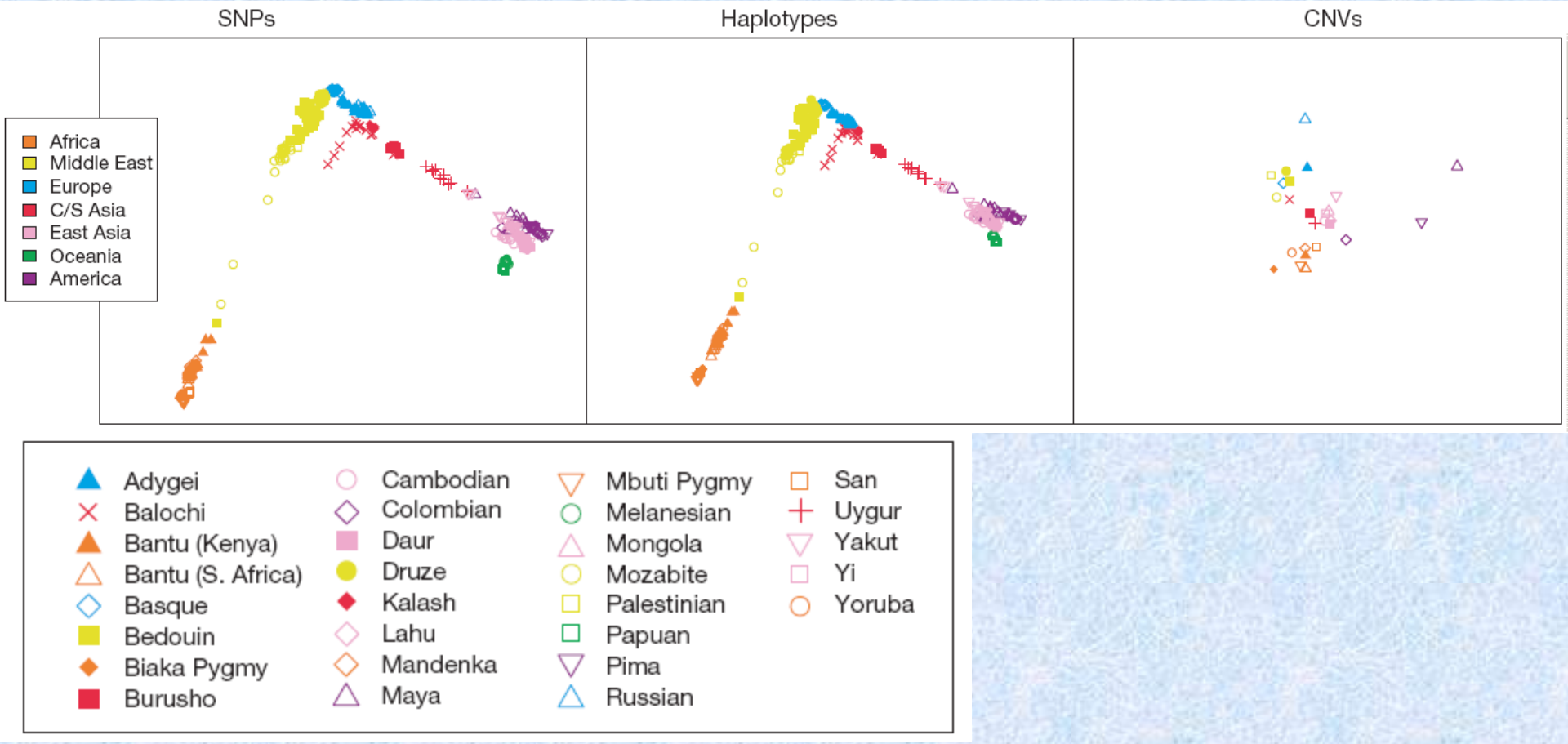

Internal branch lengths are proportional to bootstrap support. Lines of intermediate thickness represent internal branches with more than 50% bootstrap support, and the thickest lines represent more than 95% support.

