A Hierarchical Model for Simulating a Quantitative Phenotype Mediated by Gene Expression

Introduction

This document details an advanced simulation framework for a quantitative phenotype. It extends the foundational model by incorporating a biologically plausible hierarchical structure that reflects the causal chain from genetic variants to an intermediate molecular trait (gene expression) and finally to a complex organismal phenotype.

This two-stage model simulates a phenotype whose genetic architecture is mediated by the expression levels of a predefined set of genes. The expression of each gene is, in turn, regulated by its own set of genetic loci, known as expression quantitative trait loci (eQTLs). This approach enables a more nuanced examination of how genetic regulation influences complex trait variation.

The simulation will utilize real genotype data for individuals, while the gene expression levels and the final phenotype will be simulated based on the parameters specified in the model.

The Simulation Model

Stage 1: Simulating Gene Expression Levels

The expression level of gene j, GE_j, is the sum of its genetic value ($G_{GE,j}$) and an environmental/residual component ($E_{GE,j}$).

$$GE_j = G_{GE,j} + E_{GE,j}$$

The genetic value is determined by its k_j eQTLs:

$$G_{GE,j} = \sum_{i=1}^{k_j} X_{ji}\,\alpha_{ji}$$

The environmental component is drawn from a normal distribution, $E_{GE,j} \sim \mathcal{N}(0, \sigma^2_{E_{GE,j}})$, with its variance determined by the gene expression heritability, h^2_{GE}:

$$\sigma^2_{E_{GE,j}} = \mathrm{Var}(G_{GE,j})\left(\frac{1 - h^2_{GE}}{h^2_{GE}}\right)$$

Stage 2: Simulating the Final Quantitative Phenotype

To accurately model the heritability, we decompose the phenotype P into three independent components:

$$P = G_{P,\text{eQTL}} + C_{GE,\text{env}} + E_{P,\text{resid}}$$

- $G_{P,\text{eQTL}}$ **(Pure Genetic Component):** The portion of phenotypic variance strictly mediated by the genetic value of gene expression.

$$G_{P,\text{eQTL}} = \sum_{j=1}^{m} b_j \, G_{GE,j}$$

- $C_{GE,\text{env}}$ **(Downstream Non-Genetic Component):** The variance in the phenotype arising from the non-heritable component of gene expression.

$$C_{GE,\text{env}} = \sum_{j=1}^{m} b_j \, E_{GE,j}$$

- $E_{P,\text{resid}}$ **(Residual Component):** All other sources of variance independent of the m gene expression pathways.

$$E_{P,\text{resid}} \sim \mathcal{N}(0, \sigma^2_{E_{P,\text{resid}}})$$

Deriving Residual Variance and a Key Constraint

The heritability parameter $h^2_{P,\text{eQTL}}$ (your 'h2') is defined as the proportion of total phenotypic variance explained by the pure genetic component:

$$h^2_{P,\text{eQTL}} = \frac{\text{Var}(G_{P,\text{eQTL}})}{\text{Var}(P)}$$

The total non-genetic variance is therefore:

$$\text{Var}(C_{GE,\text{env}}) + \text{Var}(E_{P,\text{resid}}) = \text{Var}(G_{P,\text{eQTL}}) \left(\frac{1 - h^2_{P,\text{eQTL}}}{h^2_{P,\text{eQTL}}} \right)$$

From this, we solve for the variance of the final residual term:

$$\sigma^2_{E_{P,\text{resid}}} = \left[\text{Var}(G_{P,\text{eQTL}}) \left(\frac{1 - h^2_{P,\text{eQTL}}}{h^2_{P,\text{eQTL}}} \right) \right] - \text{Var}(C_{GE,\text{env}})$$

Condition for a Valid Simulation

For $\sigma^2_{E_{P,\text{resid}}}$ to be non-negative, the total non-genetic variance allowed by $h^2_{P,\text{eQTL}}$ must be at least as large as the non-genetic variance already introduced from the gene expression layer. This leads to a simple and critical constraint on the input parameters:

$$\boxed{h^2_{P,\text{eQTL}} \leq h^2_{GE}}$$

In words, the heritability of the final phenotype that is mediated by eQTLs cannot exceed the heritability of the intermediate gene expression traits themselves. This check must be performed prior to simulation to ensure the parameter set is mathematically and biologically plausible.