Rare Haplotypes in the Korean Population

Sepp Hochreiter, Günter Klambauer, Gundula Povysil, Djork-Arné Clevert

Institute of Bioinformatics, Johannes Kepler University Linz, Austria

Knowledge of the haplotype structure of the human genome would improve genotype calls, increase the power of association studies, and shed light on the evolutionary history of humans. Common haplotypes are found by regions of linkage disequilibrium (LD) in genotype data. The advent of new sequencing technologies also facilitates the identification of rare haplotypes. However, LD-related methods fail to extract rare haplotypes because of the high variance of LD measures. Rare haplotypes can be inferred by a region of identity by descent (IBD) in two individuals. However, IBD detection methods require sufficiently long IBD regions to avoid high false positive rates and are computationally expensive as they consider all pairs of individuals. We propose identifying rare haplotypes by HapFABIA which uses biclustering to combine LD information across individuals and IBD information along the chromosome. HapFABIA significantly outperformed IBD methods at detecting rare haplotypes on simulated genotype data with implanted rare haplotypes.

To identify rare haplotypes in the Korean population, we applied HapFABIA to data from the Korean Personal Genome Project (KPGP) supplied via Critical Assessment of Massive Data Analysis (CAMDA). Genotyping data from the KPGP was combined with those from the 1000-Genomes-Project leading to 1,131 individuals and 3.1 million single nucleotide variants (SNVs) on chromosome 1 – we only analyzed chromosome 1 to comply with the Ft. Lauderdale agreement for the use of unpublished data for method development. For biclustering such large data sets, we developed a sparse matrix algebra for the FABIA biclustering algorithm.

HapFABIA identified 113,963 different rare haplotypes marked by tagSNVs that have a minor allele frequency of 5% or less. The rare haplotypes comprise 680,904 SNVs; that is 36.1% of the rare variants and 21.5% of all variants. The vast majority of 107,473 haplotypes is found in Africans, while only 9,554 and 6,933 are found in Europeans and Asians, respectively.

HapFABIA revealed a large number of genotyping errors in the KPGP data (e.g. Figure 3). The KPGP data comprises two twin pairs and a large Korean family that contains a Caucasian female from US. In particular, genotyping errors are found as SNV disagreements at twin haplotypes (e.g. Figure 5) and by haplotypes that were observed exclusively in KPGP samples including the Caucasian female (e.g. Figure 4). We corrected for these genotyping errors by removing haplotypes that are observed in just one population and removing all relations between individuals according to the pedigree information.

We characterized haplotypes by matching with archaic genomes. Haplotypes that match the Denisova or the Neandertal genome are significantly more often observed in Asians and Europeans. Interestingly, haplotypes matching the Denisova or the Neandertal genome are also found, in some cases exclusively, in Africans. Our findings indicate that the majority of rare haplotypes from chromosome 1 are ancient and are from times before humans migrated out of Africa.

The enrichment of Neandertal haplotypes in Koreans (odds ratio 10.6 of Fisher's exact test) is not as high as for Han Chinese from Beijing, Han Chinese from South, and Japanese (odds ratios 23.9, 19.1, 22.7 of Fisher's exact test) – see also Figure 7. In contrast to these results, the enrichment of Denisova haplotypes in Koreans (odds ratio 36.7 of Fisher's exact test) is is higher than for Han Chinese from Beijing, Han Chinese from South, and Japanese (odds ratios 7.6, 6.9, 7.0 of Fisher's exact test) – see also Figure 6 and examples in Figure 1 and Figure 2.

Data Analysis steps:

1. Combine the vcf genotyping data from KPGP with those from the 1000-Genomes-Project (3.1 million SNVs on chromosome 1 of 1,134 individuals – vcftools, samtools)
2. Remove common and private SNVs
3. Transform the genotyping data into the sparse matrix format of HapFABIA
4. Apply HapFABIA to extract haplotypes
5. Base calling of Denisova and Neandertal genome at the SNV positions of KPGP and 1000-Genomes-Project
6. Analyze and annotate the haplotypes