# Results III:
# Population structure inferred by Bayesian clustering




Each individual is shown as a thin vertical line partitioned into K coloured components representing inferred membership in K genetic clusters. The bottom row provides inferred population structure for each geographic region.

# Results IV:
# Multidimensional clustering (MDS)




MDS representations of genetic distances between individuals (SNPs and haplotypes) and populations (CNVs). C/S Asia, Central/South Asia.

# Results IV:
# Multidimensional clustering (MDS)

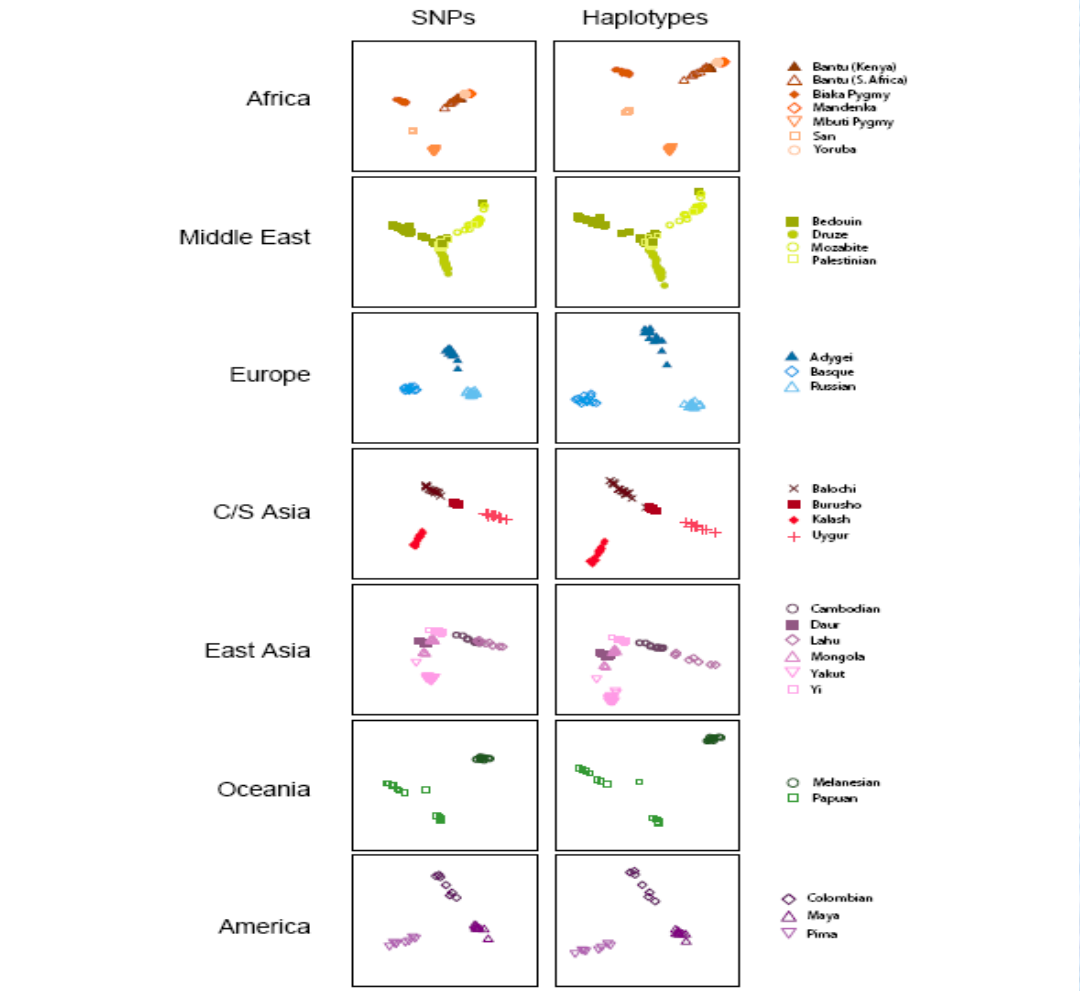



Figure S3: Multidimensional scaling representations of genetic distance matrices for individual geographic regions. Analysis was performed separately for the unphased SNP data and for the haplotype cluster data. This plot illustrates that within a given geographic region, the individuals of different populations can be further subdivided into clusters corresponding largely to distinct populations.

