

PP16 Integrating Open-Source in SAS Viya: A Bioinformatics Use Case

Samiul Haque (samiul.haque@sas.com), Stephan Weigandt, and Scott McClain
SAS Institute Inc.

ABSTRACT

Sequence processing is heavily dependent on various command line programming and open-source software resources. A divergent set of software and skillsets for these data workflows are difficult to govern and manage. In addition, modifications of the process for continuous improvements typically requires specific programming language expertise, creating processing delays. In this work, we present a contiguous polygenic scoring (PRS) solution workflow, all contained within the SAS platform. It provides flexibility for replicating, reproducing, and modifying any part of the analyses. The PRS build utilized SAS, R, Python, and Command Line code to ingest genotyping data, to visualize genome-wide association studies (GWAS), and to calculate multiple gene risk scoring for individuals with disease. The SAS PRS solution also allows investigators to query knowledge bases (e.g., Ensembl) in a no-code / low code approach. The SAS approach enables biologists, bioinformaticians, and data scientists to collaborate seamlessly in a common platform using multiple programming languages.

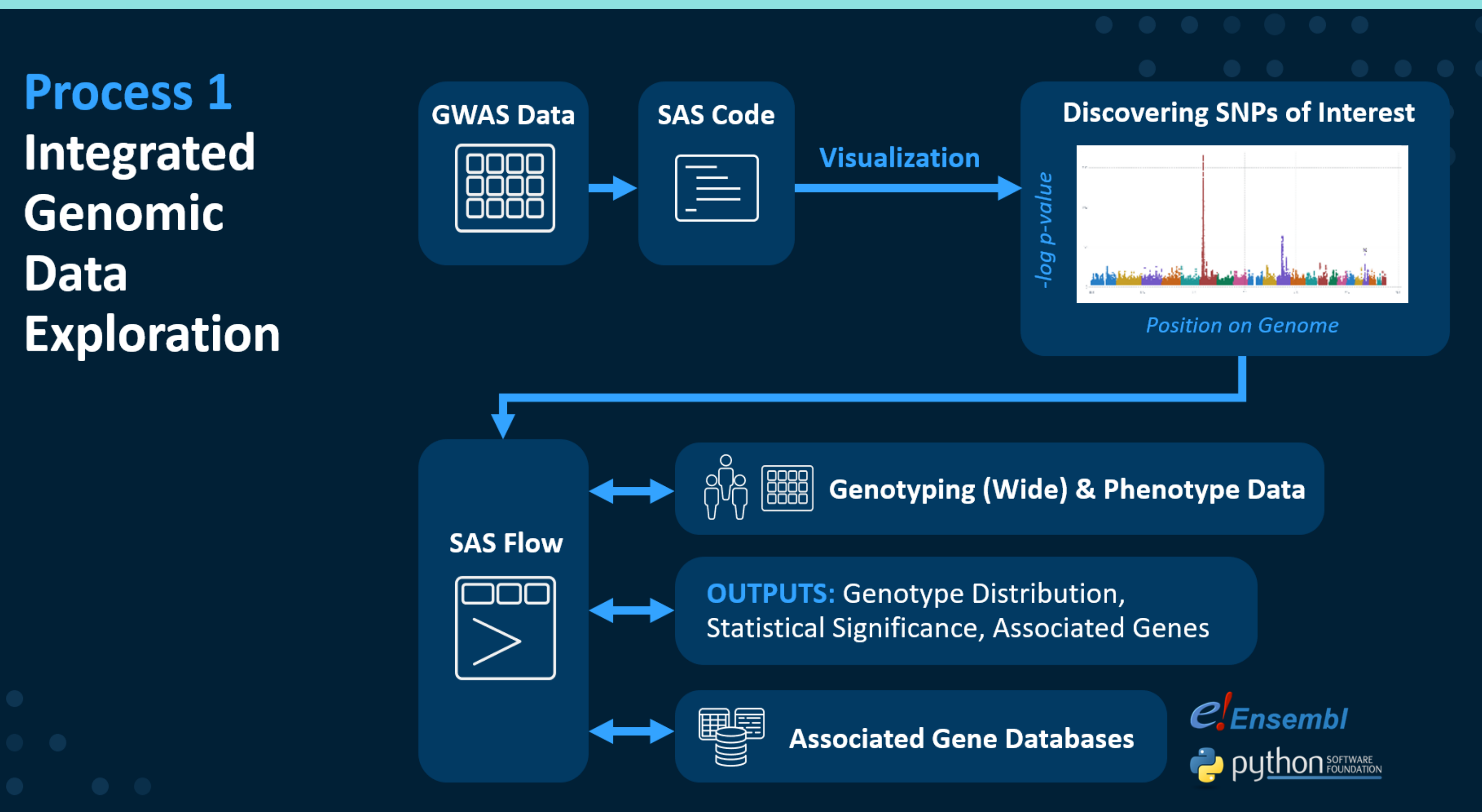
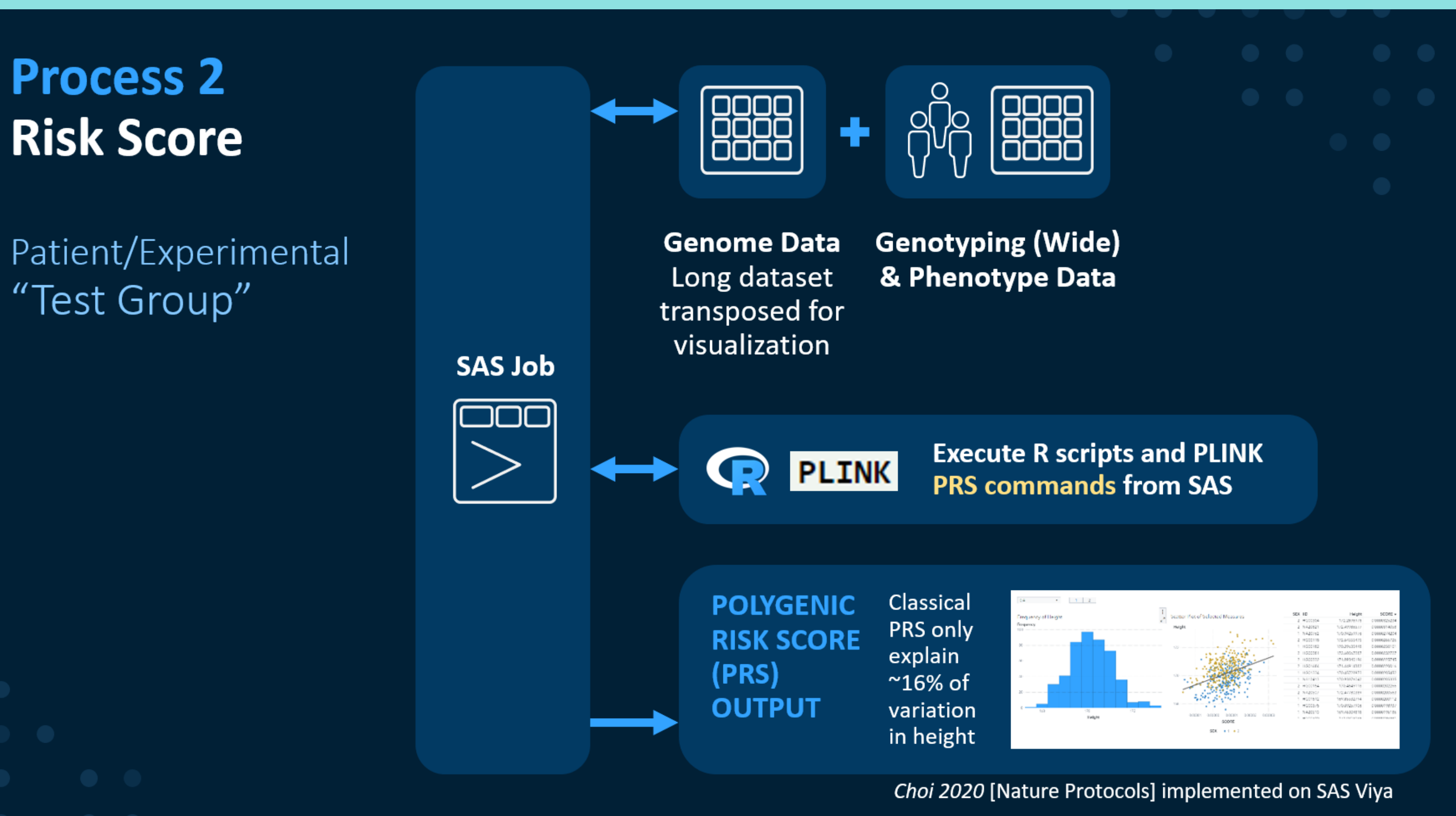
Objective

