Estimation of Genetic Predisposition

Palle Duun Rohde, Izel Fourie Sørensen & Peter Sørensen

2022-05-13

Contents

[there needs to be consistency regarding “, a, vs. (a), vs just a and also consider whether defining an abbreviation the first time e.g. ”genomic relationship matrix G” should be followed by only using G further on in the text]

1 Introduction

Estimation of genetic predisposition can play an important role in precision health. It can help inform diagnostic procedures and subsequent treatment decisions. The true genetic predisposition for an individual cannot be observed. It is only possible to measure its phenotypic value, which is influenced by variation in genotypes and the environment. Therefore, we need a way to infer the genetic predisposition from the phenotypic value. This section introduces the basic concepts in estimating genetic predisposition, including:

- The basic principle behind estimating genetic predisposition.
- Quantifying the accuracy of the estimated genetic predisposition.
- Use of genetic relationships for estimating genetic predisposition.
- The link between genetic parameters and the estimated genetic predisposition.
- Different methods, data sources and experimental designs for estimating genetic predisposition.

2 Basic principles for genetic predisposition estimation

Genetic predisposition is estimated using information on phenotypes and genetic relationships for individuals in a study population. As introduced previously, the phenotype for a quantitative trait (can be with continuous, categorical or dichotomous variation) is the sum of both genetic and environmental factors. The amount of information provided by the phenotype about the genetic predisposition is determined by the heritability, which measures the proportion of genetic variance contained in the total phenotypic variance. Furthermore, phenotypes collected from close relatives provide more information about the genetic predisposition of an individual. In this section, we will illustrate these principles using phenotypic data and genetic relationships used for estimating genetic predisposition. We will derive a general approach for predicting genetic predisposition and illustrate it with different examples.

2.1 Genetic model

The genetic predisposition is based on an assumption of a specific genetic model. The total genetic effect for an individual is the sum of both additive and non-additive effects:

$$y = \mu + a + d + i + e, \tag{1}$$

where μ is the population mean, a is the genetic predisposition (*i.e.*, additive effect), d is the dominance effect, i is the epistatic effect, and e is the environmental deviation (or residual) not explained by the genetic effects in the model. However, only the additive genetic effects are passed on to the offspring and therefore contributes to the genetic predisposition. In contrast, non-additive genetic effects (*i.e.*, dominance and epistasis) are degraded by recombination and are not inherited, but established within each new offspring where they might be important for the individual's phenotype. Furthermore, additive genetic effects explain the majority of the total genetic variance, thus, we only consider the additive genetic model as the basis for genetic predisposition estimation:

$$y = \mu + a + e.$$

[numbering of equations are not consistent] The true genetic predisposition for an individual, is the sum of all additive genetic effects that affect the quantitative trait:

$$a = \sum_{j=1}^{q} a_j,$$

where a is the total additive genetic effect, and a_j is the additive genetic effect for locus j. We assume (based on the central limit theory) that the true genetic predisposition, a, and the residual term, e, are normally distributed which means that the observed phenotype (y) is also normally distributed:

$$\begin{aligned} a &\sim N(0, \sigma_a^2), \\ e &\sim N(0, \sigma_e^2), \\ y &\sim N(\mu, \sigma_a^2 + \sigma_e^2). \end{aligned}$$

2.2 Expected genetic predisposition conditional on observed phenotype

The genetic predisposition cannot be observed but must be estimated from phenotypic data and genetic relationships between individuals from the study population. Estimation of an unobserved parameter using statistical modelling expresses the estimated quantity as a mathematical function of the observed data. [The question is what this function should look like and what properties the estimated genetic predisposition should fulfill. - I AM NOT SURE THE ANSWER OF “which properties the estimated genetic predisposition should fulfill” IS ANSWERED IN THE FOLLOWING TEXT. NOT SURE THIS SENTENCE SHOULD BE INCLUDED] Under the assumption of multivariate normality for a and y (which is justified under the central limit theorem and the assumptions of many genetic and environmental factors), the expected value of the genetic predisposition, $[E(a)]$, conditional on the observed phenotype (y) can be written as:

$$E(a|y) = E(a) + Cov(a, y)[Var(y)]^{-1}(y - E(y)) \tag{2}$$

The genetic predisposition is defined as deviation from the general mean which means that the expected value $E(a)$ of the true genetic predisposition a is $E(a) = 0$. Therefore, the expected value of the genetic predisposition is :[“is” or “becomes”?]