# Results V:
# Genetic and geographic distance and LD

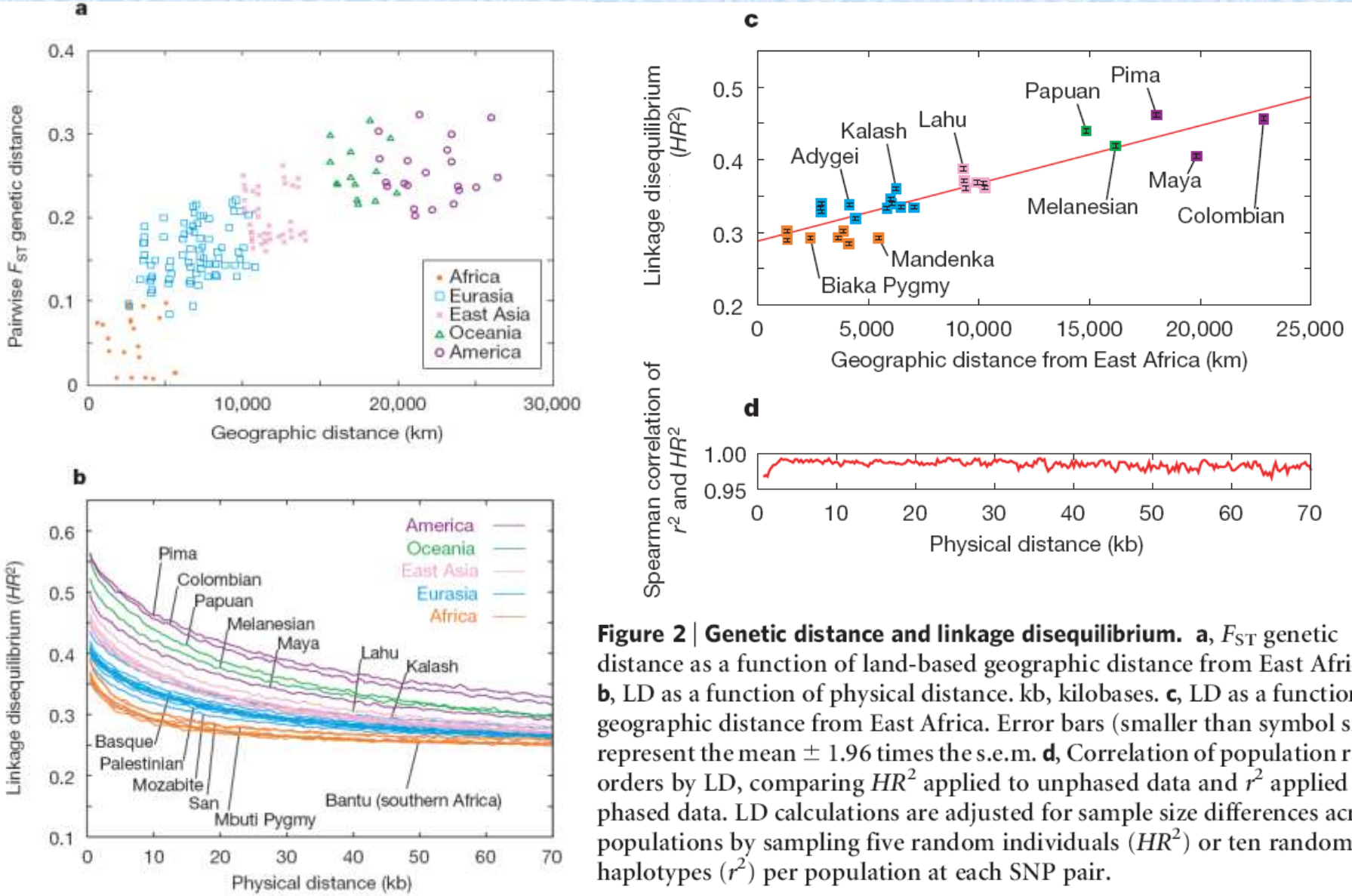



**Figure 2 | Genetic distance and linkage disequilibrium.** **a**, $F_{ST}$ genetic distance as a function of land-based geographic distance from East Africa. **b**, LD as a function of physical distance. kb, kilobases. **c**, LD as a function of geographic distance from East Africa. Error bars (smaller than symbol size) represent the mean ± 1.96 times the s.e.m. **d**, Correlation of population rank orders by LD, comparing $HR^2$ applied to unphased data and $r^2$ applied to phased data. LD calculations are adjusted for sample size differences across populations by sampling five random individuals ($HR^2$) or ten random haplotypes ($r^2$) per population at each SNP pair.

