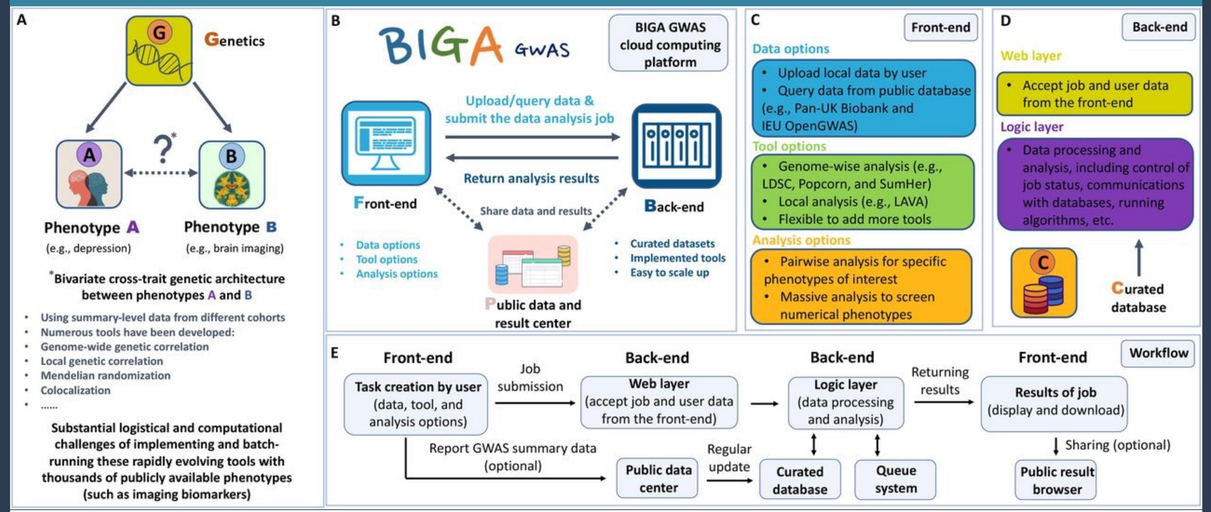
Building a Data Harmonization Pipeline for BIGA GWAS (Genome-Wide Association Studies)

Dr. Bingxin Zhao, Ethan Yu (Eng '26)
Wharton, Department of Statistics and Data Science

What is BIGA GWAS?



(A) The motivation of this project is to address the substantial logistical and computational challenges associated with implementing and batch-running the constantly evolving tools for cross-trait genetic architecture analysis. Our aim is to offer a cloud computing-based solution that can effectively overcome these challenges. (B) Overview of the BIGA GWAS platform. Users can easily upload their GWAS summary level data and submit data analysis jobs through the front-end interface. These jobs are then processed on the back-end, and the results are subsequently returned to the users. In addition, we have established public data and result centers where users can share their data and outcomes with the public. (C) The front-end interface of the BIGA GWAS platform offers users a comprehensive set of options to manage their data resources, choose the appropriate tools, and select the desired mode of data analysis. (D) Details of the back-end of the BIGA GWAS platform. (E) Overview of the analysis workflow.

Data Harmonization

Genome-wide association studies (GWAS) have proliferated rapidly, producing summary statistics on thousands of phenotypes. However, integrating and analyzing summary statistics across studies remains challenging due to inconsistencies in variant encoding, trait categories, file formats, and metadata reporting. For instance, allele codes like A/T versus 0/1 may differ, while same traits may be labeled differently across studies. These inconsistencies can lead to errors when combining datasets for meta-analysis or looking up variants across different studies. Manual curation and harmonization of variants and traits requires extensive effort and is error-prone. Automated harmonization pipelines are needed to enable rapid, accurate integration of GWAS summary statistics from diverse sources. By aligning summary statistics to a common data schema, data harmonization facilitates powerful cross-study analyses and enhances the utility of GWAS repositories. Implementing robust harmonization workflows will increase efficiency, reproducibility, and scientific insights as GWAS analyses continue to scale up in the era of big data.

Standardizing GWAS

Over a series of meetings from June 2020 to September 2021, the GWAS Catalog gathered input from diverse GWAS data generators, users, managers and bioinformaticians. Several key harmonization requirements emerged from the stakeholders' use cases and challenges:

