

Data Mining and Comparison of Measures of Informativeness for Ancestry in Admixture Mapping

Tesfaye B. Mersha

Section on Statistical Genetics, Department of Biostatistics, University of Alabama at Birmingham

ABSTRACT

Given the huge amount of single nucleotide polymorphism (SNP) data available from high-throughput sources such as HapMap, data mining is a reasonable approach to identify SNPs that are informative for genetic ancestry. The distribution and density of the SNPs across the genome of African and European populations were extensively investigated using three SNP databases of HapMap, Affymetrix and Illumina. We have exploited these resources by web mining the data available from each of these databases to prioritize potential candidate SNPs useful for admixture mapping. About 4 million SNPs were compared between Africans and Europeans using various measures of ancestry informativeness in use today *viz.* absolute allele frequency differences (δ), Fisher Information Content (FIC), Shannon Information Content (SIC), and Fixation Index (F_{ST}). Each method provides different sets of candidate ancestry informative markers (AIMs) within and across the databases. The selected SNPs represent valuable resources for admixture mapping studies. The overlap and non-overlap between selected AIMs by different measures of informativeness, and in the different platforms are discussed.