# Results VI:
# CNVs across populations

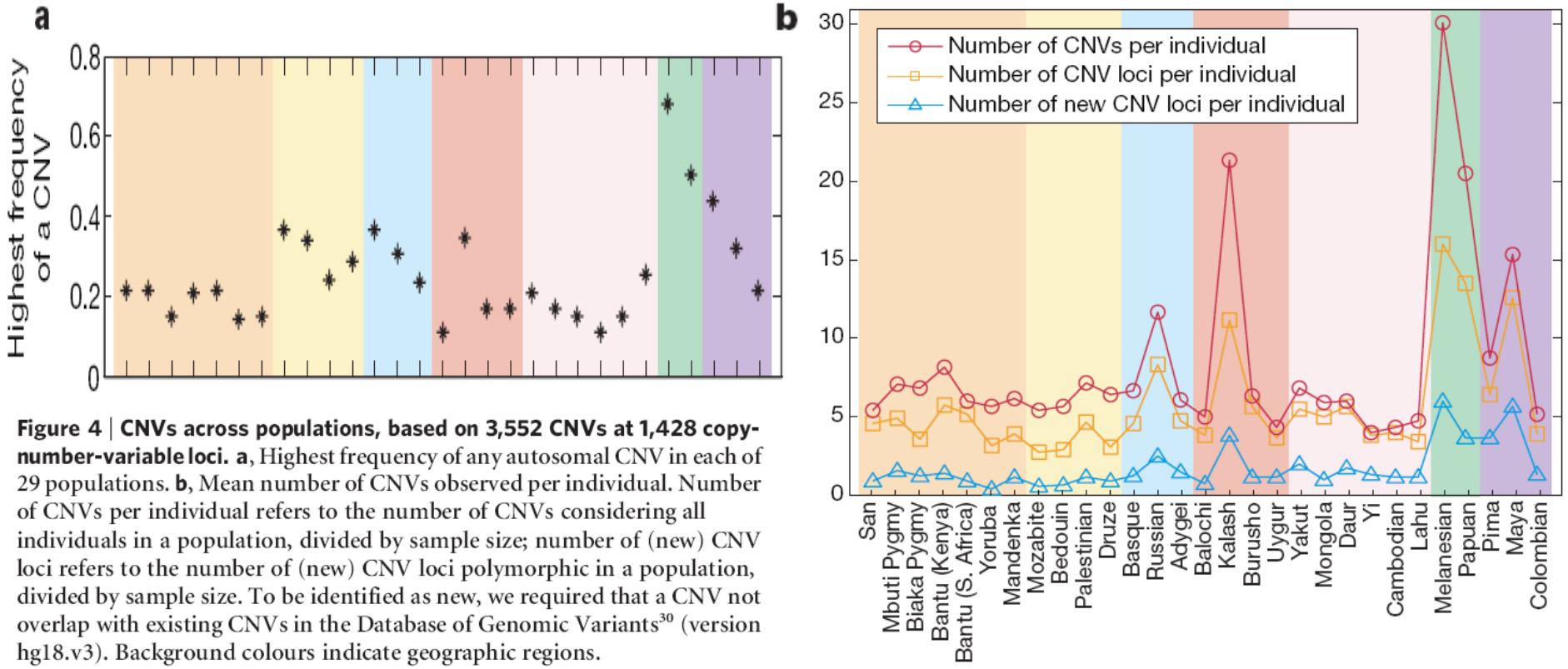



**Figure 4 | CNVs across populations, based on 3,552 CNVs at 1,428 copy-number-variable loci. a**, Highest frequency of any autosomal CNV in each of 29 populations. **b**, Mean number of CNVs observed per individual. Number of CNVs per individual refers to the number of CNVs considering all individuals in a population, divided by sample size; number of (new) CNV loci refers to the number of (new) CNV loci polymorphic in a population, divided by sample