Building the Genomic Relationship Matrix

Fei Ge

September 15, 2025

This is the foundation. It quantifies the genetic similarity between individuals based on genome-wide markers.

Step 1: Create and Center the Genotype Matrix

Input: A raw genotype matrix W $(n \times m)$ of n individuals and m markers, coded 0, 1, and 2.

1. **Calculate allele frequencies (p_i):** For each marker i, calculate the allele frequency
$$p_i = \frac{\text{total count of allele at marker } i}{2 \times (\text{number of individuals})}.$$
2. **Create the Frequency Matrix (P):** Construct a matrix P where every element in column i is the value $2p_i$. This represents the expected genotype score based on allele frequency.
3. **Center the Genotype Matrix (M):** Subtract P from the raw genotype matrix W:
$$M = W - P.$$
This adjustment centers the genotypes by allele frequencies, giving each marker a mean of zero.

Example: Suppose we have 4 individuals and 3 markers. Then the raw genotype matrix W is
$$W = \begin{bmatrix} 1 & 2 & 0 \\ 1 & 1 & 1 \\ 2 & 2 & 0 \\ 0 & 1 & 2 \end{bmatrix}.$$

1. **Calculate the allele frequencies (p_i):** The total number of alleles is $2 \times 4 = 8$. For each marker (column):
$$p_1 = \tfrac{4}{8} = 0.5, \quad p_2 = \tfrac{6}{8} = 0.75, \quad p_3 = \tfrac{3}{8} = 0.375.$$

2. **Create the frequency matrix (P):** Each column i contains the value $2p_i$ (Since each individual has two chromosomes, their expected genotype score is not p_i, but $2p_i$):

$$2p_1 = 1.0, \quad 2p_2 = 1.5, \quad 2p_3 = 0.75.$$

So

$$P = \begin{bmatrix} 1.0 & 1.5 & 0.75 \\ 1.0 & 1.5 & 0.75 \\ 1.0 & 1.5 & 0.75 \\ 1.0 & 1.5 & 0.75 \end{bmatrix}.$$

3. **Center the matrix ($M = W - P$):** Subtract P from W:

$$M = \begin{bmatrix} 1 & 2 & 0 \\ 1 & 1 & 1 \\ 2 & 2 & 0 \\ 0 & 1 & 2 \end{bmatrix} - \begin{bmatrix} 1.0 & 1.5 & 0.75 \\ 1.0 & 1.5 & 0.75 \\ 1.0 & 1.5 & 0.75 \\ 1.0 & 1.5 & 0.75 \end{bmatrix} = \begin{bmatrix} 0 & 0.5 & -0.75 \\ 0 & -0.5 & 0.25 \\ 1.0 & 0.5 & -0.75 \\ -1.0 & -0.5 & 1.25 \end{bmatrix}.$$

Step 2: Scale the Centered Matrix and Calculate G

1. **Calculate the Scaling Factor:** The denominator used for scaling is the sum of the variances of all the markers. This is calculated as:

$$k = 2 \sum_{i=1}^{m} p_i(1 - p_i)$$

where m is the total number of markers. This factor ensures that the average diagonal element of the final G matrix is close to 1, making it analogous to a traditional pedigree-based relationship matrix.

Explanation: The scaling factor k is the sum of the expected variances of all markers. The term $2p_i(1 - p_i)$ comes from the statistical variance of a single genetic marker, assuming Hardy-Weinberg equilibrium.

- **Single Allele as a Bernoulli Trial:** Consider drawing one allele from the population. It is either the allele of interest (value 1) or the other allele (value 0). The variance is

$$\text{Var}(\text{allele}) = p_i(1 - p_i)$$

- **Genotype as Sum of Two Alleles:** An individual's genotype W is the sum of two alleles (one from each parent). Assuming independence (Hardy-Weinberg Equilibrium):

$$W = \text{Allele}_1 + \text{Allele}_2$$

- **Variance of the Genotype:** Using the property that the variance of the sum of independent variables is the sum of their variances:

$$\text{Var}(W) = 2p_i(1 - p_i)$$

- **Scaling Factor k:** Summing across all markers gives

$$k = \sum_{i=1}^{m} \text{Var}(W_i) = 2 \sum_{i=1}^{m} p_i(1 - p_i)$$

2. **Calculate the Genomic Relationship Matrix G:** Once you have M and k, compute G as:

$$G = \frac{MM'}{k}$$

- M: the centered genotype matrix from Step 1.
- M': the transpose of M.
- MM': matrix multiplication that computes pairwise genomic relationships.
- k: the scaling factor calculated above.

