

A pipeline to simulate Human variants with different genetic architectures

M Santorsola¹, S Ahmed¹, D Bagordo¹, S Carpanzano¹, E Franzoso¹, L Sola¹, F Lescai¹

¹Department of Biology and Biotechnology, University of Pavia, Italy

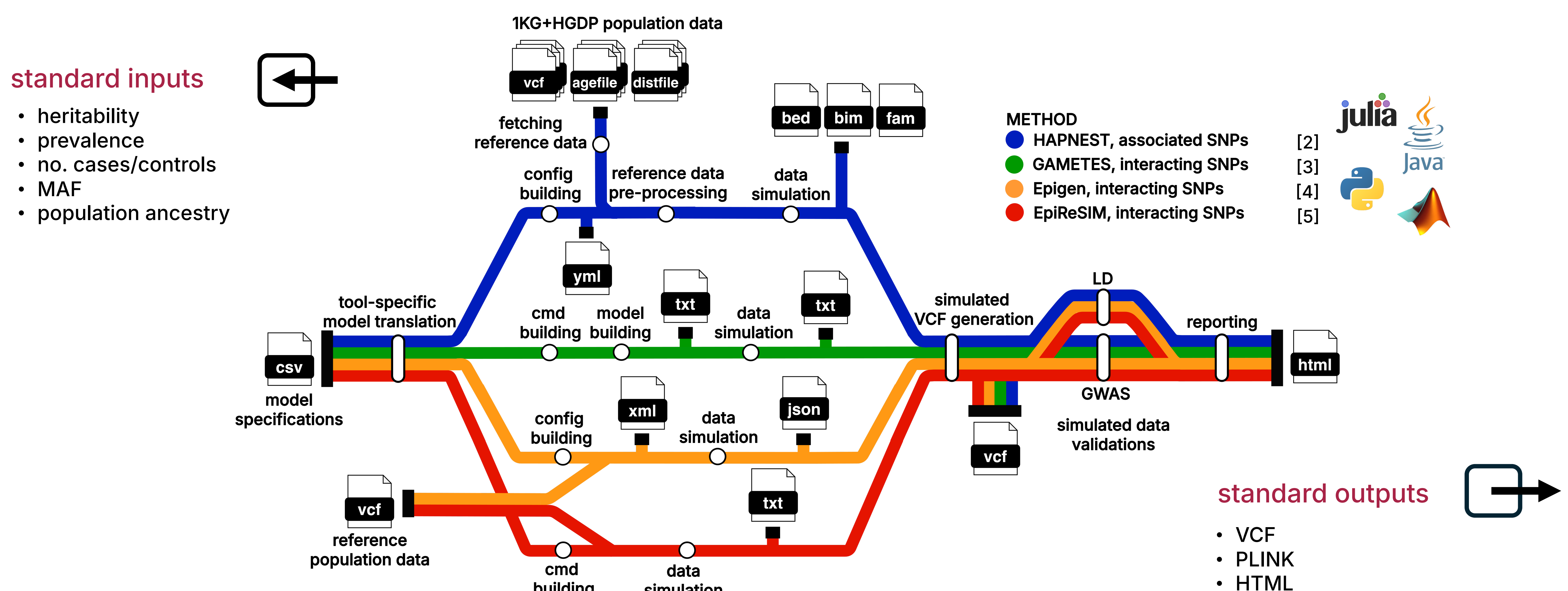
Simulating genetic data for complex traits requires standardizing existing tools

A considerable portion of the genetic basis for complex traits remains unaccounted for, known as **missing heritability** [1].

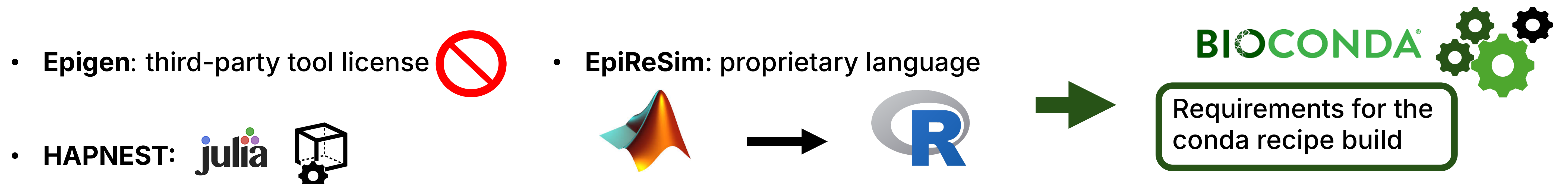
Simulated data, which closely mimic the real-world genetic architectures in Human populations, play a crucial role in developing new **statistical**, **bioinformatics**, and **deep learning** methods to address genetic complexities.