size. To be identified as new, we required that a CNV not overlap with existing CNVs in the Database of Genomic Variants[30] (version hg18.v3). Background colours indicate geographic regions.

# Results VI: Frequency of CNVs

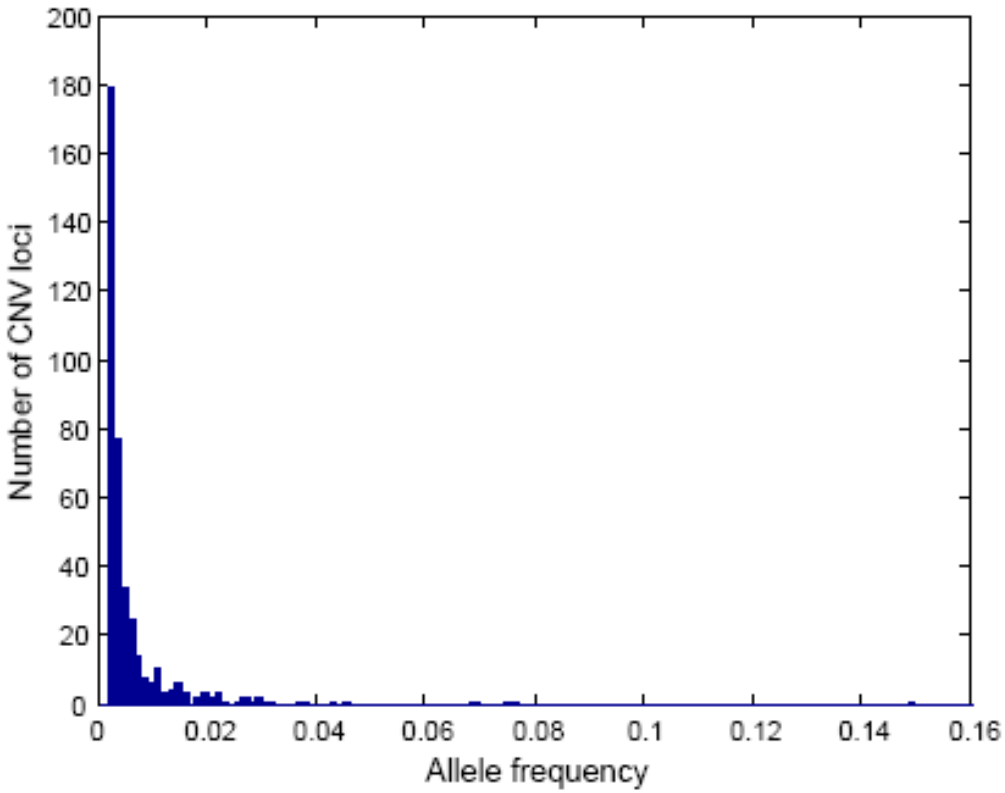



Figure S12: Allele frequency spectrum for the copy-number variants at 1302 autosomal CNV loci in 405 unrelated individuals from 29 populations. The 1/810 frequency class is not plotted and contains 906 loci. The figure illustrates that most CNVs were observed to be rare.

