A Comprehensive Guide to Simulating Phenotypes with Causal and Complex Pleiotropic Architectures

Abstract

This document provides a comprehensive guide to a generalized mathematical framework for simulating two phenotypes, P_1 and P_2, that may be linked by causality, confounding, and pleiotropy. The model's core innovation is a flexible and biologically principled representation of pleiotropy, where a subset of single-nucleotide polymorphisms (SNPs) influencing P_1 also exert a direct effect on P_2. These pleiotropic effects are modeled as random variables at the per-SNP level. The mean of the effect size distribution captures systematic, directional pleiotropy, while its variance captures random, heterogeneous pleiotropy.

This guide serves a dual purpose. For researchers, it offers a clear explanation of the principles behind the data generation process, enabling a deeper understanding of the simulated data's structure. For developers, it provides a precise, step-by-step algorithm for implementation, complete with robust handling of all logical edge cases, such as the absence of pleiotropy ($q = 0$) and the absence of a causal pathway ($\alpha = 0$). By detailing all assumptions, derivations, and calculations, this document aims to be a definitive resource for creating sophisticated and realistic phenotype simulations.

Introduction: The Need for Principled Simulation

In quantitative genetics and epidemiology, simulation is an indispensable tool. A robust simulation framework must be able to generate data that mirrors the complexities of biological reality, including causal relationships, confounding, and the multifaceted nature of pleiotropy. This document describes such a framework, providing a blueprint for generating two phenotypes, P_1 and P_2. Our goal is to provide a guide that is both conceptually clear for researchers and technically precise for software developers.

The Simulation Model

Conceptual Overview

The model defines the relationships between five core entities: two genetic factors (G_1, G_2), a common confounder (C), and two phenotypes (P_1, P_2). The key relationships are:

- **Genetic Basis:** G_1 influences P_1, and G_2 influences P_2.
- **Causality:** P_1 may have a direct causal effect on P_2.
- **Confounding:** A shared factor C may influence both P_1 and P_2.
- **Pleiotropy:** A subset of the genetic variants composing G_1 may also have a direct effect on P_2, independent of the causal path through P_1.

The Structural Equations

These relationships are formalized by the following linear equations:

$$\begin{aligned} P_1 &= G_1 + c_1 C + E_1 \\ P_2 &= \alpha P_1 + G_{\text{pleio}} + \beta G_2 + c_2 C + E_2 \end{aligned}$$

where G_1 is composed of pleiotropic and non-pleiotropic SNPs, and G_{pleio} represents the direct genetic effect on P_2 from the pleiotropic SNPs.

Detailed Component Definitions

- P_1, P_2: The final observable phenotypic values.
- G_1, G_2: The aggregate genetic scores, which are mean-centered. For a set of m causal SNPs, a genetic score is constructed as $G = \sum_{i=1}^{m} (X_i - 2p_i)\beta_i$, where X_i is the genotype (coded 0, 1, 2), p_i is the effect allele frequency, and β_i is the per-allele effect size.
- **Simplification Assumption:** For clarity, we assume each causal SNP for P_1 has a uniform effect size, $\beta_{1,i} = 1$. Therefore, the genetic score is $G_1 = \sum_{i=1}^{m_1} (X_i - 2p_i)$. This centering ensures $E[G_1] = 0$, a critical assumption for variance derivations. The variance is $\text{Var}(G_1) = \sum_{i=1}^{m_1} 2\, p_i(1 - p_i)$.
- G_{pleio}: The pleiotropic genetic score for P_2. It is constructed from the subset of m_{pleio} SNPs that are also in G_1. For each such pleiotropic SNP j, its effect on P_2 is a random variable γ_j. Thus, $G_{\text{pleio}} = \sum_{j=1}^{m_{\text{pleio}}} (X_j - 2p_j)\gamma_j$.

- γ_j: The per-SNP pleiotropic effect on P_2, modeled as a random variable for each pleiotropic SNP j: $\gamma_j \sim \mathcal{N}(\mu_\gamma, \sigma_\gamma^2)$. These effects are drawn once and are then fixed for all individuals.

- C: A common confounder, standardized with $E[C] = 0$ and $\text{Var}(C) = 1$.

- E_1, E_2: Independent, phenotype-specific residual effects, with $E[E_1] = E[E_2] = 0$.

- α, β, c_1, c_2: Fixed coefficients defining the architecture.

- q: The proportion of SNPs in G_1 that are pleiotropic. $m_{\text{pleio}} = q \cdot m_1$.

All base components are assumed to be mutually independent, except for the shared genetic variants underlying G_1 and G_{pleio}.

