hzAnalyzer: Detection, quantification, and visualization of contiguous homozygosity in human populations from high-density genotyping datasets using R and Java

Todd A. Johnson[1,2,*], Yoshihito Niimura[2], Tatsuhiko Tsunoda[1]

1. Laboratory for Medical Informatics, Center for Genomic Medicine, RIKEN Yokohama Institute, Yokohama, Kanagawa-ken, JAPAN
2. Department of Bioinformatics, Medical Research Institute, Tokyo Medical and Dental University, Tokyo, JAPAN

* Contact author: tjohnson@src.riken.jp

Keywords: genetics, polymorphisms, homozygosity, linkage disequilibrium

Research since the initiation and completion of the Human Genome Project [(4)] and its follow-up, the International HapMap Project [(1, 2)], has shown that much of human genetic variation is non-randomly organized into regions of restricted diversity in which a limited range of haplotypes can be observed. Such non-random partitioning of variation has been especially important for the design and performance of genome-wide association studies (GWAS), in which genetic variation between case and control samples are examined for associations with human diseases. In contrast to such regions where diversity is locally decreased across a population, recent reports have shown that individuals exist even in generally outbred human populations who possess extended regions of homozygosity, in which both large regions or even complete chromosomes apparently carry the same genetic variation on both chromosomes[(2, 3)]. Taken together, the locally restricted patterns that can be seen across populations and the long homozygous segments that can be seen within single individuals likely represent the extremes of a spectrum of relatedness that can be observed between individuals within human populations[(6)].

To analyze how the extent of contiguous homozygosity in high density single-nucleotide polymorphism (SNP) datasets varies within and between populations at genome-wide, chromosomal, and locally defined levels, we developed hzAnalyzer, a suite of programs using R and Java. This program uses rJava[(5)] to integrate Java code within our R programs, the combination of which allowed for a synergy with R contributing its ready-made statistical components, ease of scripting, and quick proto-typing and Java contributing enhanced performance in dealing with large datasets and an object-oriented construction that allowed us to represent some of the complexity inherent in population genetic data such as the fact that our sample population consisted of data subsets such as sample populations, families, and individuals. The functions making up hzAnalyzer can be broken down into three categories: 1) Homozygous segment detection and processing, 2) Quantification of variation in the extent of contiguous homozygosity within individuals and populations at different resolutions (i.e. genome, chromosome, local chromosomal regions), and 3) Visualization of raw segment positions and summarized/aggregated data. Our presentation will provide details about the functions that make up hzAnalyzer as well as figures using real human genotyping data from the International HapMap Project.

References

Altshuler D *et al* (2005). A haplotype map of the human genome. *Nature*, 437**,** 1299-1320.

Frazer KA *et al* (2007). A second generation human haplotype map of over 3.1 million SNPs. *Nature*, 449**,** 851-861.

Gibson J *et al* (2006). Extended tracts of homozygosity in outbred human populations. *Hum Mol Genet*, 15**,** 789-795.

Lander ES *et al* (2001). Initial sequencing and analysis of the human genome. *Nature*, 409**,** 860-921.

Urbanek S (2008). *rJava: Low-level R to Java interface,*
http://www.rforge.net/rJava/.

Weir BS *et al* (2006). Genetic relatedness analysis: modern data and new challenges. *Nat Rev Genet*, 7**,** 771-780.