# Results VI: Length of CNVs

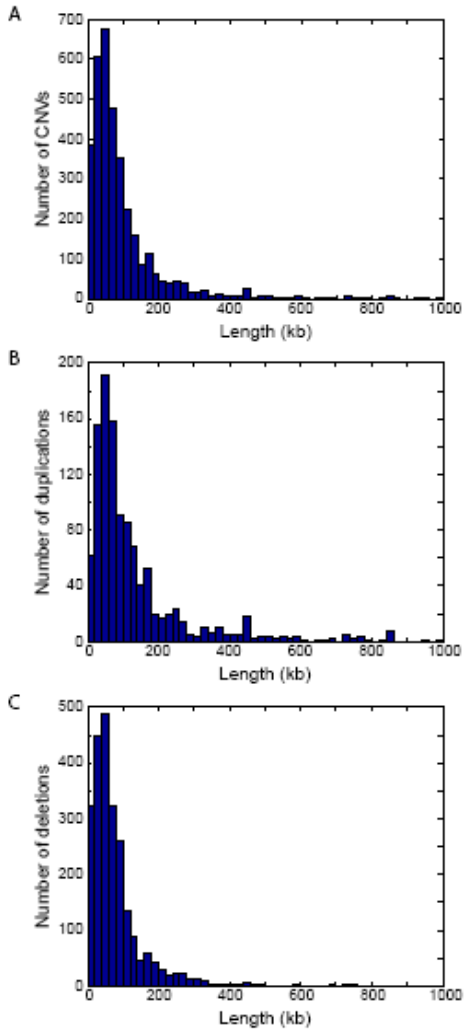



Figure S22: Distribution of the lengths of 3503 autosomal CNVs. (A) All CNVs. (B) Duplications (1117). (C) Deletions (2386). In kilobases, the bins in each histogram are (0, 20], ... , (980, 1000].

# References:

Jakobsson M, Scholz SW, Scheet P, Gibbs JR, VanLiere JM, Fung HC, Szpiech ZA, Degnan JH, Wang K, Guerreiro R, Bras JM, Schymick JC, Hernandez DG, Traynor BJ, Simon-Sanchez J, Matarin M, Britton A, van de Leemput J, Rafferty I, Bucan M, Cann HM, Hardy JA, Rosenberg NA, Singleton AB. **Genotype, haplotype and copy-number variation in worldwide human populations.** Nature. 2008 Feb 21;451(7181):998-1003.

# Multidimensional clustering (MDS)

1. A matrix of pairwise distances was constructed for the 443 unrelated HGDP-CEPH individuals, using the 512,762 autosomal SNPs.
2. Between-individual distances were obtained using allele-sharing distance, P0+P1/2, where Pk represents the proportion of loci at which the individuals shared exactly k alleles identical in state.
3. The overall distance between individuals was obtained as the average across loci.
4. Classical metric multidimensional scaling was applied to the individual distance matrix to provide a representation of the matrix in two dimensions.
5. The resulting coordinates were then rotated 225 to place the populations in an approximate geographic orientation.

# Multidimensional clustering (MDS), an exapmle in aaaaaRRRR!:

```
loc <- cmdscale(eurodist)
x <- loc[,1]
y <- -loc[,2]
plot(x, y, type="n", xlab="", ylab="", main="cmdscale(eurodist)")
text(x, y, rownames(loc), cex=0.8)
```

# LD with $HR^2$

- Linkage disequilibrium (LD) was measured for the unphased data using the $HR^2$ statistic, a measure analogous to the $r^2$ statistic for phased data that for a pair of SNPs considers a normalized squared difference between the proportion of double homozygotes expected under linkage equilibrium and the proportion of double homozygotes observed.