Documentation

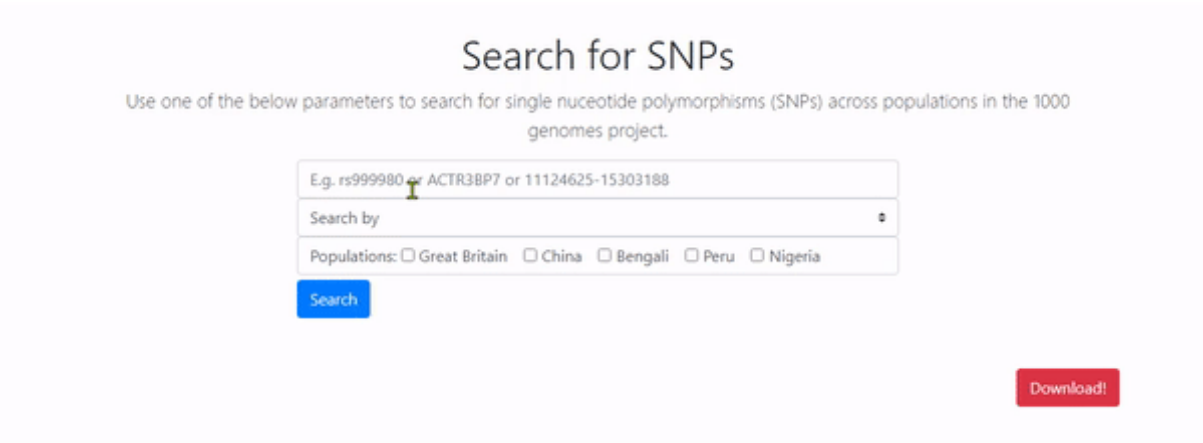
Methodology

Aim

The SNPip webserver provides a broad database containing SNP information from chromosome 22. The main functions of SNPip relies on basic data search by chromosomal location, gene name and SNP name for a subset of 5 populations. The search query outputs basic statistics for allele frequency and genotype frequency for every subpopulation. Additionally, if multiple SNPs are selected it contains summary statistics for genetic diversity, haplotype diversity and Tajimas D. If multiple populations are selected it contains statistics of pairwise genetic distance and delta Tajimas-D.

Search by:

- Gene annotation
- SNP annotation
- Reference and Alternate allele



Output:

- Alternate allele frequency
- Genotype frequency
- Shannon Diversity
- Haplotype Diversity
- Tajimas-D
- Delta Tajimas-D between two populations

Data

SNPip uses 1000 Genomes Project Phase 3 variants and also its definition of the super populations of European, East Asian, Southeast Asian, South American and African. Super populations are defined using the 1000G panel file.
The database contains biallelic SNPs derived from pre-processing the phase 3 genotype data. SNPip contains only single nucleotide variants (SNPs). Indels and larger structural variants have been removed.
The data only contains high confidence calls for ALT alleles and positions that have passed all filters.
SNPip contains SNPs located on chromosomes 22 (February 2022)

All the 1000G variants has been pre-processed with bcftools before creating the database using the following criteria:

1. Remove SNPs with duplicate rsIDs.
2. Remove Indels, and multiallelic sites
3. Run PLINK variant filtering
4. Remove SNPs with duplicate chromosomal coordinates.

Population and Sub-population

SNPip super population definitions are based on the following 1000G cohorts:

- European (EUR)
  - **British in England and Scotland (GBR)**
- East Asian (EAS)
  - **Han Chinese in Beijing, China (CHB)**
- Africa (AFR)
  - **Esan in Nigeria (ESN)**
- Southeast Asian (SAS)
  - **Bengali in Bangladesh (BEB)**
- American Ancestry (AMR)
  - **Peruvian in Lima, Peru (PEL)**

Statistics

Shannon's Diversity index

Shannon diversity index (SDI) is widely used in ecology to estimate species diversity and can be applied to population genetic studies. SNpip calculates genetic diversity using SDI; the allele frequency for a SNP variant in each population is used to estimate the genetic diversity between each population.

$H = -\sum[(pi) * \ln(pi)]$

- H = Shannon Diversity index
- pi = allele frequency(af) where the proportion of individuals with a particular alternate allele(n) found divided by the total number of individual alleles(N) found.
    - pi = n/N or af = al/ an
    - al = allele count an = allele number
- ∑= sum of the calculation
- ln = natural logarithm

SDI calculated using a python algorithm with the SDI formula:

1. Select allele frequency for the SNP variant for each population.
2. For each population, multiply the proportion by the logarithm of the proportion.
3. Sum all the numbers from step 2.
4. Multiply the sum by -1.

Output files

- Hudson_FST.csv
    - If 2 or more populations are selected during search there will be statistics otherwise will be an empty file.
- Tajima_D.csv
- Shannon_Diversity.csv
- Haplotype_Diversity.csv