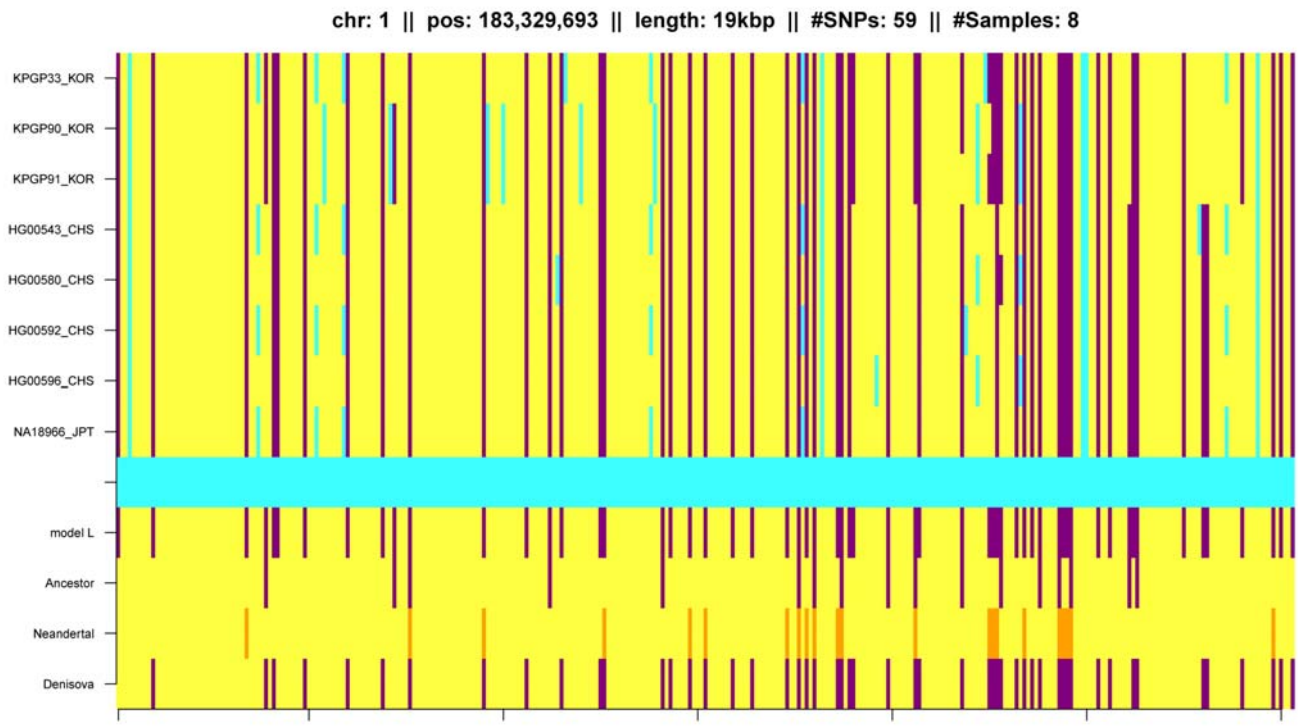

Figure 1: Example of a haplotype matching the Denisova genome found exclusively in Asians including Koreans. The y-axis gives all chromosomes that have the haplotype and the x-axis consecutive SNVs/Indels/SVs. Major alleles are shown in yellow, minor alleles of tagSNVs in violet, and minor alleles of other SNVs in cyan. The row labeled "model L" indicates tagSNVs identified by HapFABIA in violet. The rows "Ancestor", "Neandertal", and "Denisova" show bases of the respective genomes in violet if they match the minor allele of the tagSNVs (in yellow otherwise). Missing Neandertal tagSNV bases are shown in orange.

