

# Heritability

Nuno Carvalho

July 14, 2025

## 1 SNP Heritability

These derivations are based on the Methods of [Yang et al., 2010].

### 1.1 Phenotype model

We can define a quantitative phenotype  $y$  as:

$$\mathbf{y} = \mathbf{X}_c \boldsymbol{\beta} + \boldsymbol{\epsilon}$$

Where:

- $\mathbf{y}$ : phenotypes.  $N \times 1$  vector. Centered so that  $E[\mathbf{y}] = 0$ .
  - $N$ : number of samples.
- $\mathbf{X}_c$ : normalized genotypes for causal variants.  $N \times M_c$  matrix.
  - $M_c$ : number of causal variants.
  - Normalized according to  $\mathbf{X}_{c,i} = \frac{\mathbf{X}'_{c,i} - 2f_i}{\sqrt{2f_i(1-f_i)}}$ .
    - $\mathbf{X}'_c$ : allele dosages, taking on values of 0, 1, 2.
    - $f_i$ : true population allele frequency for variant  $i$ .
    - Such that for each row (variant),  $E[\mathbf{X}_{c,i}] = 0$  and  $\text{Var}[\mathbf{X}_{c,i}] = 1$ .
- $\boldsymbol{\beta}$ : per-normalized-genotype causal effects.  $M_c \times 1$  vector.
  - Assume infinitesimal model.
  - Drawn from  $\boldsymbol{\beta} \sim \mathcal{N}(\mathbf{0}, \mathbf{I}\sigma_\beta^2)$ .
    - $\mathbf{I}$ :  $M_c \times M_c$  identity matrix.
    - $\sigma_\beta^2$ : variance of causal effects.
- $\boldsymbol{\epsilon}$ : residual effects (i.e. error or noise term).  $N \times 1$  vector.
  - Drawn from  $\boldsymbol{\epsilon} \sim \mathcal{N}(\mathbf{0}, \mathbf{I}\sigma_\epsilon^2)$ .
    - $\mathbf{I}$ :  $N \times N$  identity matrix.
    - $\sigma_\epsilon^2$ : residual variance.

We assume  $\mathbf{X}_c$ ,  $\boldsymbol{\beta}$ , and  $\boldsymbol{\epsilon}$  are all independent from each other. We can define the genetic effects as a single term,  $\mathbf{g} = \mathbf{X}_c \boldsymbol{\beta}$ , meaning that:

$$\mathbf{y} = \mathbf{g} + \boldsymbol{\epsilon} \quad \text{where} \quad \mathbf{g} \sim \mathcal{N}(\mathbf{0}, \mathbf{I}\sigma_g^2) \quad \text{where} \quad \sigma_g^2 = M_c \sigma_\beta^2$$

We interpret  $\sigma_g^2$  as variance of total additive genetic effects on the phenotype.

## 1.2 Variance of the phenotype

By making use of the independence between terms, we can define the variance-covariance matrix of  $\mathbf{y}$  as:

$$\begin{aligned}
 \text{Var}[\mathbf{y}] &= \text{Var}[\mathbf{X}_c\boldsymbol{\beta} + \boldsymbol{\epsilon}] \\
 &= \text{Var}[\mathbf{X}_c]\text{Var}[\boldsymbol{\beta}] + \text{Var}[\boldsymbol{\epsilon}] \\
 &= (\mathbf{X}_c\mathbf{X}_c^\top)\sigma_g^2 + \mathbf{I}\sigma_\epsilon^2 \\
 &= (\mathbf{X}_c\mathbf{X}_c^\top)\frac{\sigma_g^2}{M_c} + \mathbf{I}\sigma_\epsilon^2 \\
 &= \mathbf{G}\sigma_g^2 + \mathbf{I}\sigma_\epsilon^2
 \end{aligned}$$

Where we define  $\mathbf{G} = \frac{\mathbf{X}_c\mathbf{X}_c^\top}{M_c}$  as the  $N \times N$  genetic relationship matrix (GRM) between individuals. The  $G_{ii}$  element is the variance of individual  $i$ 's normalized genotype vector, while the  $G_{ij}$  element is the covariance of individuals  $i$  and  $j$ 's normalized genotype vectors.