Tools for simulating data for specific study designs (e.g., case/control) and genetic architectures (e.g., rare or interacting variants) have a steep learning curve due to varying programming languages, unique model definitions, input formats, and parameter settings. An **automated** and **user-friendly pipeline** is needed to streamline **simulations**.

Pipeline schema to simulate Human variants



Challenges in the tool implementation



Simulated data closely mirror real-world genomic patterns

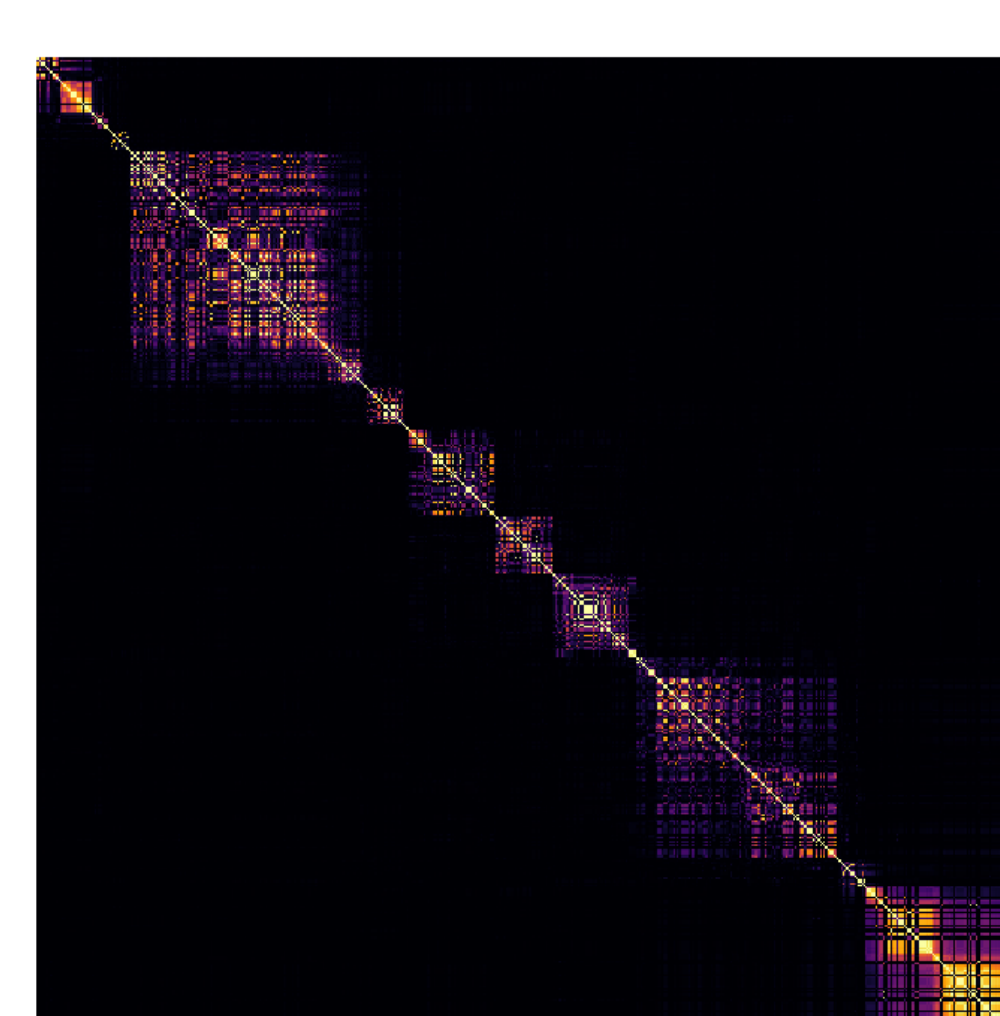
Pipeline validation goal:

