HUMBOLD UNIVERSITÄT ZU BERLIN

Cross-Cohort Integration of Genetic Data for Disease Risk Analysis

Sedra Abou Ghaloun
Matrikelnr.: 633117

January 4, 2025

Studienproject Exposé

1 INTRODUCTION

Genetic variants account for much human phenotypic diversity, ranging from eye colour and height to disease risk. Environmental factors further shape these traits. A person's genotype is the sum of all the genetic material found in their cells, and differences in this genetic code are fundamental to understanding human diversity at the molecular level. A central goal of genetic research is identifying these variations and determining their relationship to specific phenotypes, including disease risk [4].

One application of machine learning (ML) in medicine is using genotyping data to predict disease risk. This involves training models on large datasets, such as the UK Biobank (UKB), which contains genetic and phenotypic data from half a million participants [5]. These models can then be used to predict disease risk based on a person's genetic profile.
While ML models have shown great promise, it's crucial to ensure they generalize beyond the population they were trained on. Although the UKB dataset is large and includes some diversity, it mainly consists of individuals of European ancestry. Models trained on such datasets may not perform as well when applied to populations with different demographics, environmental factors, or data characteristics [8].
Cross-cohort evaluations test whether a ML model has learned the underlying patterns rather than overfitting to cohort-specific biases. The All of Us Research Program provides a unique opportunity to address this limitation. With a mission to collect health data from more than one million people across the United States, All of Us emphasizes the inclusion of traditionally underrepresented populations in biomedical research [7]. This makes it an ideal test population for evaluating the generalizability of disease prediction models trained on the UKB dataset [1].

2 Scope of project

We currently have a workflow that applies feature selection (FS) methods on the UK Biobank (UKB) variants to generate UKB datasets with n selected variants. This workflow begins by sorting genetic variants based on their effect sizes, which are derived from PRS (Polygenic Risk Scores) [3] or GWAS (Genome-Wide Association Studies) [6]. These effect sizes are used to rank the variants. We apply different FS methods to select subsets of variants from this ranked list. One of the most effective FS methods we have employed so far selects the top variants with the highest effect sizes from each gene, selecting n variants per gene across all genes. We usually select 100,000 variants from a total of 93 million variants. After selecting the top 100,000 variants, the workflow generates BGEN files using PLINK2 [2]. These files contain genotype information for the selected variants from patient profiles.

The workflow converts this data into a NumPy array, where rows represent patients and columns correspond to the selected variants. This array serves as input for the ML models.
The goal of this project is to adapt this workflow to create new datasets from the All of Us (AoU) data that aligns with the structure, variant selection, and order of our existing UKB datasets.

3 Challenges

3.1 AoU Cloud-Based Platform

The genomic data in the All of Us (AoU) workbench is stored within a curated data repository (CDR) located in a cloud-based infrastructure (the Researcher Workbench) for analyzing genomic, clinical, and lifestyle data collected from diverse participants. While the underlying infrastructure uses Google Cloud Platform, the Workbench is designed specifically for scientific research and is provided as a non-commercial service to approved researchers. This setup introduces operational limitations. Only Jupyter notebooks are persistent within the workspaces, meaning that any temporary files, virtual environments, or intermediate results must be carefully managed to avoid data loss. This requires users to save all essential outputs explicitly to workbench buckets.
To access the data, users must first create a workspace within the AoU platform. Within this workspace, a cloud analysis environment must then be configured for performing computations, which incurs additional costs based on resource usage. Users are provided with free credits when they apply to the All of Us program.

3.2 Limited Memory Capacity

The disk space in the AoU workspace is limited and expensive, making it difficult to process large genomic datasets. To overcome this, intermediate results should be stored in the workbench bucket, and the data can be processed in smaller chunks, such as by chromosome, to manage storage more efficiently.

3.3 Variants Mapping

1. **Data Format**: AoU provides genomic data in BGEN files for the ACAF threshold callset, which includes variants with a population-specific allele frequency (AF) > 1% or allele count (AC) > 100. However, the data lacks reference SNP IDs (rsIDs), making direct variant mapping more complex.

2. **Shared Variants**: The variants in AoU do not fully overlap with those in the UKB, and vice versa. This requires careful selection and ordering of variants to ensure that they align with the UKB datasets. To align the AoU datasets with UKB ones, the first step is to ensure that only variants present in both the UKB and AoU datasets are selected. Approximately 17 million variants are shared between the two datasets.

3. **Genome Version**: AoU genomic data uses the GRCh38 (hg38) genome version, whereas UKB datasets might be based on GRCh37 (hg19). We need to map the variants using CrossMap [9].

4 Methodology and Workflow

First, an appropriate workspace will be created with defined persistent disk (PD) storage, and a controlled tier dataset will be set up to gain access to the genotyping data. Next, the UKB variants will be mapped from hg19 to hg38, and the UKB variants will be imported into the AoU cloud analysis environment to identify the intersection between the variants from the UKB dataset and the ACAF variants in the AoU dataset.
After identifying the shared variants, their IDs will be exported, and the genotyping feature selection (FS) workflow will be adapted to accommodate the intersection. Typically, 100,000 variants will be selected for the datasets. After performing feature selection on the intersecting variants and finalizing the 100,000 variants for genotyping datasets, these selected variants will be imported back into the AoU cloud analysis environment.
To evaluate the selected variant lists, the Jaccard index will be calculated to measure the similarity between different feature selection

methods, the number of covered genes will be compared, linkage disequilibrium (LD) patterns will be analyzed, and the variance within the datasets will be examined..
Additionally, the workflow will be modified to write genotyping datasets for the AoU cohorts using the BGEN files from the ACAF threshold callset and the 100,000 selected variants. Finally, demographic differences between the UKB and AoU datasets will be compared, focusing on key traits such as ancestry distribution, age, and gender. Since models will often be trained to predict disease risk for conditions like type 2 diabetes or coronary heart disease, the prevalence and distribution of these diseases within the datasets will also be compared.

REFERENCES

[1] Oscar Aguilar et al. “Integrative machine learning approaches for predicting disease risk using multi-omics data from the UK Biobank”. In: *bioRxiv* (2024).

[2] Christopher C Chang et al. “Second-generation PLINK: rising to the challenge of larger and richer datasets”. In: *Gigascience* 4.1 (2015), s13742–015.

[3] Samuel A Lambert et al. “The Polygenic Score Catalog as an open database for reproducibility and systematic evaluation”. In: *Nature Genetics* 53.4 (2021), pp. 420–425.

[4] Nazli G Rahim et al. “Genetic determinants of phenotypic diversity in humans”. eng. In: *Genome biology* 9.4 (2008), pp. 215–215. ISSN: 1465-6906.

[5] Cathie Sudlow et al. “UK biobank: an open access resource for identifying the causes of a wide range of complex diseases of middle and old age”. In: *PLoS medicine* 12.3 (2015), e1001779.

[6] Emil Uffelmann et al. “Genome-wide association studies”. In: *Nature Reviews Methods Primers* 1.1 (2021), p. 59.

[7] All of Us Research Program Investigators. “The “All of Us” research program”. In: *New England Journal of Medicine* 381.7 (2019), pp. 668–676.

[8] Erik Widen et al. “Machine learning prediction of biomarkers from SNPs and of disease risk from biomarkers in the UK Biobank”. In: *Genes* 12.7 (2021), p. 991.

[9] Hao Zhao et al. “CrossMap: a versatile tool for coordinate conversion between genome assemblies”. In: *Bioinformatics* 30.7 (2014), pp. 1006–1007.