Fast Genome-Wide Association Study Visualization with C++

Project Description: In Genome-Wide Association Study (GWAS), one of the staple outputs is a p-value for each single nucleotide mutation (SNP) that indicates the significance of the association between that SNP and the trait being studied. A modern GWAS study with whole genome sequencing can have several million, or more, SNPs pass quality control and these are frequently post-processed and then visualized in the form of a 'Manhattan Plot'. This is a plot with the chromosomal position of the SNPs on the 'x' axis and the –log10 of the p-value plotted on the 'y' axis. With the advent of biobank cohorts with many measured traits, there is also a need to produce a sort of summary Manhattan Plot across the traits, which there can be thousands of. This leads to the need to plot hundreds of millions of points. To reduce this burden it has become standard to 'thin' (essentially sampling) the less significant SNPs as they are uninteresting for downstream analysis. Even with the sampling, the overall process is slow. The project here is to refine a series of steps for pre- and post-processing in a mix of R, C++, and Bash scripts into a single, fast, and flexible C++ program plotting with SFML 2.4 and integrating the PlotGenCpp (STBimage included) and RichText for SFML2 libraries to support creating plots very quickly with configurable labelling of top SNPs (positions need to be matched with nearest genes, filtered to not have too many overlapping, only 'lead' SNPs, genes in Italic other parts of the label not, etc...) to be near publication ready and output as a high resolution png embedded into an SVG so that final tweaks (e.g. font, legend placement, etc...) can be made easily in an SVG editor (e.g. Inkscape) for publication. Stretch goals: potentially supporting other relevant plot types (like LocusZoom style plots or mirrored Manhattan plots where the negative portion of the 'y' axis is used to plot a second trait), update everything to work with SFML 3.

Project Type: Research or Engineering

Internship Batch:

- **Batch 1:** May 10 to July 9, suitable for Education City students, i.e., CMUQ, TAMUQ and HBKU students
- **Batch 2:** May 31 to July 23, suitable for QU university students

Duties/Activities: Write and document C++ code to create the appropriate plots needed. Track new and updated code with Git. Write documentation of code function into comments, every function should explicitly state what it requires and modifies, and its effects. Meet daily to discuss progress.

Required Skills: C++ coding and debugging. Extras: Git, Valgrind, Make/Cmake, experience with writing portable code.

Preferred Intern Academic Level: Any

Learning Opportunities: This is a chance to learn about the downstream processing of genetic data and scientific programming. It is also will give some experience with SFML, which is frequently used for game development.

Expected Team Size: 2-4

Mentors

Name: Khalid Kunji email: kkunji@hbku.edu.qa