Develop a bioinformatics workflow that integrates different tools for processing genotyping data. This workflow has the following features

- Low code / no code interface for end user
- Interactive visualization through dashboard
- Integration of SAS, R, and Python in same workflow environment
- End users have ability to change different parameters
- Complete log of different pipeline runs

Use Case:

- We implemented a Polygenic Risk Score analysis approach originally proposed by Choi 2022 in Nature Protocols.
- We used GWAS data on simulated phenotype (height) and individual-level simulated genotype-phenotype from the 1000 Genomes Project. These datasets can be obtained from <https://choishingwan.github.io/>



Execute R, Python, and SAS code using SAS studio Flow

Standard GWAS QC

Page 1 Node Notes

Select Plink Execution

- ☒ Standard GWAS QC
- ☐ Remove highly correlated SNPs
- ☐ Compute Heterozygosity with Cleanup
- ☐ Perform Sex Check
- ☐ Relatedness
- ☐ Final QCed Target Data File

bfile

EUR

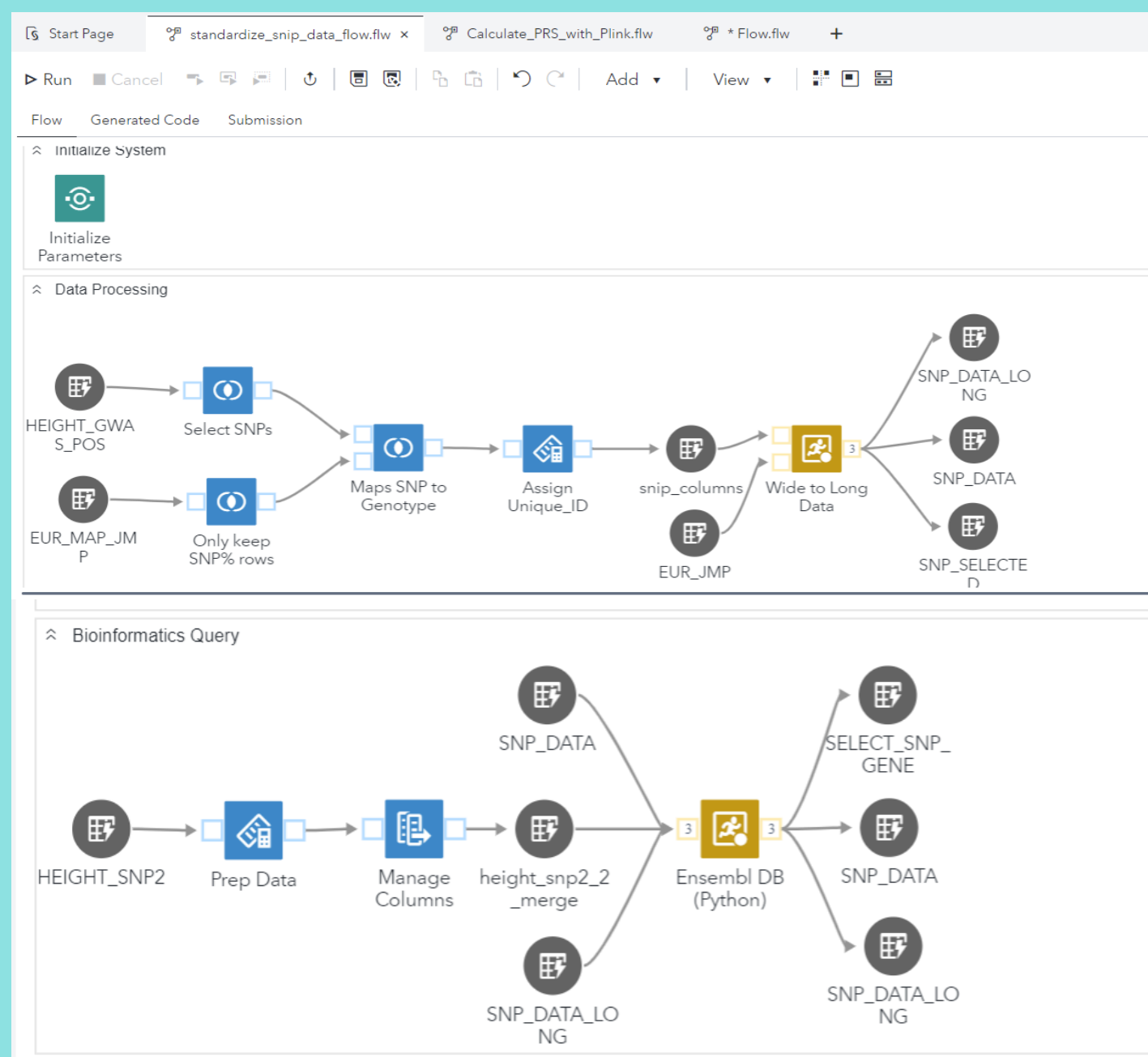
maf

0.01

hwe

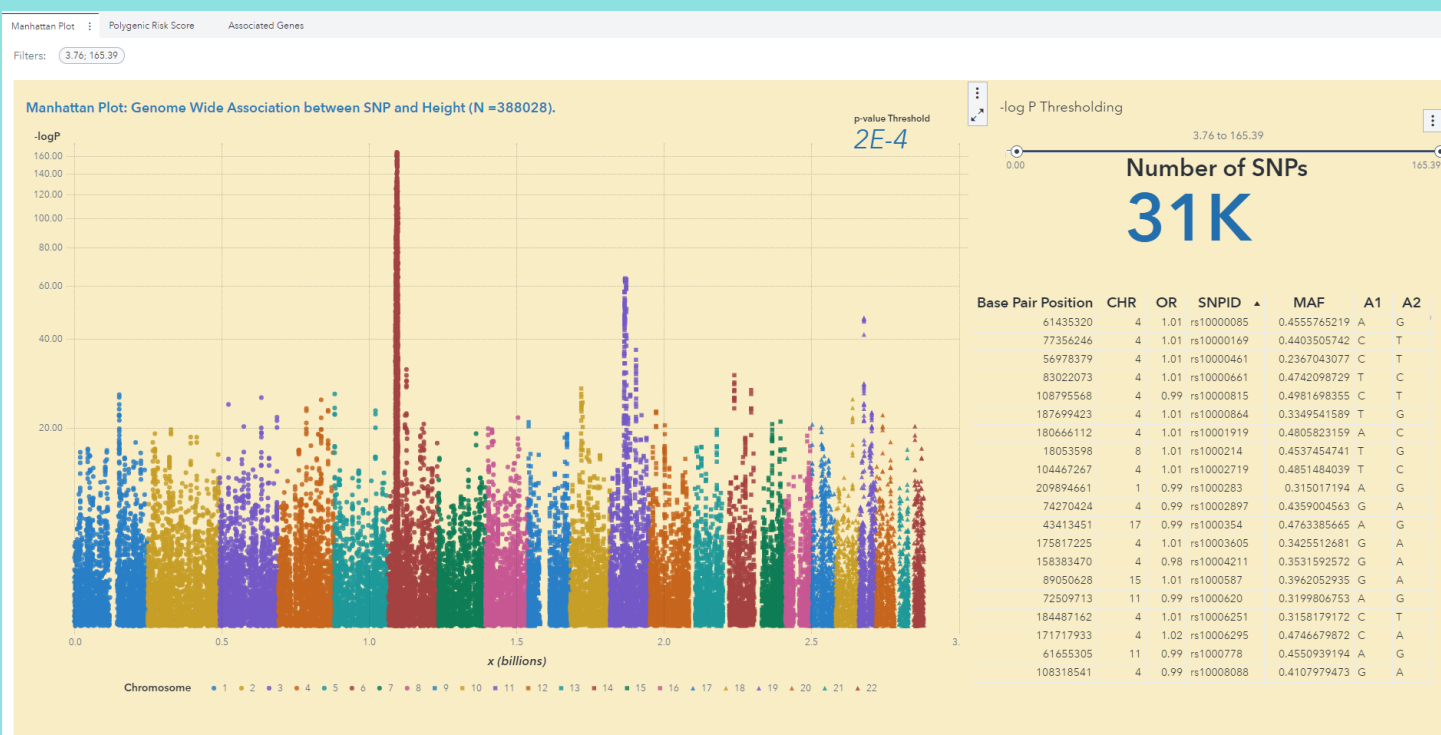
1e-6

SAS Studio Custom Node that allows user To run the workflow while changing important parameters.

