ColdBrew: Clustering on Genetics Relationship Matrix with High Variability Region

Presenters:
Hong-Sheng Lai, Shijie Tang, Xirui Liu, Zhiwen Bian, Hairuo Wang
Advisor: Dr. Ben Busby

[1]Ray and Stephanie Lane Computational Biology Department,
School of Computer Science, Carnegie Mellon University
[2]Department of Biological Sciences,
Mellon College of Science, Carnegie Mellon University

Outline

- Introduction
- Problem Statement
- Methods
  - Data Preparation
  - Genetic Relationship Matrix in GENESIS
  - Clustering
- Results
  - Allele Variability
  - PCA analysis
  - Clustering
  - Allele Variability
- Discussion
- Reference

Introduction

- Genetic Relationship Matrices (GRM) are crucial tools in genetic studies for quantifying genetic similarities between individuals, helping in understanding population structures, genetic diversity, and evolutionary relationships.
- Computational efficiency is essential for matrix calculation. Our goal is to develop or utilize an approach that can handle the high variability and compute GRMs effectively without extensive resource demands.
- Clustering GRMs allows us to detect population-specific patterns or genetic substructures, which may provide insights into population genetics, ancestry, and genetic susceptibility profiles for diseases.
- Investigating the influence of variability on GRM sparsity may reveal how genetic diversity impacts GRM properties, potentially affecting downstream analyses in genetic studies and applications in disease prediction or treatment personalization.

Problem Statement

- We select two regions of chromosome 6: HLA-A and TNF genes, both in 3000~3500 bp.
- Our goal is to use computationally efficient method to create a GRM for over 300 individuals from 3 populations.
- HLA regions have high variability.
- We want to find out does the variability affect the sparsity of the GRM.
- The significance of clustering the GRM across different populations.

Methods

Data Preparation

- Phased population VCF from 1kGP
  - British in England and Scotland (GBR) 107 samples
  - Puerto Rican in Puerto Rico (PUR) 150 samples
  - Chinese Dai in Xishuangbanna, China (CDX) 109 samples
- Use bcftools to extract the given region: HLA-A, TNF
- We use gene map information, and pick the most close blocks.

Genetic Relationship Matrix in GENESIS

Why GENESIS?

- Matrix calculation is slow, GENESIS use sparse, block-diagonal matrix storage and calculation.
- Implementation of **PC-AiR** Partition the sample into an “unrelated subset” and a “related subset”. Run PCA on the ‘unrelated subset’ and predict PC values for the “related subset”.
- Flexible data inputs (VCF, Plink)

Clustering

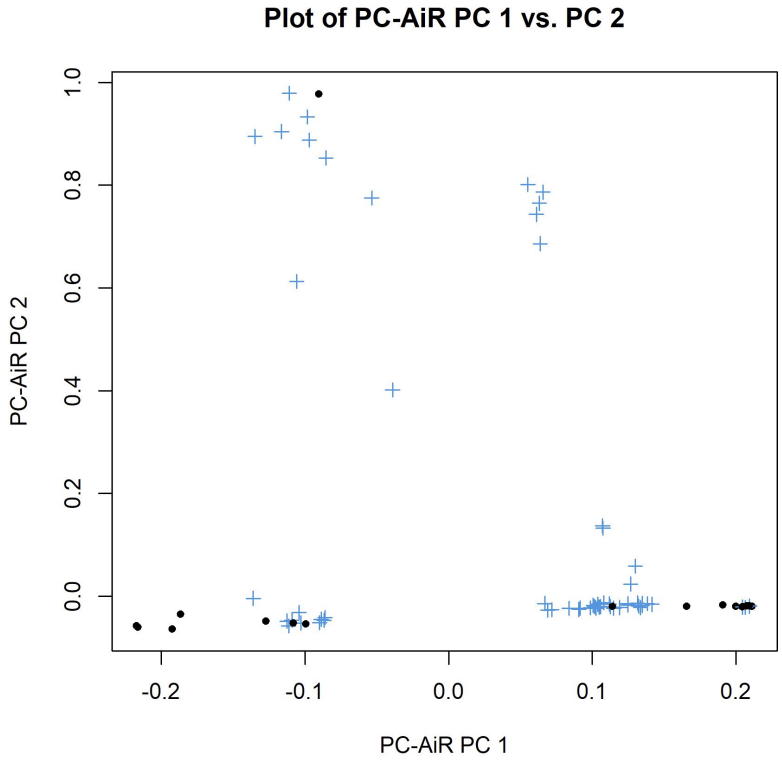
- We build the clustering based on the GRM matrix.
- Cluster can further explore the substructure of the input dataset and providing ancestry information.
- When we have the phenotype, we could further find the significance of each group.