Ensure simulated data preserve key Human genomic patterns by comparing real and simulated data

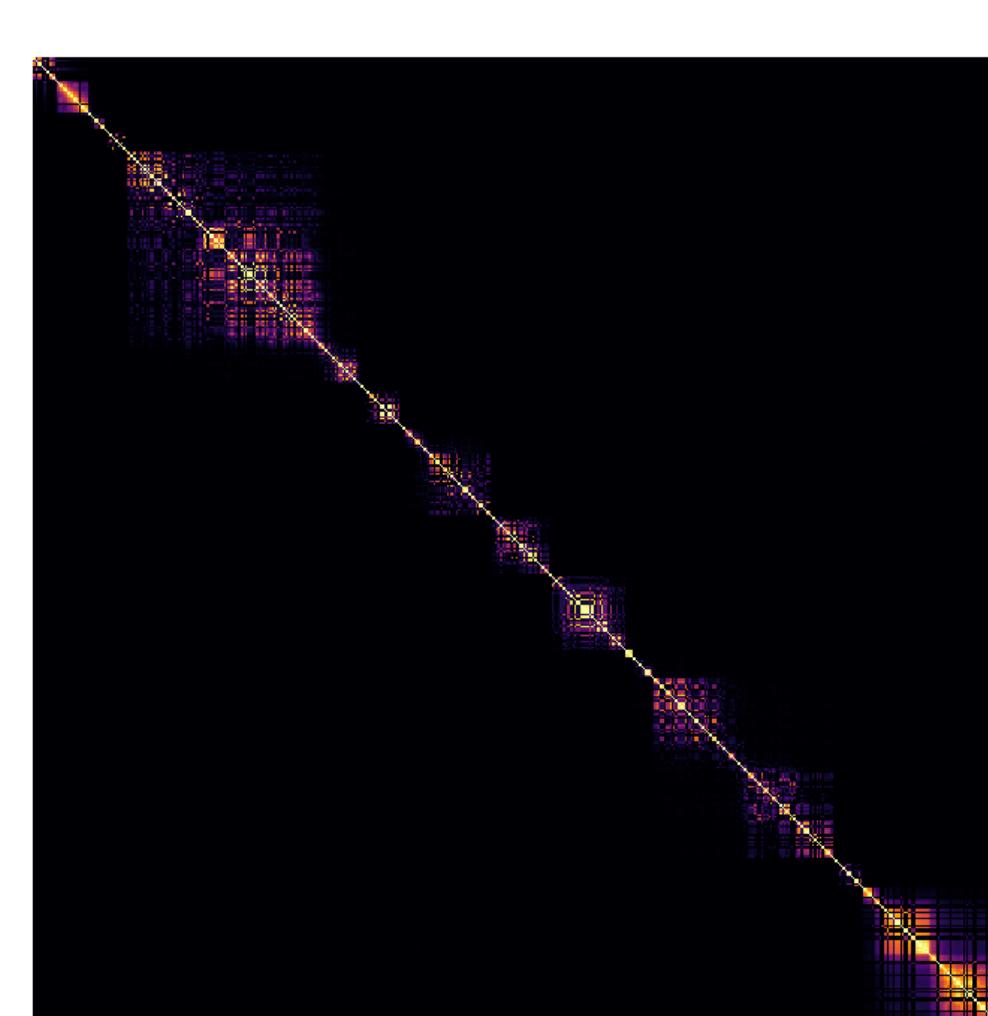
- GWAS
- LD

LD correlation matrices for 500 random contiguous SNPs on Human chromosome 21

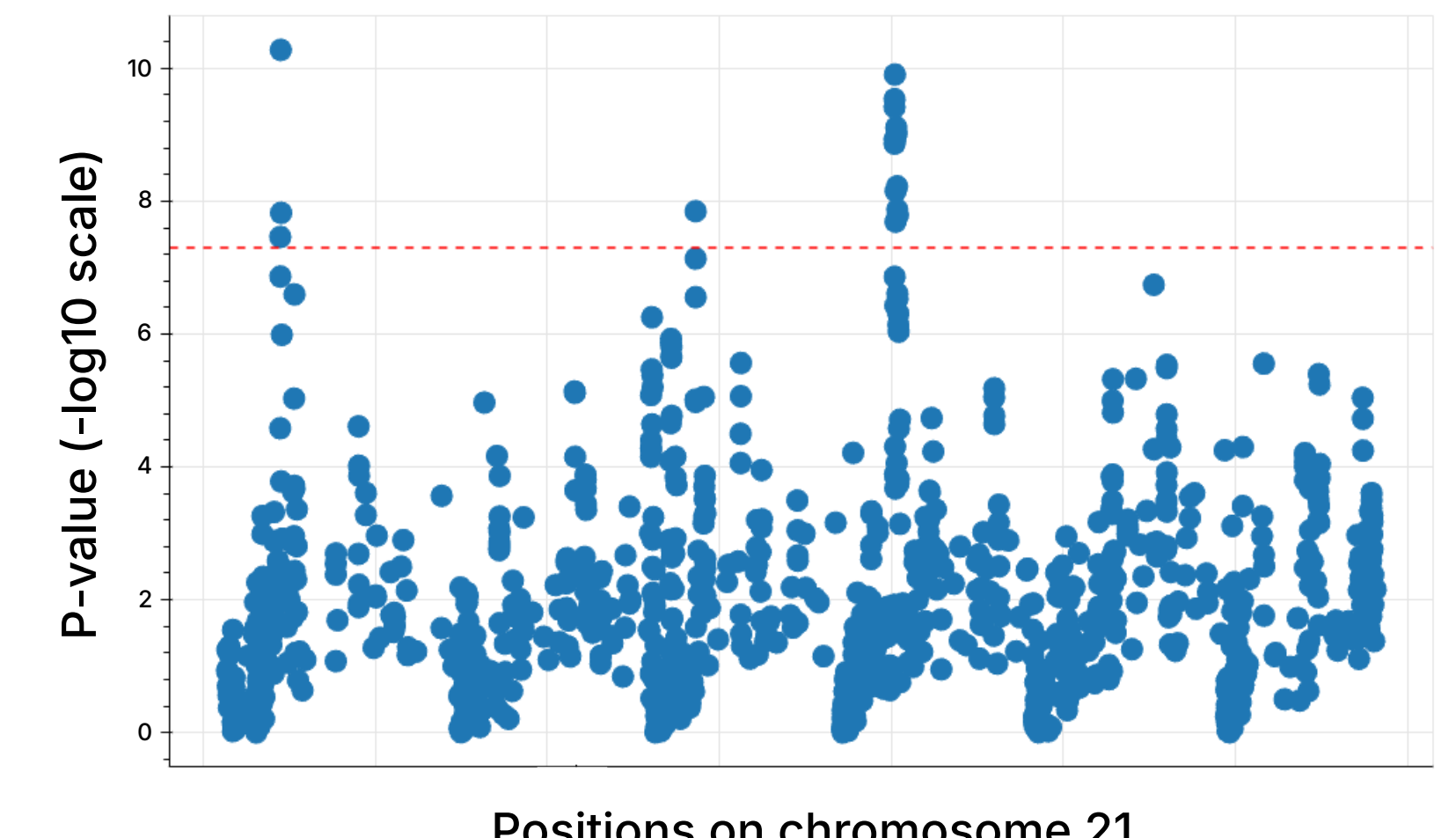
1KG+HGDP reference (N=4,000)



simulated data



GWAS ran on the simulated data



github repository



Legend:
LD: Linkage Disequilibrium
GWAS: Genome-wide association studies
1KG: 1000 Genomes Project
HGDP: Human Genome Diversity Project

References:
1. Maher, B. Personal genomes: The case of the missing heritability, Nature 2008.
2. Wharrie et al. HAPNEST: efficient, large-scale generation and evaluation of synthetic datasets for genotypes and phenotypes, Bioinformatics 2023.
3. Urbanowicz et al. GAMETES: a fast, direct algorithm for generating pure, strict, epistatic models with random architectures, BioData 2012.
4. Blumenthal et al. EpiGEN: an epistasis simulation pipeline, Bioinformatics 2020.
5. Shang et al. EpiReSIM: A Resampling Method of Epistatic Model without Marginal Effects Using Under-Determined System of Equations, Genes 2022.