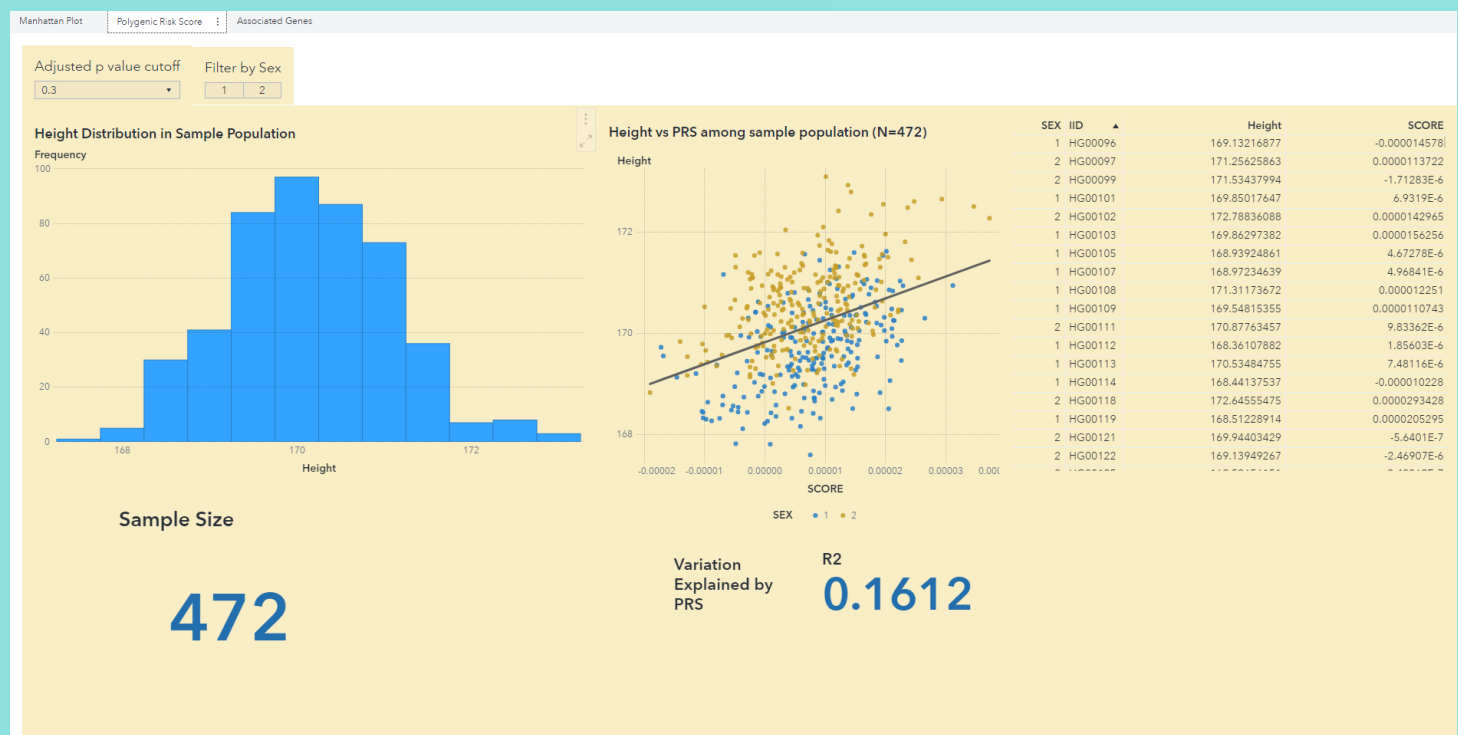


SAS Studio Flow representing parts of the bioinformatics workflow

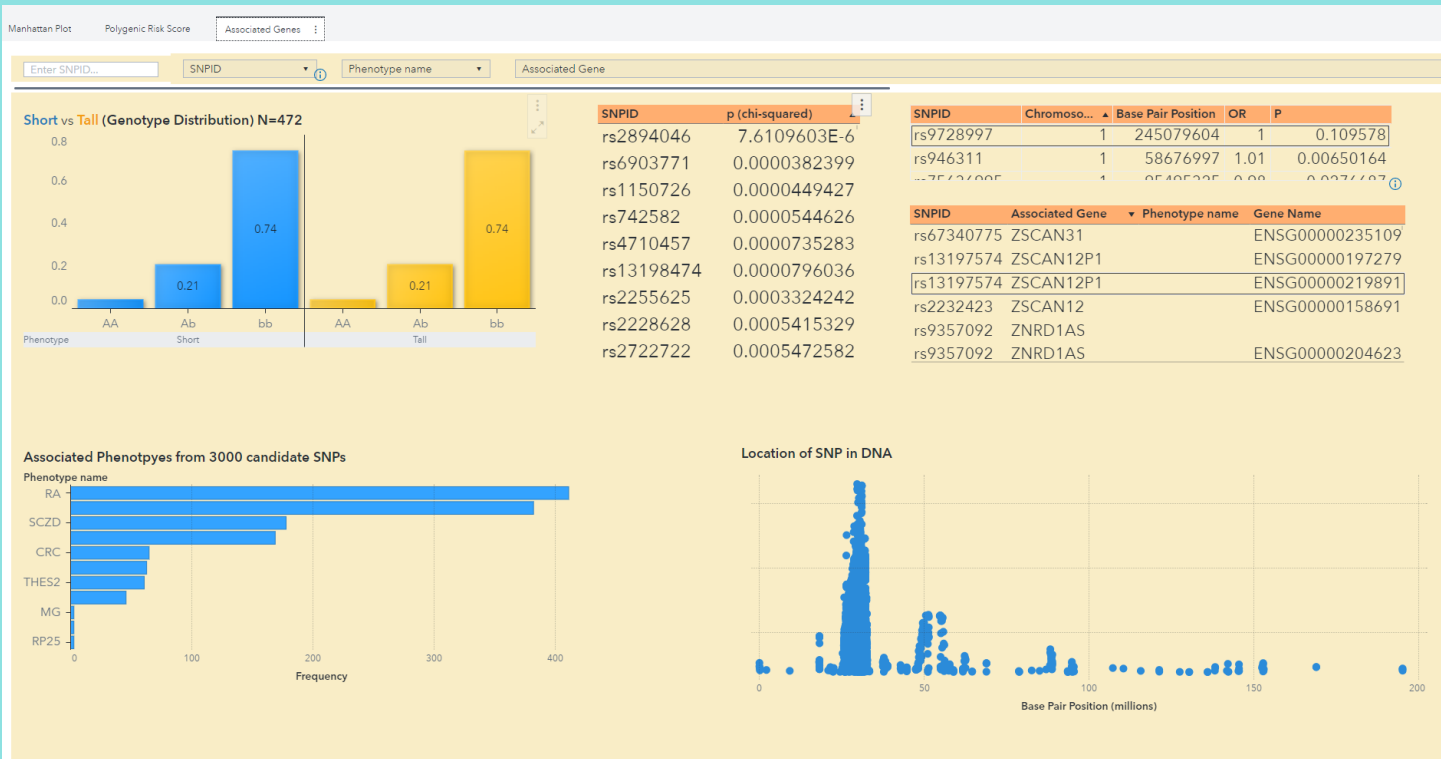
Explore Results on Interactive Dashboard



Manhattan plot showing association between SNP and height



Polygenic risk score associated with height



Genotype distribution across SNPs, associated genes, and phenotypes

References

- Choi, S.W., et al. Tutorial: a guide to performing polygenic risk score analyses. Nature protocols, 15(9), pp.2759-2772, 2020.
- Kinsella, R.J., et al.. Ensembl BioMart: a hub for data retrieval across taxonomic space. Database, 2011.
- Purcell, S., et al. PLINK: a tool set for whole-genome association and population-based linkage analyses. The American journal of human genetics, 81(3), pp.559-575., 2007

