

Identification of Short and Rare Haplotype Clusters in Korean Genomes

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

Institute of Bioinformatics, Johannes Kepler University Linz, Austria

We developed HapFABIA to identify rare haplotypes in large sequencing data by biclustering which combines LD information across individuals and IBD information along the chromosome. For biclustering large data sets, we developed a sparse matrix algebra which is implemented in HapFABIA. HapFABIA significantly outperformed IBD methods at detecting rare haplotypes on simulated genotype data with implanted rare haplotypes.

We used HapFABIA to extract rare haplotypes from sequencing data from the Korean Personal Genome Project (KPGP). The genotyping data from 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. 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.

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). In contrast to these results, the enrichment of Denisova haplotypes in Koreans (odds ratio 36.7 of Fisher's exact test) 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).