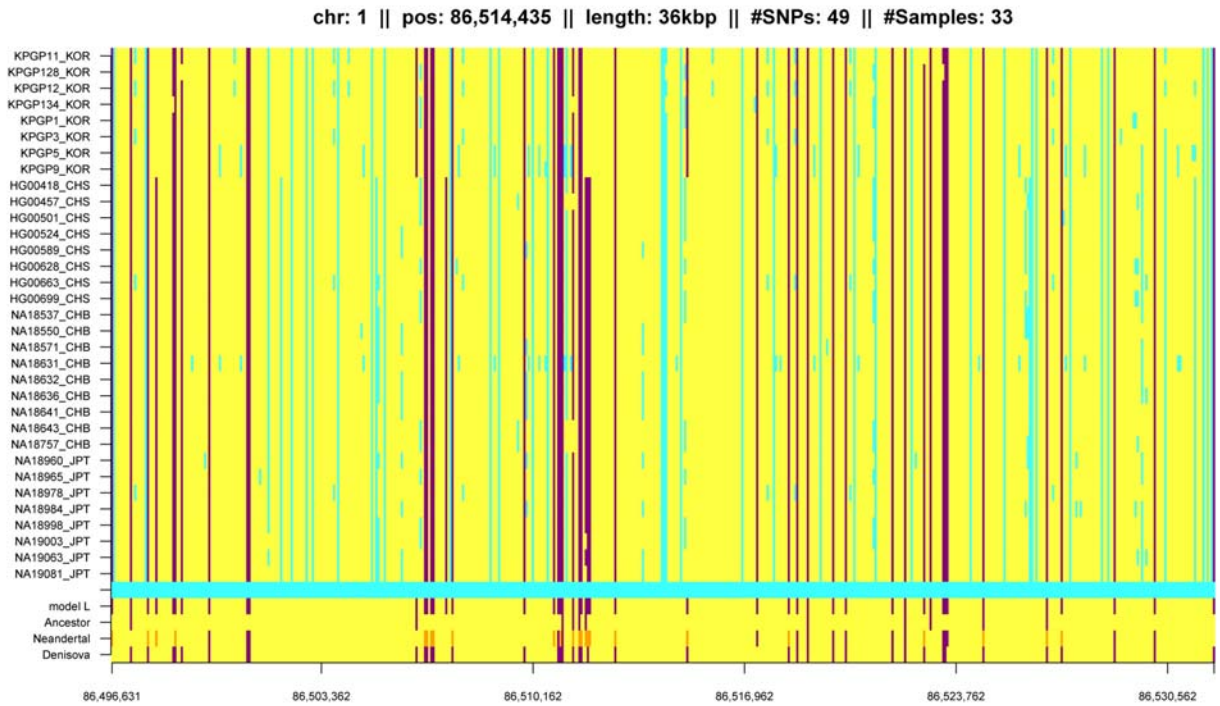

Figure 2: Another example of a haplotype matching the Denisova genome including Koreans.