$$E(a|y) = Cov(a, y)[Var(y)]^{-1}(y - E(y)) \tag{3}$$

The expression for the estimate of the genetic predisposition consists of two parts. The term $y - E(y)$ shows that the observed phenotypic values are corrected for the fixed effects represented by μ [where is the description of μ? - where does it fit in the above equation? Earlier on μ was called the population mean - is that the same as fixed effects?]. The term $b_{a|y} = Cov(a, y)[Var(y)]^{-1}$ often referred to as the regression coefficient is a weighting factor with which the corrected phenotypic values are multiplied.

To be able to estimate the genetic predisposition we need to determine the values for $E(a)$, $E(y)$, $Var(y)$, and $Cov(a, y)$ from the expression above. It is possible to derive simple formula’s for these terms based on:

- adjusted phenotypic observations for the quantitative trait of related individuals,
- heritability of the quantitative trait,
- knowledge of inheritance laws and genetic relationships (*e.g.* parents, grandparents, siblings) for individuals with phenotypic observations of the quantitative trait.

We will distinguish between true and estimated genetic predisposition using the following notation:

$$\begin{aligned} a &= \text{additive genetic value} = \text{true genetic predisposition.} \\ \hat{a} &= E(a|y) = \text{estimated additive genetic value} = \text{estimated genetic predisposition.} \end{aligned}$$

2.2.1 Estimated genetic predispositions are unbiased

Below we show that $\hat{a}$ is an unbiased estimator of a. The expected value of the predicted genetic predisposition ($E(\hat{a})$) can be computed as:

$$\begin{aligned} E(\hat{a}) &= E(Cov(a, y)[Var(y)]^{-1}(y - E(y))) \\ &= Cov(a, y)[Var(y)]^{-1}E((y - E(y))) \\ &= Cov(a, y)[Var(y)]^{-1}(E(y) - E(y)) = 0 \end{aligned}$$

Because we have already specified that $E(\hat{a}) = 0$, it follows that $E(\hat{a}) = E(a) = 0$. This means that $\hat{a}$ is an unbiased estimator of a.

2.2.2 Variance of estimated genetic predisposition ($\hat{a}$)

In a similar fashion, we can derive expressions for the variance of the genetic predisposition ($Var(\hat{a})$) and the covariance between the true and the estimated genetic predisposition ($Cov(a, \hat{a})$):

$$\begin{aligned}
Var(\hat{a}) &= Var(Cov(a,y)[Var(y)]^{-1}(y - E(y))) \\
&= Cov(a,y)[Var(y)]^{-1}Var((y - E(y)))[Var(y)]^{-1}Cov(a,y) \\
&= Cov(a,y)[Var(y)]^{-1}Var(y)[Var(y)]^{-1}Cov(a,y) \\
&= Cov(a,y)[Var(y)]^{-1}Cov(a,y)
\end{aligned}$$

$$\begin{aligned}
Cov(a,\hat{a}) &= Cov(a, Cov(a,y)[Var(y)]^{-1}(y - E(y))) \\
&= Cov(a,y)[Var(y)]^{-1}Cov(a,(y - E(y))) \\
&= Cov(a,y)[Var(y)]^{-1}Cov(a,y) \\
&= Cov(a,y)[Var(y)]^{-1}Cov(a,y) \\
&= Var(\hat{a})
\end{aligned}$$