Narrow-sense heritability is defined as the proportion of phenotypic variance,  $\sigma_P^2$ , explained by additive genetic effects:

$$h^2 = \frac{\sigma_g^2}{\sigma_P^2} = \frac{\sigma_g^2}{\sigma_g^2 + \sigma_\epsilon^2}$$

## 1.3 Estimating the GRM

In practice, we likely do not know the exact set of causal variants and instead must estimate the GRM using a set of genotyped SNPs:

$$\mathbf{A} = \frac{\mathbf{X}\mathbf{X}^\top}{M}$$

Where  $\mathbf{A}$  is the estimated GRM,  $\mathbf{X}$  is the normalized genotype matrix of our genotyped SNPs, and  $M$  is the number of genotyped SNPs. Note that because we are also working with a sample,  $\mathbf{X}$  is normalized using sample allele frequencies,  $\mathbf{p}$ :

$$\mathbf{X}_i = \frac{\mathbf{X}'_i - 2p_i}{\sqrt{2p_i(1 - p_i)}}$$

However, this equation for  $\mathbf{A}$  ignores the sampling error associated with each SNP. Let's consider the covariance computation between two individuals for SNP  $i$ , which is then

summed across  $M$  SNPs to get the value for  $A_{jk}$ . When  $j \neq k$ :

$$\begin{aligned} A_{ijk} &= x_{ij}x_{ik} \\ &= \frac{x'_{ij} - 2p_i}{\sqrt{2p_i(1-p_i)}} \frac{x'_{ik} - 2p_i}{\sqrt{2p_i(1-p_i)}} \\ &= \frac{(x'_{ij} - 2p_i)(x'_{ik} - 2p_i)}{2p_i(1-p_i)} \end{aligned}$$

Because  $x'_{ij}$  and  $x'_{ik}$  are independent from each other, the expected value of this is:

$$\begin{aligned} E[A_{ijk}] &= E\left[\frac{(x'_{ij} - 2p_i)(x'_{ik} - 2p_i)}{2p_i(1-p_i)}\right] \\ &= \frac{(E[x'_{ij}] - 2p_i)(E[x'_{ik}] - 2p_i)}{2p_i(1-p_i)} \\ &= \frac{(2p_i - 2p_i)(2p_i - 2p_i)}{2p_i(1-p_i)} \\ &= 0 \end{aligned}$$

This makes sense if our sample is of unrelated individuals. If the raw genotypes of two individuals at a SNP are independent from each other, then the covariance of their adjusted genotypes should also be zero. Furthermore, the variance of  $A_{ijk}$  is given by:

$$\begin{aligned} \text{Var}[A_{ijk}] &= \text{Var}\left[\frac{(x'_{ij} - 2p_i)(x'_{ik} - 2p_i)}{2p_i(1-p_i)}\right] \\ &= \frac{\text{Var}[(x'_{ij} - 2p_i)]\text{Var}[(x'_{ik} - 2p_i)]}{(2p_i(1-p_i))^2} \\ &= \frac{\text{Var}[(x'_{ij})]\text{Var}[(x'_{ik})]}{(2p_i(1-p_i))^2} \\ &= \frac{(2p_i(2-p_i))(2p_i(2-p_i))}{(2p_i(1-p_i))^2} \\ &= 1 \end{aligned}$$

So, the variance in  $A_{jk}$  is independent of allele frequency, which is a desirable property. But do these properties hold up when  $j = k$  (i.e. variance of an individual's genotype)?

$$\begin{aligned} A_{ijj} &= x_{ij}^2 \\ &= \left(\frac{x'_{ij} - 2p_i}{\sqrt{2p_i(1-p_i)}}\right)^2 \\ &= \frac{(x'_{ij} - 2p_i)^2}{2p_i(1-p_i)} \end{aligned}$$