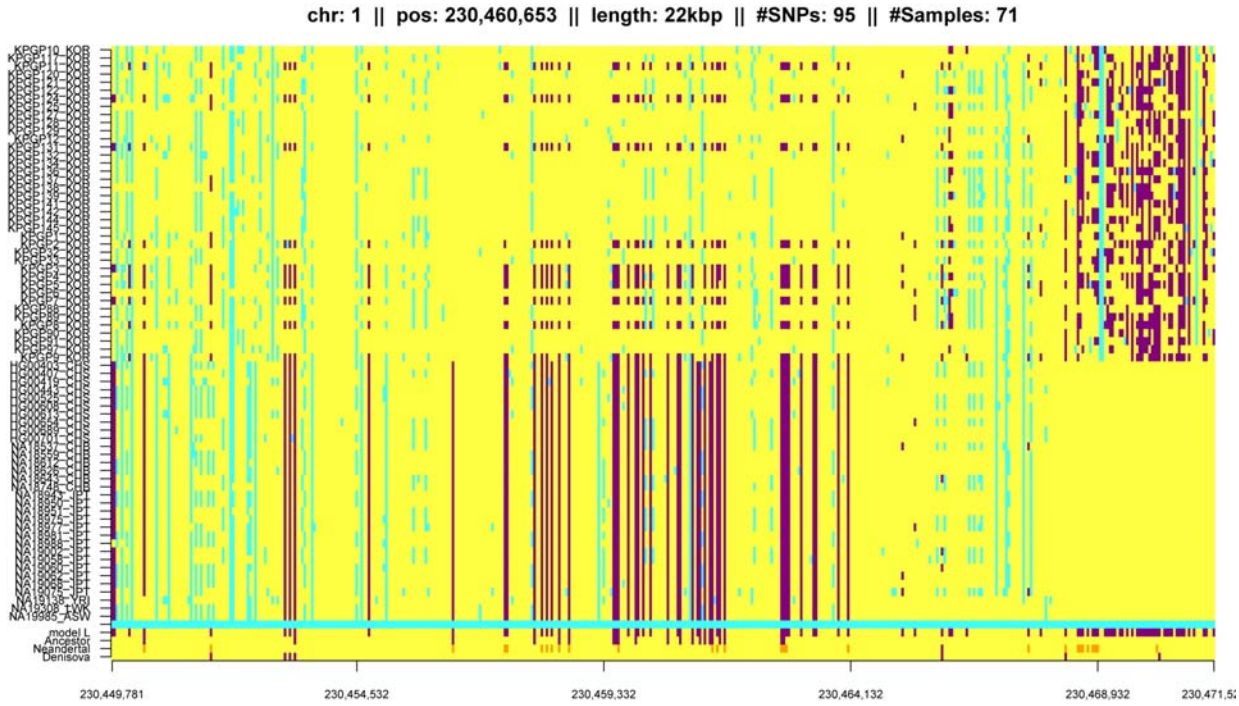

Figure 3: Example of a haplotype that contains genotyping errors (right border).

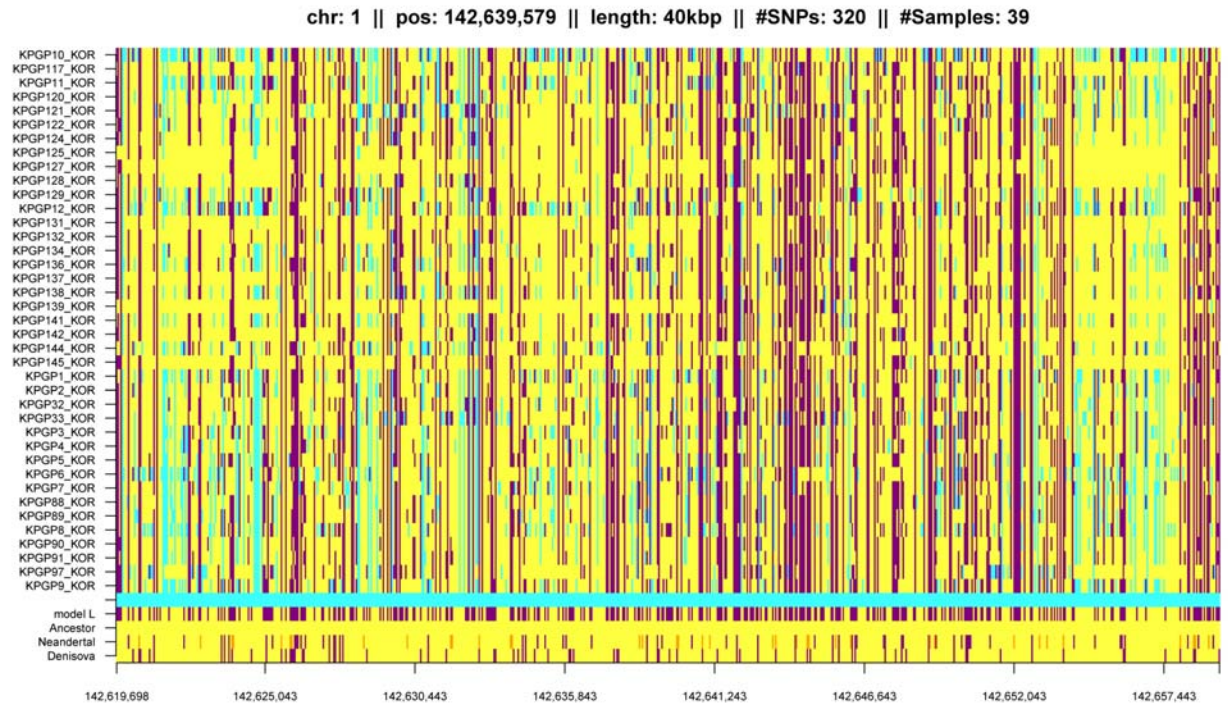

Figure 4: Example of a haplotype representing genotyping errors. The haplotype is exclusively observed in KPGP samples, however KPGP10 is an US American Caucasian female sequenced by KPGP. Many tagSNVs are inconsistent across samples.