2.2.3 Conditional density of estimated genetic predisposition ($\hat{a}$)

In some cases, *e.g.*, for specifying confidence intervals of true genetic predisposition, it might be interesting to have a look at the conditional density $f(a|\hat{a})$. This density is a multivariate normal density with expected value $E(a|\hat{a})$ and variance $Var(a|\hat{a})$. These values can be computed based on the theory of conditional multivariate normal densities.

$$\begin{aligned}
E(a|\hat{a}) &= E(a) + Cov(a,\hat{a})[Var(\hat{a})]^{-1}(\hat{a} - E(\hat{a})) \\
&= 0 + Var(\hat{a})[Var(\hat{a})]^{-1}(\hat{a} - 0) \\
&= \hat{a} \\
Var(a|\hat{a}) &= Var(a) - Cov(a,\hat{a})[Var(\hat{a})]^{-1}Cov(a,\hat{a}) \\
Var(a|\hat{a}) &= Var(a)(1 - Cov(a,\hat{a})^2 Var(a)^{-1} Var(\hat{a})^{-1}) \\
Var(a|\hat{a}) &= Var(a)(1 - r^2_{a,\hat{a}})
\end{aligned}$$

[in the beginning of this section “we need to determine the values for $E(a)$, $E(y)$, $Var(y)$, and $Cov(a, y)$” it looks like you will show how to determine these values - but I can only see the calculation of E(a), the variance of a and the covariance of a and hat{a}]

2.3 Accuracy of genetic predisposition estimates

Estimates of genetic predisposition ($\hat{a}$) are estimates of the true genetic predisposition (a), which cannot be observed directly. It is therefore important to quantify how well we have estimated the genetic predisposition in relation to the true genetic predisposition. This can be done using accuracy or reliability.

Accuracy is the correlation between the estimated and the true genetic predisposition:

$$r_{a,\hat{a}} = \frac{\mathrm{Cov}(a,\hat{a})}{\sqrt{\mathrm{Var}(a)\,\mathrm{Var}(\hat{a})}} \tag{4}$$

Reliability is the squared correlation, $r^2_{a,\hat{a}}$, between the estimated genetic predisposition and the true genetic predisposition.

A high correlation means that the estimated genetic predisposition is very accurate.

To be able to compute the accuracy or reliability of the estimated genetic predisposition we need to determine the values for the terms, $Cov(a, \hat{a})$, $Var(\hat{a})$, and $Var(a)$ in the expression above. It can be shown that the variance of the estimated genetic predisposition is the same as the covariance between the true and estimated genetic predisposition (i.e. $Cov(a, \hat{a}) = Var(\hat{a})$). Therefore, the reliability can be expressed as:

$$r^2_{a,\hat{a}} = \frac{\mathrm{Var}(\hat{a})}{\mathrm{Var}(a)} \tag{5}$$

[where did the square root of the denominator go? - see equation 4 above] The reliability ($r^2_{a,\hat{a}}$) can be interpreted as the part of the genetic variation that is explained by the estimated genetic predisposition whereas the remainder (1 - $r^2_{a,\hat{a}}$) is the uncertainty. Reliability of the genetic predisposition ($r^2_{a,\hat{a}}$) is important because it determines how well we can predict an individual's genetic predisposition. If $r^2_{a,\hat{a}}$ is low then we might consider more phenotypic records, in order to make better-informed precision health decisions.