Results

Allele Variability

- In this batch of aggregate VCF, we have 2548 samples.
- There are 270 variants (4,340 total possible alleles) in HLA-A.
- There are 82 variants in TNF.
- Both genes are in 3000~3500 bp.

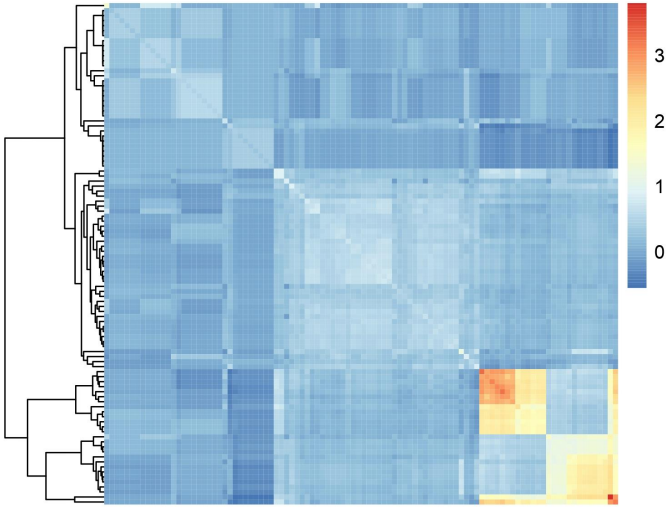
PCA analysis for PC-AiR



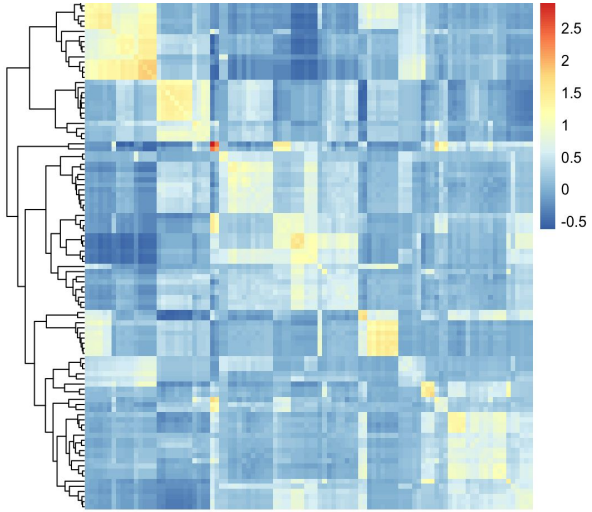
Gene: HLA-A
Population: CDX

● Unrelated Dataset
+ Related Dataset

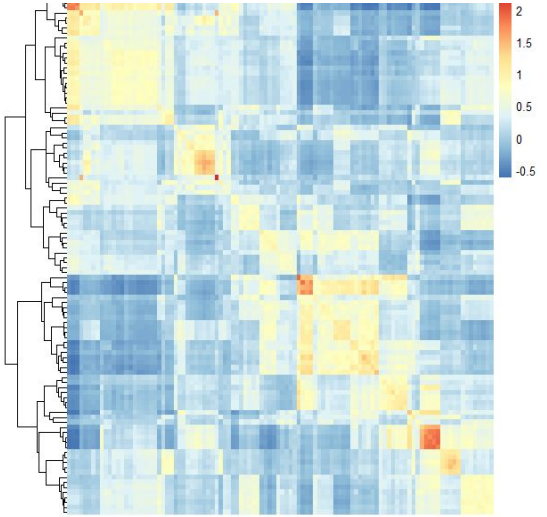
Clustering for GRM in HLA-A



Population: CDX

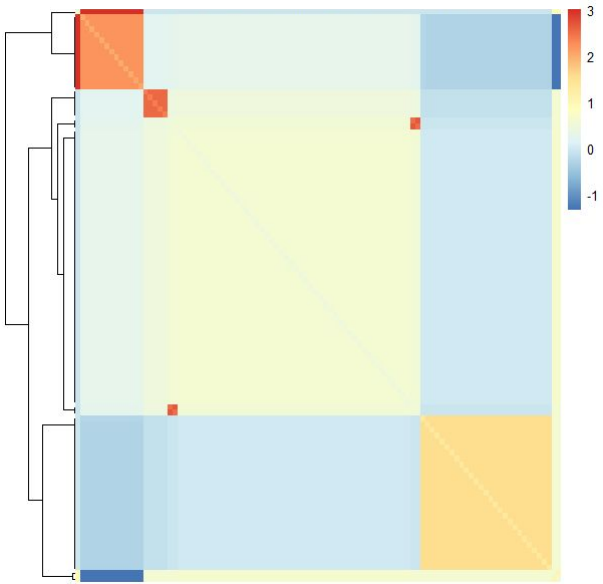


Population: GBR

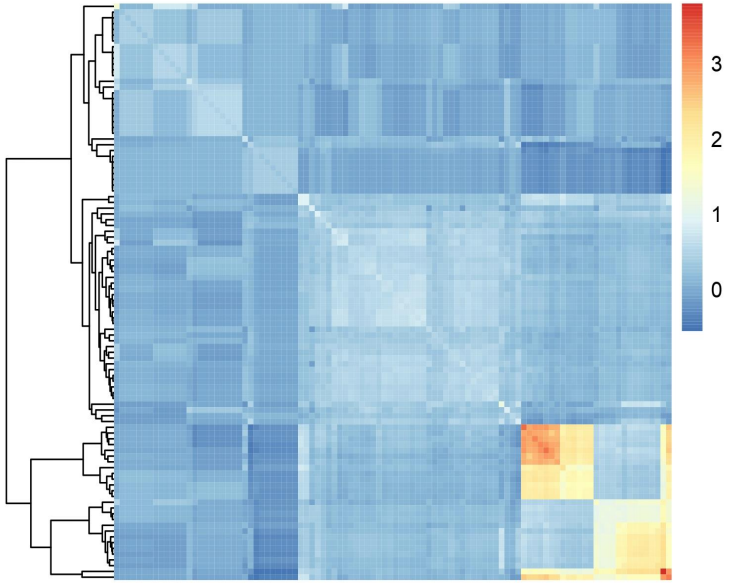


Population: PUR

Clustering for GRM from CDX in TNF and HLA-A



TNF



HLA-A

Allele Variability

- ~80% of the cells in GRM for TNF gene are NaN.
- High Variability leads to sparse matrix because there are too many possible alleles for each sample.
- However, for less variability region (like TNF), because it would have less variants in the same size of the window, it would not even have heterozygous calls to calculate the similarity.

Discussion

Discussion

- We can now speed up the sparse block-diagonal matrix calculation in GPU using Google JAX https://jax.readthedocs.io/en/latest/quickstart.html.
- After our procedure, if we have the phenotypes, we might have better inference when we have a new genotype results.
- Kinship estimation are based on the heterozygous calls. To prevent this, we might need larger variant sets (real haplotype blocks).

Reference

[1]. Link, V., et al., Tree-based QTL mapping with expected local genetic relatedness matrices. The American Journal of Human Genetics, 2023. 110(12): p. 2077-2091.

[2]. Gogarten, S.M., et al., Genetic association testing using the GENESIS R/Bioconductor package. Bioinformatics, 2019. 35(24): p. 5346-5348.

[3]. Conomos M.P., Miller M., & Thornton T. (2015). Robust Inference of Population Structure for Ancestry Prediction and Correction of Stratification in the Presence of Relatedness. Genetic Epidemiology, 39(4), 276-293.