Derivation of Variance Components

The derivation of variance relies on the independence and mean-zero property of the base components. The variance of P_1 is:

$$\text{Var}(P_1) = \text{Var}(G_1) + c_1^2 + \text{Var}(E_1)$$

To derive the variance of P_2, we first substitute P_1 into the equation for P_2:

$$P_2 = \alpha(G_1 + c_1 C + E_1) + G_{\text{pleio}} + \beta G_2 + c_2 C + E_2$$

Let G_1 be the sum of its non-pleiotropic ($G_{1,np}$) and pleiotropic ($G_{1,p}$) parts. The genetic components influencing P_2 are αG_1, G_{pleio}, and βG_2. The total genetic variance in P_2 is $\text{Var}(\alpha G_{1,np} + (\alpha G_{1,p} + G_{\text{pleio}}) + \beta G_2)$. Through derivation (see Appendix), this simplifies elegantly. The full variance equation for P_2 is:

$$\begin{aligned}\text{Var}(P_2) = &\left[\alpha^2(1-q) + ((\alpha + \mu_\gamma)^2 + \sigma_\gamma^2)q\right]\text{Var}(G_1) + \beta^2\text{Var}(G_2) \\ &+ (\alpha c_1 + c_2)^2 + \alpha^2\text{Var}(E_1) + \text{Var}(E_2)\end{aligned}$$

A Step-by-Step Guide to Parameter Calculation

This section provides a complete algorithm for calculating all unknown model coefficients from a set of intuitive, user-defined parameters.

User-Specified Parameters

- $\text{Var}(G_1)$,$\text{Var}(G_2)$: Base genetic variances, calculated from the number of causal SNPs and their allele frequencies.

- α: The causal effect coefficient. A value of 0 indicates no causal path.

- q: The proportion of causal SNPs for P_1 that are pleiotropic.
- $h_1^2, v_{c,1}^2$: Heritability and confounder variance proportion for P_1.
- $h_2^2, v_{c,2}^2$: Heritability and confounder variance proportion for P_2.
- $h^2_{\text{pleio-dir}}, h^2_{\text{pleio-rand}}$: Variance proportions for directional and random pleiotropy, respectively. These represent the proportion of $\text{Var}(P_2)$ explained by the mean-centered and variance components of the pleiotropic effects. Must be 0 if $q = 0$.
- **Conditional Parameter for Causal Path:**
    - Only required if $\alpha \neq 0$: $v^2_{P_1 \to P_2}$, the proportion of variance in P_2 explained by the causal path from P_1.

The Calculation Algorithm

Step 1: Calculate Parameters for Phenotype P_1

Purpose: To fully define the scale and composition of P_1, which may act as a causal source of variance for P_2.

1. **Total variance of P_1:** $\text{Var}(P_1) = \text{Var}(G_1)/h_1^2$.
2. **Confounder scaling coefficient:** $c_1 = \sqrt{v_{c,1}^2 \cdot \text{Var}(P_1)}$.
3. **Residual variance:** $\text{Var}(E_1) = \text{Var}(P_1)(1 - h_1^2 - v_{c,1}^2)$.

Step 2: Determine the Scale and Parameters of Phenotype P_2

Purpose: To establish the absolute variance of P_2 and solve for its genetic parameters. The logic branches based on the presence of a causal path.

Case 1: Causal Path Exists ($\alpha \neq 0$)

Principle: The causal effect from the now-defined P_1 provides a natural, data-driven anchor to determine the total variance of P_2.

1. **Calculate Total Variance of P_2:**

$$\text{Var}(P_2) = \frac{\alpha^2 \text{Var}(P_1)}{v^2_{P_1 \to P_2}}$$

2. **Calculate Pleiotropic Parameters ($\mu_\gamma, \sigma_\gamma^2$):** If $q > 0$, solve using the known $\text{Var}(P_2)$. The variance from directional pleiotropy is $\mu_\gamma^2 \cdot q \cdot \text{Var}(G_1)$, and from random pleiotropy is $\sigma_\gamma^2 \cdot q \cdot \text{Var}(G_1)$.

$$\mu_\gamma^2 = \frac{h^2_{\text{pleio-dir}} \cdot \text{Var}(P_2)}{q \cdot \text{Var}(G_1)}, \quad \sigma_\gamma^2 = \frac{h^2_{\text{pleio-rand}} \cdot \text{Var}(P_2)}{q \cdot \text{Var}(G_1)}$$

(Note: We solve for μ_γ^2; its sign is arbitrary and can be set to positive by convention.) If $q = 0$, then $\mu_\gamma = 0, \sigma_\gamma^2 = 0$.