2.4 Prediction error variance of the estimated genetic predisposition

Every prediction is associated with an error, and the same is true for the estimated genetic predisposition $\hat{a}$. The variability of the error for the predicted genetic predisposition is quantified by the prediction error variance (PEV) computed as:

$$\begin{aligned} Var(a - \hat{a}) &= Var(a) - 2Cov(a, \hat{a}) + Var(\hat{a}) = Var(a - \hat{a}) \\ &= Var(a)(1 - Var(\hat{a})Var(a)^{-1}) \\ &= Var(a)(1 - r^2_{a,\hat{a}}) \end{aligned}$$

The standard error of prediction (SEP) can be a useful quantity. SEP corresponds just to the square root of PEV. Therefore,

$$\begin{aligned} SEP(\hat{a}) &= \sqrt{Var(a - \hat{a})} \\ &= \sqrt{Var(a)(1 - r^2_{a,\hat{a}}))} \\ &= \sigma_a\sqrt{(1 - r^2_{a,\hat{a}}))} \end{aligned}$$

3 Estimation of genetic predisposition using pedigree information

Genetic predisposition is estimated using information on phenotypes and genetic relationships among individuals in a study population. We will illustrate the basic principles of genetic predisposition estimation using some simple examples where the trait has been measured on the individuals themselves or close relatives.

3.1 Estimation of genetic predisposition and accuracy based on own phenotype

An estimate of the genetic predisposition (a) based on own phenotype (y) can be calculated as:

$$\begin{aligned}
E(a|y) &= E(a) + Cov(a,y)[Var(y)]^{-1}(y - E(y)) \\
E(a|y) &= 0 + \sigma_a^2[\sigma_a^2 + \sigma_e^2]^{-1}(y - \mu) \\
E(a|y) &= h^2(y - \mu)
\end{aligned}$$

Thus, the estimated genetic predisposition using own phenotypic records can be computed based on an estimate of the trait heritability (h^2) and the observed phenotype deviation ($y - \mu$).

The expression for expected value terms ($E(a)$ and $E(y)$) in the equation above is based on rules for expected value of a sum of (normally distributed) random variables:

$$\begin{aligned}
\mathrm{E}(a) &= 0 \\
\mathrm{E}(e) &= 0 \\
\mathrm{E}(y) &= \mathrm{E}(\mu + a + e) \\
&= \mathrm{E}(\mu) + \mathrm{E}(a) + \mathrm{E}(e) \\
&= \mu + 0 + 0 \\
&= \mu
\end{aligned}$$

The expression for (co)variance terms ($Var(y)$, and $Cov(a,y)$) in the equation above is based on rules for the variance of a sum of (normally distributed) random variables:

$$\begin{aligned}
\mathrm{Var}(y) &= Var(a) + Var(e) + 2Cov(a,e) \\
\mathrm{Var}(a) &= \sigma_a^2 \\
\mathrm{Var}(e) &= \sigma_e^2 \\
\mathrm{Cov}(a,e) &= 0 \\
\mathrm{Var}(y) &= \sigma_a^2 + \sigma_e^2 \\
\mathrm{Cov}(a,y) &= Cov(a, a+e) \\
&= Cov(a,a) + Cov(a,e) \\
&= \sigma_a^2 + 0 \\
&= \sigma_a^2
\end{aligned}$$

The accuracy for the genetic predisposition based on own phenotype (y) can be calculated as:

$$\begin{aligned}
r_{a,\hat{a}} &= \frac{\mathrm{Cov}(a,\hat{a})}{\sqrt{\mathrm{Var}(a)}\sqrt{\mathrm{Var}(\hat{a})}} \\
r_{a,\hat{a}} &= \frac{(h^2)^2\sigma_y^2}{\sqrt{h^2\sigma_y^2}\sqrt{(h^2)^2\sigma_y^2}} \\
r_{a,\hat{a}} &= \sqrt{h^2}
\end{aligned}$$

The variance of the estimated genetic predisposition, $Var(\hat{a})$, can be expressed as:

$$\begin{aligned} Var(\hat{a}) &= Var(h^2(y-\mu)) \\ Var(\hat{a}) &= (h^2)^2 Var(y-\mu) = (h^2)^2 Var(y) = (h^2)^2 \sigma_y^2 \end{aligned}$$

The variance of the true genetic predisposition, $Var(a)$, can be expressed by the heritability and phenotypic variance:

$$\sigma_a^2 = (\sigma_a^2)/(\sigma_y^2)\sigma_y^2 = h^2\sigma_y^2.$$

Estimation of genetic predisposition based on own phenotype is only possible when the trait in question can be measured (directly or indirectly) on the individual.