Explanation: Dot Product = Similarity

The reason we multiply M by its transpose M' comes from linear algebra and statistics:

Dot Product: A Measure of Similarity

Each row of M is a vector representing an individual's deviation from population allele frequencies across all markers. - A large positive dot product between two rows indicates similar deviations (genetic similarity). - A dot product near zero indicates no correlation. - A large negative dot product indicates opposite deviations (genetic dissimilarity).

Link to Covariance

The covariance between two vectors X and Y is

$$\text{Cov}(X, Y) = \sum_{i} (X_i - \bar{X})(Y_i - \bar{Y})$$

Since M is mean-centered, the dot product of rows i and j is exactly the covariance between individuals i and j.

Gram Matrix Interpretation

- MM' produces an $n \times n$ matrix of dot products between individuals (rows). - The diagonal elements $(MM')_{ii}$ measure an individual's total variance. - The off-diagonal elements $(MM')_{ij}$ measure the genetic covariance between individuals.

In other words, MM'/k is a scaled **Gram matrix**, giving the genomic relationship matrix used in GBLUP.

Example: Continue from Step 1

Step 2: Calculate the G Matrix

a) **Calculate the Numerator (MM'):** Multiply the centered genotype matrix M by its transpose:

$$M = \begin{bmatrix} 0 & 0.5 & -0.75 \\ 0 & -0.5 & 0.25 \\ 1 & 0.5 & -0.75 \\ -1 & -0.5 & 1.25 \end{bmatrix}, \quad M' = \begin{bmatrix} 0 & 0 & 1 & -1 \\ 0.5 & -0.5 & 0.5 & -0.5 \\ -0.75 & 0.25 & -0.75 & 1.25 \end{bmatrix}$$

$$MM' = \begin{bmatrix} 0.8125 & -0.4375 & 0.8125 & -1.1875 \\ -0.4375 & 0.3125 & -0.4375 & 0.5625 \\ 0.8125 & -0.4375 & 1.8125 & -2.1875 \\ -1.1875 & 0.5625 & -2.1875 & 2.8125 \end{bmatrix}$$

b) **Calculate the Scaling Factor k:** Using the formula

$$k = 2 \sum_{i=1}^{m} p_i(1 - p_i),$$

we compute the variance component for each marker:

$$\begin{aligned} &\text{Marker 1: } 2 \times 0.5 \times (1 - 0.5) = 0.5 \\ &\text{Marker 2: } 2 \times 0.75 \times (1 - 0.75) = 0.375 \\ &\text{Marker 3: } 2 \times 0.375 \times (1 - 0.375) = 0.46875 \end{aligned}$$

Summing these gives the total scaling factor:

$$k = 0.5 + 0.375 + 0.46875 = 1.34375$$

c) **Calculate the Final G Matrix:** Divide each element of MM' by the scaling factor $k = 1.34375$:

$$G = \frac{MM'}{k} \approx \begin{bmatrix} 0.605 & -0.326 & 0.605 & -0.884 \\ -0.326 & 0.233 & -0.326 & 0.419 \\ 0.605 & -0.326 & 1.349 & -1.628 \\ -0.884 & 0.419 & -1.628 & 2.093 \end{bmatrix}$$

Interpreting the G Matrix

- **Diagonal Elements:** Represent an individual's relationship with itself. Values greater than 1 (e.g., Ind3 and Ind4) indicate more homozygosity than average.
- **Off-Diagonal Elements:** Represent the estimated genomic relationship between two individuals:
  - Ind1 and Ind3: 0.605 → genetically similar (they share similar alleles).
  - Ind1 and Ind4: -0.884 → genetically dissimilar (opposite deviations from average).