

Homework 1 : Linkage Disequilibrium

Analysis of Linkage Disequilibrium (LD) of the *Lactase* (*LCT*) Gene in African (AFR), European (EUR), and Asian (ASN) Populations Using 1000 Genomes Phase 1 Data

LCT Gene Location (GRCh37/hg19):

chr2: 136,545,420 – 136,594,754

Analysis Steps:

1. Download Data from 1000 Genomes Project FTP.
2. Convert PED to BED format.
3. Filter and extract three subgroups: AFR, EUR, and ASN.
4. Perform variant quality control (QC) for each subgroup:
 - Minor Allele Frequency ($MAF > 0.05$)
 - Hardy-Weinberg Equilibrium (HWE $p\text{-value} > 10^{-3}$)
 - Retain only SNPs
5. Identify the intersection of SNPs, keeping only those that pass QC in all three subgroups.
6. Calculate pairwise Linkage Disequilibrium (LD).
7. Generate LD plots and provide interpretations.

Note: Lactase gene on human chromosome 2 is important because a mutation in the promoter, found at high frequencies in European populations, results in the ability to digest lactose in milk persisting into adulthood; in other populations this ability ceases between 5 and 10 years of age (Swallow, 2003). The high frequency of this mutation in Europe is thought to have been the result of strong selection for lactose persistence associated with the innovation of dairy farming. (From *Handbook of Statistical Genomics*, page 54)