3.2 Genetic relationship used for estimating genetic predisposition

Related individuals share genes and thus resemble each other, i.e., they have correlated phenotypic values to an extent that depends on additive genetic relationships. Consider a simple parent-offspring example. The offspring get one copy of the genome (*i.e.*, half of the diploid genome) from each parent, and therefore the genetic predisposition for the offspring is the average of the parents' genetic predisposition (a), plus Mendelian deviation ($a_{\text{mendelian}}$), which is the part of the genetic predisposition that is due to random segregation of the genes from each parent:

$$a_{\text{child}} = \frac{1}{2}a_{\text{father}} + \frac{1}{2}a_{\text{mother}} + a_{\text{mendelian}}$$

The term $a_{\text{mendelian}}$ is necessary, because two fullsibs i and j, both having the same parents, receive different random samples of parental alleles. Hence, the genetic predisposition, a_i and a_j, of fullsibs, i and j, are not going to be the same. The Mendelian deviation reflects the random contribution of (Mendelian) segregation to genetic predisposition of individuals.

In this example, the $\frac{1}{2}$ refers to the additive genetic relationship which indicates that the offspring receives half of its alleles from each parent. The weight given to a specific source of information depends on the ["its" may be a better word - then it becomes: The weight given to a specific source of information depends on its additive genetic relationship with the individual] additive genetic relationship with the individual. Examples of different types of additive genetic relationships (A_{ij}) between the various sources (j) and the individual itself, *i.e.*, the individual to be evaluated (i), can be seen in the table below.

Table 1: Examples on additive genetic relationship (A_{ij}) between individual i and j.

Type of relative	A_{ij}
Self	1.0
Unrelated	0
Mother	0.5
Father	0.5
Grandparent	0.25
Child	0.5
Full-sib	0.5
Half-sib	0.25
Twins (MZ/DZ)	1/0.5
Cousin	0.0625

4 Best Linear Unbiased Prediction for estimating genetic predisposition using pedigree information

Genetic predisposition can be estimated using the Best Linear Unbiased Prediction (BLUP) method. BLUP estimates genetic predisposition using phenotypic information for individuals from a general pedigree (with arbitrary relationships among them). The estimation of genetic predisposition based on multiple sources of information must correct for the redundancy between the different sources of information (*e.g.*, redundant information provided by parents and grandparents). Moreover, it is necessary to adjust for environmental factors that may affect the phenotypes in the study populations.

The key idea behind BLUP is to estimate the identifiable environmental factors as fixed effects and to predict the genetic predisposition as random effects simultaneously in a linear mixed model. Here, mixed refers to the presence of two types of effects: fixed effects (identifiable effects from environmental or genetic factors) and random effects (non-identifiable effects from segregating genetic factors and fluctuating environmental conditions). The **BLUP** method was developed by Henderson (e.g. Henderson 1973a and Henderson 1975) and the properties of this method are directly incorporated into its name:

- **B** stands for **best** which means that the correlation between the true (a) and the predicted genetic predisposition ($\hat{a}$) is maximal or the prediction error variance ($Var(a - \hat{a})$) is minimal.
- **L** stands for **linear** which means the predicted genetic predisposition are linear functions of the observations (y)
- **U** stands for **unbiased** which means that the expected values of the predicted genetic predisposition are equal to the true genetic predisposition
- **P** stands for **prediction**

BLUP is based on a linear mixed model methodology, and estimates of genetic predisposition can be obtained by solving mixed model equations. The BLUP method also requires a genetic relationship matrix and estimates of variance components (e.g., σ_a^2 and σ_e^2).