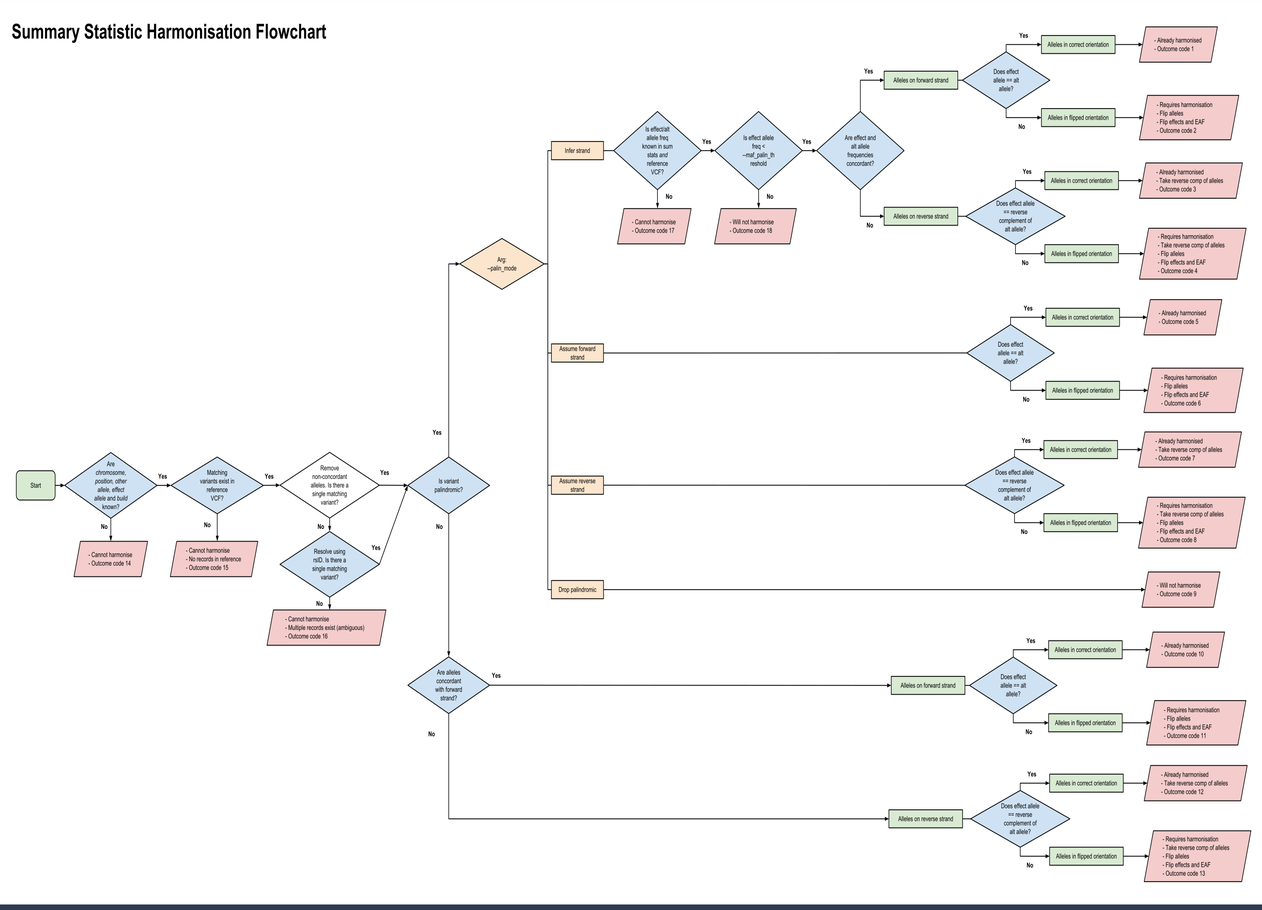
1) consistent data representation for interoperability;
2) accessible metadata to enable interpretation and reuse
3) unambiguous variant reporting for standard annotation
4) mandatory fields providing essential data for analyses like MR and PGS
5) 5) recommended standard headers to encourage harmonization
6) a balance of essential but achievable requirements to maximize uptake
7) low bioinformatics needs to accommodate the user community.

By directly engaging the GWAS community, this process underscored the critical need for harmonization to enable effective data sharing and cross-study analyses. Our harmonization pipeline aims to address the stakeholders' requirements to maximize downstream data utility.

Creating a Pipeline

The pipeline implements systematic harmonization of heterogeneous GWAS summary statistics to enable meta-analyses and polygenic scoring applications. It starts by parsing input files in various formats to extract key fields including variant identifiers, chromosomes, positions, alleles, effect sizes, and p-values. Variant IDs are converted to rsIDs mapped to the latest genome build using the Ensembl API; variants without valid rsID mappings are lifted over to the current build, while unmappable variants are dropped. For palindromic variants with ambiguous strand, allele orientation is inferred by sampling 10% of non-palindromic variants to determine forward/reverse strand consensus based on Ensembl VCF allele alignment. If consensus is unclear, all palindromes are dropped. Each variant is then harmonized by querying the Ensembl VCF to orient alleles to the forward strand, flip alleles and effect sizes if swapped from the reference, or replace alleles with NA when no match exists. Invalid rows missing required variant, chromosome, position or p-value data are filtered out. Final harmonized summary statistics are output in standard MINIMSS format with added harmonization codes indicating processing details. Tabix indexed files and detailed pipeline reports are generated.

Data Harmonization Pipeline (Orientation Alignment Portion)



Results

To address the need for scalable GWAS data harmonization, we developed an automated Python pipeline using the Pandas data analysis library. The pipeline will take summary statistics files in varied formats as input and outputs standardized CSV files compliant with the MINIMSS schema. Key processing steps include inspecting and parsing file headers, detecting data types, handling missing data, mapping variant identifiers to unique rsIDs, converting trait labels to EFO ontology terms, and applying schema mappings to homogenize columns. Custom Pandas functions enable efficient large-scale data transformations and validation checks. The modular design allows adding enhancements like support for additional file formats or integration of new ontologies. In initial tests, the pipeline successfully harmonized GWAS summary statistics from major repositories including the GWAS Catalog. By leveraging Python's extensive scientific computing libraries, this pipeline provides a robust and extensible data harmonization solution to unlock the full value of disparate GWAS resources

Follow-ups and Acknowledgements

The pipeline is not complete yet. The final step of MINIMSS output with processing detail metadata will be completed once the validity of the variant ID mapping, allele orientation, and filtering/QC is confirmed.

Code can be found at the following public BIGAGWAS github repository: https://github.com/ethayu/BIGAGWAS-Data-Harmonization/tree/main

This project was made possible under:
PURM
Penn Undergraduate Research Mentoring Program