For deriving  $E[A_{ijj}]$ , we make use of:

$$\begin{aligned}
E[x'_{ij}] &= \sum_{n=0}^2 E[x'_{ij}|x'_{ij} = n] \Pr(x'_{ij} = n) \\
&= (0)^2(1 - p_i)^2 + (1)^2 2p_i(1 - p_i) + (2)^2(p_i)^2 \\
&= 2p_i - 2p_i^2 + 4p_i^2 \\
&= 2p_i^2 + 2p_i
\end{aligned}$$

Allowing us to calculate  $E[A_{ijj}]$ :

$$\begin{aligned}
E[A_{ijj}] &= E\left[\frac{(x'_{ij} - 2p_i)^2}{2p_i(1 - p_i)}\right] \\
&= \frac{E[x'_{ij}^2 - 4p_i x'_{ij} + 4p_i^2]}{2p_i(1 - p_i)} \\
&= \frac{E[x'_{ij}^2] - 4p_i E[x'_{ij}] + 4p_i^2}{2p_i(1 - p_i)} \\
&= \frac{2p_i^2 + 2p_i - 4p_i(2p_i) + 4p_i^2}{2p_i(1 - p_i)} \\
&= \frac{2p_i - 2p_i^2}{2p_i(1 - p_i)} \\
&= \frac{2p_i(1 - p_i)}{2p_i(1 - p_i)} \\
&= 1
\end{aligned}$$

So, the expected variance of an individual's normalized genotype is 1 and independent of the allele frequency, which is a desirable property. However, this frequency-independence does not hold for the variance. For simplicity, let's denote  $P = 2p_i(1 - p_i)$  and make use of  $\text{Var}[Y] = E[Y^2] - E[Y]^2$ :

$$\begin{aligned}
\text{Var}[A_{ijj}] &= \text{Var}\left[\frac{(x'_{ij} - 2p_i)^2}{P}\right] \\
&= \frac{\text{Var}[(x'_{ij} - 2p_i)^2]}{(P)^2} \\
&= \frac{E[((x'_{ij} - 2p_i)^2)^2] - E[(x'_{ij} - 2p_i)^2]^2}{(P)^2} \\
&= \frac{(P) - (P)^2}{(P)^2} \\
&= \frac{(P)(1 - P)}{(P)^2} \\
&= \frac{1 - P}{P} \\
&= \frac{1 - 2p_i(1 - p_i)}{2p_i(1 - p_i)}
\end{aligned}$$

The full derivation for why  $E[(x'_{ij} - 2p_i)^2] = E[(x'_{ij} - 2p_i)^2] = 2p_i(1 - p_i)$  is very lengthy algebraically, but can be shortcutted by using the formula for the [higher moments of a binomially distributed variable](#), where  $n = 2$  and  $p = p_i$ . Importantly, the variance of  $A_{jj}$  therefore depends on the allele frequencies of the SNPs, even after normalization.  $\mathbf{A}$  would be a better estimator of the GRM if for the same individual, its variance did not depend on the allele frequency, so we want to adjust  $A_{ijj}$  so that  $E[A_{ijj}] = 1$  like before, but also that  $\text{Var}[A_{ijj}] = 1$  like for unrelated individuals.

To be frank, I am not sure how the authors derived it, but this equation holds both properties:

$$A_{ijj} = 1 + \frac{x'^2_{ij} - (1 + 2p_i)x_{ij} + 2p_i^2}{2p_i(1 - p_i)}$$

We now explicitly define  $\mathbf{A}$  as follows:

$$A_{jk} = \frac{1}{N} \sum_{i=1}^M A_{ijk} = \begin{cases} \frac{1}{N} \sum_{i=1}^M \frac{(x'_{ij} - 2p_i)(x'_{ik} - 2p_i)}{2p_i(1 - p_i)}, & j \neq k \\ 1 + \frac{1}{N} \sum_{i=1}^M \frac{x'^2_{ij} - (1 + 2p_i)x'_{ij} + 2p_i^2}{2p_i(1 - p_i)}, & j = k \end{cases}$$