Reflective Report

Role Played

The primary responsibility included selecting appropriate populations (Gujarati Indians from Houston [GIH] and Punjabi from Lahore [PJL]) and relevant genomic datasets (VCFs, genetic maps, and curated T2D-associated SNPs) important for the positive selection analysis. These populations were chosen due to their known high prevalence of T2D and suitable sample sizes, ensuring statistical power and relevance to the project needs. Additionally, the role encompassed choosing the integrated haplotype score (iHS) method and the selscan software for identifying recent positive selection out of a wide pool of alternatives, given their reliability and wide usage within population genetics. Further responsibilities involved preprocessing steps, such as filtering VCFs to retain only population focused, biallelic SNPs and ensuring genetic maps were appropriately interpolated to match SNP positions precisely, important for running selscan for iHS. Finally running selscan on the preprocessed files as well as normalizing and interpreting the results.

Challenges Faced

Certain challenges were encountered throughout the project. Initially, a challenge was the lack of readily available, detailed workflows or guidelines specifically for positive selection analysis, resulting in considerable initial time taken to design an appropriate analytical workflow and gain confidence the choices made. Deciding on an optimal positive selection method, particularly the selection between selscan, SweepFinder, Tajima's D and others, required literature review and testing due to the initial unfamiliarity and uncertainty regarding their practical application. The decision to interpolate genetic maps to address density limitations was challenging and perhaps results in lessening biological relevance but necessary to prevent loss of key T2D-associated SNPs. Additionally, the minor allele frequency (MAF) filtering implemented by selscan excluded many relevant SNPs, primarily due to the highly curated nature of the T2D association data chosen, leading to concerns about sufficient data representation. Working under time constraints presented difficulties, leading to parallel analysis with alternative methods (Tajima's D and SweepFinder) as a contingency which were later not chosen as selscan finally worked, albeit later than expected. Furthermore, another challenge came due to the project's distribution of work, which initially limited the opportunity for team members to gain in-depth knowledge of each other's contributions. This separation sometimes led to difficulties, especially when integrating individual components into a unified team outcome. For instance, while the responsibility for generating and interpreting the positive selection

results fell to one member, other members managed separate aspects such as visualization. This division meant the insights necessary for producing accurate and biologically meaningful graphs were fairly isolated. Regular team meetings helped bridge those gaps.
Finally, the work entailed handling and processing large data files for which a common low spec computer can't handle very efficiently. This was improved on by implementing HPC for the workflow.

Technical Skills

The project enhanced certain technical skills involved in bioinformatics, particularly in genomic data manipulation, utilizing command-line tools such as bcftools for VCF file filtering, and Python libraries (pandas, NumPy, SciPy) for data preprocessing and interpolation. Better understanding of positive selection software/ workflows (particularly haplotype and allele frequency based methodologies) and how they work. Additionally, the project enhanced skills related to utilizing high-performance computing (HPC) platforms, particularly Apocrita, improving overall proficiency in working within such environments.

Transferable Skills

Growth in transferable skills included critical thinking, problem-solving, and strategic decision-making, particularly through resolving methodological challenges and workflow planning difficulties. Collaborative skills were notably improved through effective communication and coordination during regular meetings. Additionally, managing tight deadlines significantly developed effective time management, adaptability, and organizational capabilities crucial for future projects.

Reflective Insights

The contributions outlined above were crucial in shaping the project's scientific outcomes and what data the web-app displays, which fulfills the software's requirement of calculating and providing positive selection statistics for two distinct populations.