3. **Calculate Genetic Scaling Coefficient β:** The total genetic variance in P_2 is $\text{Varg}(P_2) = h_2^2 \cdot \text{Var}(P_2)$. We solve for β^2:

$$\beta^2 = \frac{\text{Varg}(P_2) - \left[\alpha^2(1-q) + ((\alpha + \mu_\gamma)^2 + \sigma_\gamma^2)q\right]\text{Var}(G_1)}{\text{Var}(G_2)}$$

Case 2: No Causal Path ($\alpha = 0$)

Principle: With no causal anchor, the scale of P_2 is indeterminate from proportions alone. We establish the scale by creating a direct, standardized link to its own unique genetic basis. By convention, we set $\beta = 1$.

1. **Set the Genetic Scaling Coefficient β:** $\beta = 1$.

2. **Derive Total Variance of P_2:** The absolute variance from G_2 is $\text{Var}(\beta G_2) = \text{Var}(G_2)$. The proportional variance from G_2 is the total heritability minus the pleiotropic contributions: $h^2_{G_2 \to P_2} = h_2^2 - (h^2_{\text{pleio-dir}} + h^2_{\text{pleio-rand}})$. Equating these gives:

$$h^2_{G_2 \to P_2} = \frac{\text{Var}(\beta G_2)}{\text{Var}(P_2)} \Longrightarrow \text{Var}(P_2) = \frac{\text{Var}(G_2)}{h_2^2 - h^2_{\text{pleio-dir}} - h^2_{\text{pleio-rand}}}$$

This result elegantly determines the total variance of P_2 from the input proportions and the variance of its own genetic component.

3. **Calculate Pleiotropic Parameters ($\mu_\gamma, \sigma_\gamma^2$):** Now that $\text{Var}(P_2)$ is known, we calculate the pleiotropic parameters as in the causal case (with $\alpha = 0$).

Step 3: Calculate Remaining Parameters for P_2 (Common to both cases)

Purpose: To solve for the final non-genetic coefficients now that the scale of P_2 is fully determined.

1. **Confounder Scaling Coefficient c_2:** The confounder variance in P_2 is $(\alpha c_1 + c_2)^2 = v_{c,2}^2 \cdot \text{Var}(P_2)$.

$$c_2 = \sqrt{v_{c,2}^2 \cdot \text{Var}(P_2)} - \alpha c_1$$

(Note: If $\alpha = 0$, this simplifies to $c_2 = \sqrt{v_{c,2}^2 \cdot \text{Var}(P_2)}$).

2. **Residual Variance** Var(E_2)**:** This term accounts for all remaining variance.

$$\text{Var}(E_2) = \text{Var}(P_2)(1 - h_2^2 - v_{c,2}^2) - \alpha^2\text{Var}(E_1)$$

   (Note: If $\alpha = 0$, this simplifies to $\text{Var}(E_2) = \text{Var}(P_2)(1 - h_2^2 - v_{c,2}^2)$).

Simulation and Interpretation

Once all model parameters have been calculated, phenotypes can be simulated for a population of individuals. The crucial step for pleiotropy is:

1. **Before the simulation loop:** For each of the m_{pleio} pleiotropic SNPs, draw a single effect size γ_j from the distribution $\mathcal{N}(\mu_\gamma, \sigma_\gamma^2)$. These values are now fixed constants for the duration of the simulation.

2. **For each individual:** Calculate P_1 and P_2 using the structural equations. The pleiotropic term G_{pleio} is calculated as the sum of the individual's genotypes at the pleiotropic loci, weighted by their corresponding fixed γ_j values.

This generalized framework elegantly captures simpler models as special cases:

- **Fixed Correlated Pleiotropy:** Achieved by setting $h_{\text{pleio-rand}}^2 = 0$. This forces $\sigma_\gamma^2 = 0$, making each γ_j a constant equal to μ_γ.

- **Random Uncorrelated Pleiotropy:** Achieved by setting $h_{\text{pleio-dir}}^2 = 0$. This forces $\mu_\gamma = 0$, and each γ_j is drawn from a zero-mean distribution.

- **Mixed-Effect Pleiotropy:** Setting both $h_{\text{pleio-dir}}^2 > 0$ and $h_{\text{pleio-rand}}^2 > 0$ models a scenario with both a systematic directional effect and random heterogeneity around it.

Conclusion: A Blueprint for Principled Simulation

This document has laid out a comprehensive and robust framework for phenotypic simulation. By providing a clear distinction between user-defined intuitive parameters and the derived model coefficients, it enables the creation of complex datasets where the underlying ground truth is precisely known. The biologically principled, per-SNP model of pleiotropy and the careful handling of different causal scenarios make it a powerful tool for methodological development and for teaching the core principles of quantitative genetics. Adherence to the steps outlined here will ensure that simulated data is not only complex and realistic but also internally consistent and reproducible.