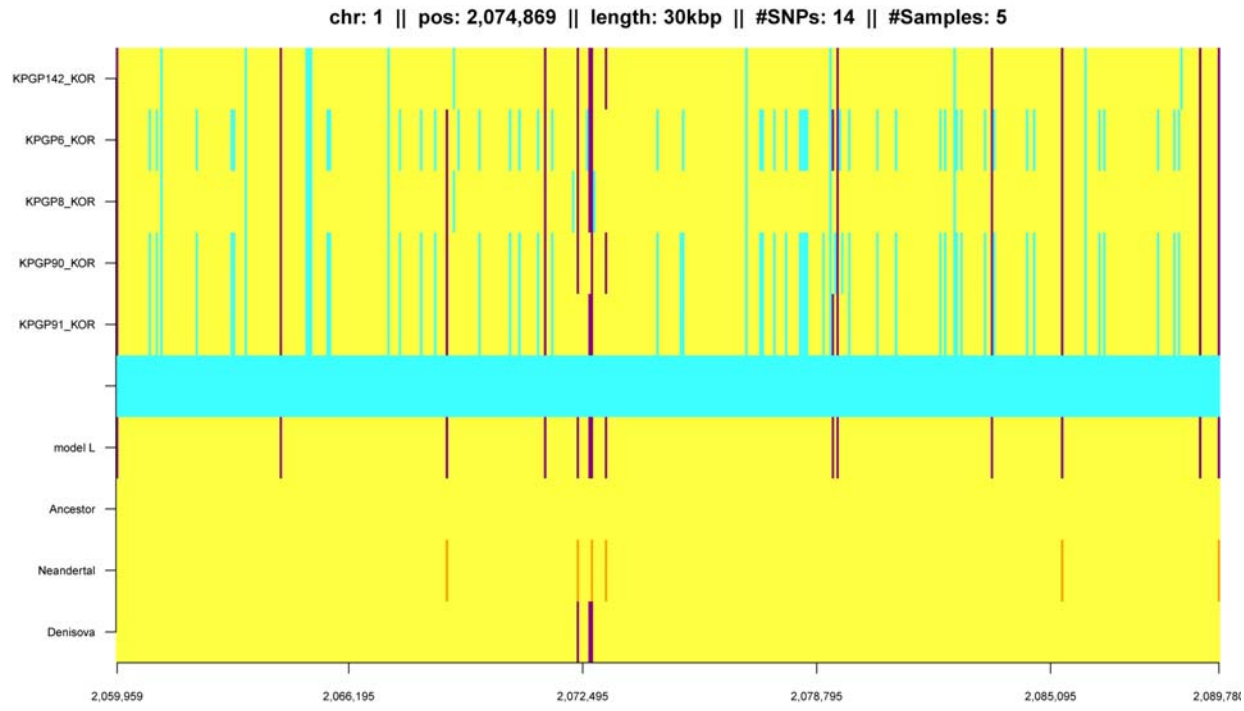

Figure 5: Example of a haplotype possessed by the twins KPGP90 and KPGP91. Differences in the twins are presumably genotyping errors.