BLUP approaches are widely used for estimation of genetic predisposition. The popularity of BLUP is not only due to the theoretical foundations of the method, but also the efficient algorithms developed by Henderson for computing predicted genetic predisposition, even in very large study populations.

4.1 Linear mixed model

The linear mixed model contains the observation vector for the trait(s) of interest (y), the **fixed effects** [should there be added (b)?] that explain systematic differences in y, and the **random genetic effects** a and **random residual effects** e.

A matrix formulation of a general model equation is:

$$y = Xb + a + e,$$

where

y : is the vector of observed values of the trait,

b : is a vector of fixed effects,

a : is a vector of random genetic effects,

e : is a vector of random residual effects,

X : is a known design matrix that relates the elements of b to their corresponding element in y.

In this statistical model the random effects (a and e) and the phenotypes (y) are considered to be random variables which follow a multivariate normal (MVN) distribution. In general terms the expectations of these random variables are:

$$\begin{aligned} a &\sim MVN(0,G) \\ e &\sim MVN(0,R) \\ y &\sim MVN(Xb,V) \end{aligned} \tag{6}$$

where $G = A\sigma_a^2$, and $R = I\sigma_e^2$ are square matrices of genetic and residual (co)variances among the individuals, respectively, and $V = A\sigma_a^2 + I\sigma_e^2$ is the overall phenotypic covariance matrix.

4.2 Estimating fixed and random effects in the linear mixed model

The goal of linear mixed model analysis is to estimate the fixed effects, b, and random genetic effects, a. This can be done using the BLUE and BLUP equations shown below.

The best linear unbiased estimator (BLUE) of $\hat{b}$ is:

$$\hat{b} = (X'V^{-1}X)^{-1}X'V^{-1}y. \tag{7}$$

The BLUP of a is:

$$\hat{a} = GV^{-1}(y - X\hat{b}). \tag{8}$$

.

The BLUP equation for the estimate of the genetic predisposition consists of three parts: the term, $y - X\hat{b}$, which shows that the observed phenotypic values are corrected for the fixed effects represented by $X\hat{b}$; $G = Cov(a, y)$ which is the covariance between the true genetic predisposition (a) and phenotypes (y); and $V^{-1} = [Var(y)]^{-1}$ which is the inverse of the phenotypic covariance matrix. which is similar to the expression shown earlier for the expected value of the genetic predisposition conditional on the observed phenotype y:

$$E(a|y) = Cov(a, y)[Var(y)]^{-1}(y - E(y)). \tag{9}$$

4.3 Mixed model equations

Estimates of the fixed and random effects in the linear mixed model (*i.e.*, $\hat{b}$ and $\hat{a}$) can also be obtained by solving the following system of equations simultaneously:

$$\begin{bmatrix} X^TR^{-1}X & X^TR^{-1}Z \\ Z^TR^{-1}X & Z^TR^{-1}Z + G^{-1} \end{bmatrix} \begin{bmatrix} \hat{b} \\ \hat{a} \end{bmatrix} = \begin{bmatrix} X^TR^{-1}y \\ Z^TR^{-1}y \end{bmatrix}. \tag{10}$$

The expression above is called a **mixed model equation** (MME). It no longer contains the inverse V^{-1} and hence the MME is much simpler to solve. Instead, the MME contains the inverses R^{-1} and G^{-1}, which are easier to invert: $R = I\sigma_e^2$ is often a very simple matrix, and $G = A\sigma_a^2$ is usually smaller than (or the same size as) V. As a consequence, whenever we have to estimate genetic predisposition using BLUP we will usually use the mixed model equation shown in [NB NB here is a reference to an equation - should be removed if equations in this document no longer are numbered] - Eq. (10).

