Bioinformatics Summer School Report

Name: Suliman Alraqebah
Internship Organization: Genetics & Genome Biology Program, The Hospital for Sick Children
Supervisor's Name: Lisa Strug
Project title: GWAS of cystic fibrosis lung disease using a variable number of tandem repeats called from long-read sequencing.

1. Introduction

Cystic fibrosis (CF) is an autosomal recessive disease that affects 70,000-100,000 individuals globally.[1] Variations in the cystic fibrosis transmembrane conductance regulator (CFTR) gene cause e the disease, which affects multiple organs, including the lungs.[2] A genome-wide association study (GWAS) identified five loci associated with lung function in CF, including a region on chromosome 5 near the solute carrier family 9 member A3 gene *(SLC9A3).*[2]This region contains numerous repeats, some expanding hundreds of base pairs, and it is unclear if these repeats contribute to the GWAS signal.[2] The **aim** of this study is to use HiFi technology from Pacific Biosciences (PacBio) to explore the association between repeat lengths genome-wide, particularly in the *SLC9A3* region, and lung function in CF patients.

2. Materials and Methodology

Subjects with CF are recruited into the Canadian Cystic Fibrosis Gene Modifier Study (CGMS) from 35 sites across 9 Canadian provinces. High-molecular-weight DNA from these individuals is sequenced by PacBio whole-genome long-read HiFi sequencing (30X). Tandem Repeat Genotyping Tool (TRGT)[3] and its repeat catalog (adotto_repeats.hg38.bed v0.0.2) is used to genotype repeats. GrafPop is used to infer subject ancestry from ~100,000 SNPs extracted using PLINK2.4.4 and plotted using Python.3.12.0. Principal Component Analysis (PCA) was conducted using PCAir to calculate PCs to include as covariates in the model. Tracy-Widom test was used to confirm significance in R.02.3. Subsequently, the associations of repeat lengths (average of the two haplotypes) among individuals with CF were tested with lung function, measured by Saknorm, a mortality-adjusted function of forced expiratory volume in 1 second (FEV1) in the European subset using linear regression in R.02.3.

3. Results and Discussion

GrafPop Results (Identifying Europeans)

This analysis included an initial cohort of 126 individuals with CF. GrafPop results showed that 93% of our sample are Europeans (N=117). Five, three, and one individual were from Latin American, South Asian, and African American ancestries, respectively.

Principal Component Analysis

The scree plot shows the eigenvalues for the first 32 PCs. The Tracy-Widon test suggested using the first 4 PCs, which we included as covariates in the model.

Association of VNTRs with lung function

The analysis examined a total of VNTRs across chromosomes 1-22, focusing on repeats with motifs greater than 6 base pairs. The genomic inflation factor (GC lambda) was calculated at 0.98, indicating no significant inflation for type I error. While no VNTRs reached genome-wide significance for association with SaKnorm ($p < 5E-8$), several VNTRs showed associations at a suggestive level ($p < 1E-5$)

***SLC9A3* Region.**

In the *SLC9A3* region, 36 VNTRs were identified. One VNTR (Chr5: 474056-474898) was nominally associated with SaKnorm (Beta = 0.002, SE = 0.0008, P = 0.04). Interestingly, this VNTR has been associated with *SLC9A3* expression in CF nasal epithelia.[2]

4. Skills and Knowledge Gained

- **Utilize the terminal shell to work with SLURM**, a workload manager, to handle computing jobs on high-performance computing (HPC) clusters.
- **Leveraged PLINK 2.4.4,** an enhanced version of the widely used, open-source whole-genome association analysis toolset, to perform faster Genome-Wide Association (GWA) calculations, including LD-based pruning and association analysis.[4]
- **Performed positions lifting from hg19 to hg38** using the UCSC LiftOver tool, which converts genome coordinates and annotation files from one assembly to another based on alignment.[5]
- **Gained proficiency in uploading data to LocusZoom,** a platform that enables the generation of interactive, shareable plots of GWAS summary statistics.[6]
- **Enhanced skills in statistical software,** including Python and R, for data analysis and visualization.
- **Developed stronger communication, teamwork,** and problem-solving skills through collaborative projects and departmental meetings.

5. Challenges and Solutions

Initially, troubleshooting memory errors and time limitations while working with large datasets on the HPC was challenging. However, with continuous guidance and support from the team, we successfully overcame these issues

6. Conclusion

My experience has been transformative. It has enhanced my technical skills, deepened my knowledge and skills in genomic data analysis, and confirmed my interest in bioinformatics. Key takeaways include the importance of collaboration, adaptability, and continuous learning.

7. Acknowledgments

I want to express my gratitude to KAUST Academy for their financial support. The GWAS of cystic fibrosis lung disease using a variable number of tandem repeats called from a long-read sequencing project is supported by the Data Sciences Institute, University of Toronto. I am deeply thankful to my supervisor, Dr. Lisa Strug, for her continuous guidance and support. I would also like to thank Strug's lab team members, especially Dr. Delnaz Roshandel, Shalvi Chirmade, Scott Mastromatteo, and all Graduate Students, for their generous knowledge and expertise.

8. References (if applicable)

1- Corvol, H., Blackman, S. M., Boëlle, P. Y., Gallins, P. J., Pace, R. G., Stonebraker, J. R., Accurso, F. J., Clement, A., Collaco, J. M., Dang, H., Dang, A. T., Franca, A., Gong, J., Guillot, L., Keenan, K., Li, W., Lin, F., Patrone, M. v., Raraigh, K. S., … Knowles, M. R. (2015). Genome-wide association meta-analysis identifies five modifier loci of lung disease severity in cystic fibrosis. *Nature Communications*, *6*, 23. https://doi.org/10.1038/NCOMMS9382

2- Roshandel, D., Mastromatteo, S., Wang, C., Gong, J., Thiruvahindrapuram, B., Sung, W. W. L., Wang, Z., Hamdan, O., Whitney, J., Panjwani, N., Lin, F., Keenan, K., Chen, A., Esmaeili, M., Halevy, A., Avolio, J., Ratjen, F., Celedón, J. C., Forno, E., … Strug, L. J. (2022). A cystic fibrosis lung disease modifier locus harbors tandem repeats associated with gene expression. *MedRxiv*, 2022.03.28.22272580. https://doi.org/10.1101/2022.03.28.22272580

3- Dolzhenko, E., English, A., Dashnow, H., de Sena Brandine, G., Mokveld, T., Rowell, W. J., Karniski, C., Kronenberg, Z., Danzi, M. C., Cheung, W. A., Bi, C., Farrow, E., Wenger, A., Chua, K. P., Martínez-Cerdeño, V., Bartley, T. D., Jin, P., Nelson, D. L., Zuchner, S., … Eberle, M. A. (2024). Characterization and visualization of tandem

4- *PLINK 2.0*. (n.d.). Retrieved August 13, 2024, from https://www.cog-genomics.org/plink/2.0/

5- *Lift Genome Annotations*. (n.d.). Retrieved August 13, 2024, from https://genome.ucsc.edu/cgi-bin/hgLiftOver

6- Boughton, A. P., Welch, R. P., Flickinger, M., Vandehaar, P., Taliun, D., Abecasis, G. R., & Boehnke, M. (2021). LocusZoom.js: interactive and embeddable visualization of genetic association study results. *Bioinformatics*, *37*(18), 3017–3018. https://doi.org/10.1093/BIOINFORMATICS/BTAB186