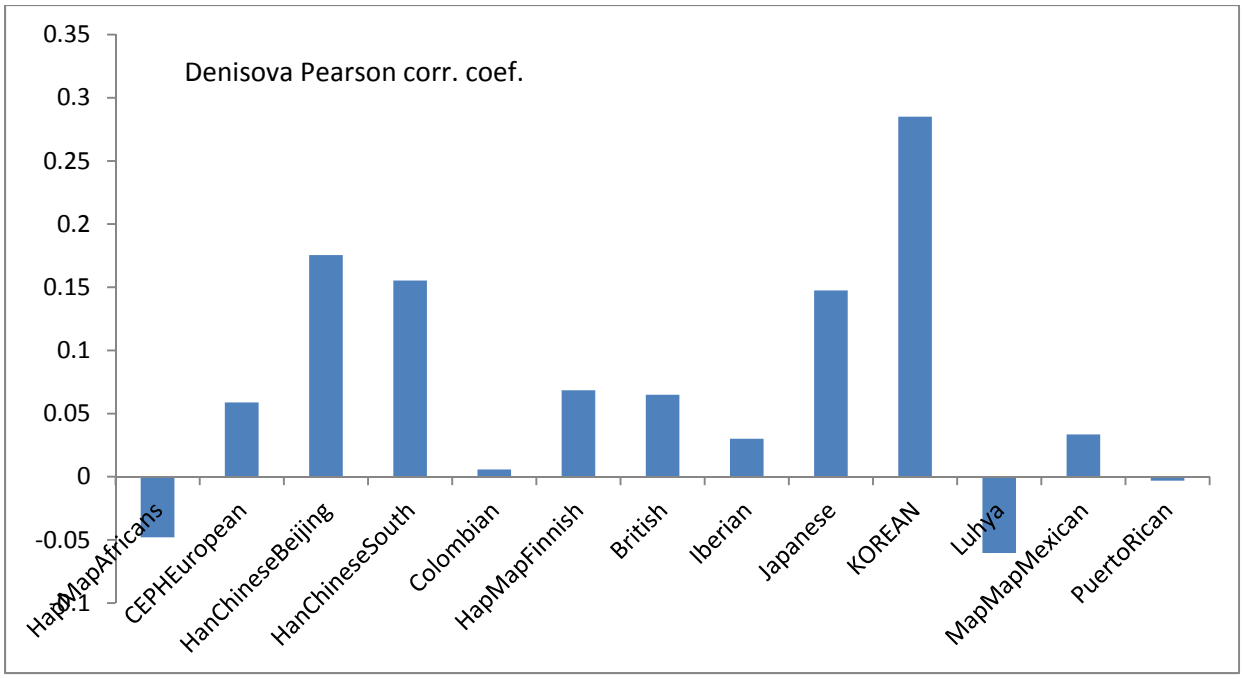

Figure 6: Persons correlation coefficient between Denisovian SNVs and subpopulations across all haplotypes of chromosome 1.

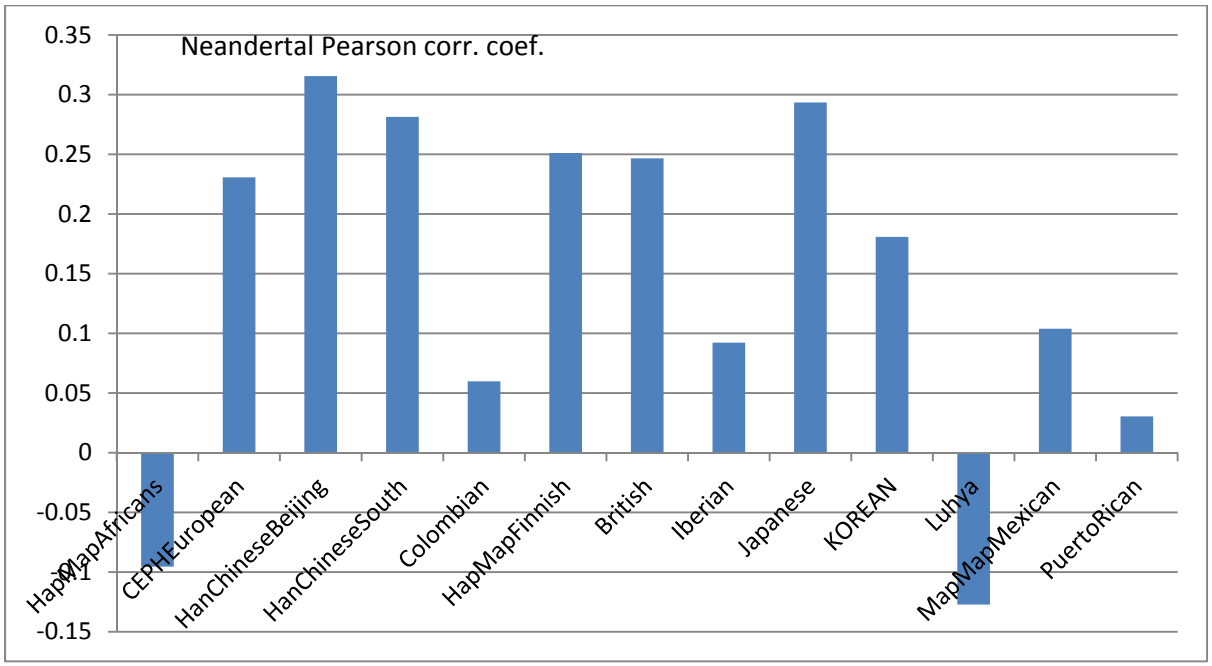

Figure 7: Persons correlation coefficient between Neandertal SNVs and subpopulations across all haplotypes of chromosome 1.