**I. Projects on genomic data analysis**

**Project 1.**

The 1000 Genomes Project (1000GP):

The 1000GP released a human genetic variation catalog including single nucleotide variants (SNVs), short insertions and deletions (INDELs) and structure variants (SVs) based on sequencing of 2,504 genomes from 26 populations. Read the following paper and answer the following questions:

1. What techniques has been used to sequence these genomes, mapped the sequencing reads, called, genotyped and phased the genetic variants of various types?
2. What data formats are used to represent the aligned reads and genetic variants?
3. What are the most frequent SNVs, INDELs, and SVs in the African populations? Note you can write code to parse out the VCF files to gather this information.

[1] The 1000 Genomes Project Consortium (2015), “A global reference for human genetic variation”, Nature volume 526, pages 68–74.

<https://www.nature.com/articles/nature15393>

[2] The 1000 Genomes Project data: <http://www.internationalgenome.org/data/>

**Project 2.**

Genotype associations and genome-wide association studies (GWAS):

1. Using GWAS catalog [1], find the genetic variations associated with eye color [2] reported in published GWAS studies.

2. Using the 1000GP data, find the allele frequency of genetic variation associated with eye color [1,2] in African, European and Asian populations.

3. Using openSNP data [3], find the allele frequency of genetic variation associated with eye color.

[1] <https://www.ebi.ac.uk/gwas/>

[2] <https://www.snpedia.com/index.php/Eye_color>

[3] <https://opensnp.org/>

**Project 3.**

GWAS meta-analysis:

1. Using GWAS catalog [1], find the genetic variations associated with human height reported in published GWAS studies.

2. Using the Genetic Investigation of ANthropometric Traits (GIANT) consortium data and publications [2], find the allele frequency of all genetic variants associated with height in European populations.

[1] <https://www.ebi.ac.uk/gwas/>

[2] <https://portals.broadinstitute.org/collaboration/giant/index.php/GIANT_consortium>

**Project 4.**

Expression quantitative trait locus (eQTL) analysis project:

The GEUVADIS [1] project has sequenced the mRNA of lymphoblastoid cell lines of 465 individuals covering five populations in the 1000 Genomes Project. SNV eQTL analysis has been performed to assess how SNV genotypes are associated with gene expression. Read the papers and browse the data and answer the following questions.

1. What are the most significant SNV that’s assocociated with gene expression changes as reported in the GEUVADIS publication [2] and what gene is that?
2. Elaborate the workflow for the eQTL analysis conducted in the GEUVADIS publication [2].
3. Using the same strategy, perform the eQTL analysis using the SV genotypes provided in the 1000GP [3,4,5] and gene expression quantifications provided in the GEUVADIS [1,2] to assess how SVs are associated with gene expression changes.

[1] <https://www.ebi.ac.uk/Tools/geuvadis-das/>

[2] Lappalainen et al. (2013), “Transcriptome and genome sequencing uncovers functional variation in humans”, Nature 501, pages 506–511. <https://www.nature.com/articles/nature12531>

[3] The 1000 Genomes Project Consortium (2015), “A global reference for human genetic variation”, Nature volume 526, pages 68–74. <https://www.nature.com/articles/nature15393>

[4] The 1000 Genomes Project data: <http://www.internationalgenome.org/data/>

[5] Sudmant et al. (2015), “An integrated map of structural variation in 2,504 human genomes”, Nature, 2015. 562(7571): 75-81.

**Project 5.**

Epistasis analysis:

Instead of assessing individual genetic variants’ effect in classical GWAS and eQTL analysis, epistasis

epistasis analysis investigates the joint of non-linear effect of two or multiple genetic variants on a particular phenotype or trait. Through epistasis analysis, we can dissect the complex genetic architecture underlying human quantitative traits and complex diseases.

1. Using the pipeline in the paper [1], conduct epistasis analysis to see how two SNPs are associated with a particular trait in yeast using the yeast data [1].
2. Using the The Cancer Genome Atlas (TCGA) [2] and the pipeline we developed for epistasis analysis [3], find two microRNAs with joint effect on pathological stages of a particular cancer type.

[1] Bloom et al. (2015), “Genetic interactions contribute less than additive effects to quantitative trait variation in yeast”, Nat Commun. 2015; 6: 8712.

<https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4635962/>

[2] The Cancer Genome Atlas (TCGA).

<https://tcga-data.nci.nih.gov/>.

[3] Wen J et al. (2017), "An Empirical Bayesian Elastic Nets Method for Epistasis Analysis of microRNAs on Pathological Stages in Colon Cancer", BMC Genomics 18 (Suppl 7):756, 2017.

<https://github.com/shilab/EBEN-epistasis/>

<https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5657052/>