4.4 Prediction error variance, reliability and accuracy

The precision of estimates of genetic predisposition can be calculated from the inverse of the left hand side of the mixed model equation (Eq. (10)):

$$\left[\begin{array}{cc} X'X & X'Z \\ Z'X & Z'Z + G^{-1}\lambda \end{array}\right]^{-1} = \left[\begin{array}{cc} C^{11} & C^{12} \\ C^{21} & C^{22} \end{array}\right], \tag{11}$$

where $\lambda = \frac{\sigma_e^2}{\sigma_2^2} = \frac{1-h^2}{h^2}$, C^{11} corresponds to the fixed effects $(X'X)$ and C^{22} to the random genetic effects $(Z'Z + A^{-1}\lambda)$.

The prediction error variance (PEV) for the genetic predisposition can be calculated as:

$$PEV = C^{22}\sigma_e^2, \tag{12}$$

or expressed for individual i is:

$$PEV_i = C_{ii}^{22}\sigma_e^2, \tag{13}$$

which is the diagonal element of the inverse of the coefficient matrix corresponding to individual i, multiplied by the residual variance.

To compute the reliability, $r_{\hat{a},a}^2$, we use the following relationship:

$$PEV_i = (1 - r_{\hat{a},a}^2)\sigma_a^2 = C_{ii}^{22}\sigma_e^2. \tag{14}$$

Therefore the reliability can computed as:

$$r_{\hat{a},a}^2 = 1 - C_{ii}^{22}\frac{\sigma_e^2}{\sigma_a^2} \tag{15}$$

$$= 1 - C_{ii}^{22}\lambda, \tag{16}$$

and the accuracy computed as:

$$r_{\hat{a},a} = \sqrt{1 - C_{ii}^{22}\lambda}. \tag{17}$$

4.5 [consider moving this section - or change the heading] BLUPs are useful for ranking and decision making in precision health

[I didn't look at this section] BLUP estimates of genetic predisposition, especially from the linear mixed model including all relationships, is a useful tool in precision health. Identification of individuals with an extreme genetic predisposition based on BLUP estimates maximizes the probability for correct ranking of individuals and decisions in precision health. There are many factors contributing to this:

- The linear mixed model which makes full use of information from all relatives increases accuracy (precision) of the estimated genetic predisposition.

- The genetic predisposition are adjusted for systematic environmental effects in an optimal way.
- Several traits can be analyzed simultaneously.

However, estimation of genetic predispostion is based on phenotypic observations, and that regardless of how great the BLUP procedure may be, it cannot compensate for bad data. So a good health recording system is necessary for a reliable genetic evaluation. Furthermore, BLUP assumes that the genetic parameters used are the true ones. In practice, this means that genetic predisposition will only be accurate if the estimated genetic parameters are close enough to their true value.

5 Estimation of genetic predisposition using genomic information

In recent, years much attention has been given to genomic information due to the technological advancements in genotyping platforms. Today, dense genetic maps of single nucleotide polymorphisms (SNP) are available, and they enable us to divide the entire genome into thousands of relatively small chromosome segments.

Genetic predisposition is conceptually simple to calculate. First, the entire genome is divided into small chromosome segments by dense markers. Second, the additive effects of each chromosome segment are estimated simultaneously. Finally, the genetic predisposition is calculated as the sum of all chromosome segment effects. The chromosome segment effects are estimated for a group of individuals (*i.e.*, a reference or training population). For any remaining individual, only a blood or tissue sample is needed to determine its genetic predisposition. For use in precision health, it is desirable that the genetic predisposition can be estimated accurately and early in the individual's life. The effect of each of these small chromosome segments can be estimated if we have phenotypes and genotypes from many individuals (from several hundreds to hundreds of thousands or even millions). With sufficiently dense marker maps, the chromosome segment effects capture the genomic variability in the population in which they were estimated, because markers are in linkage disequilibrium with the causal gene that they bracket.

6 BLUP for estimating genetic predisposition using genomic information

We will present two approaches that are commonly used to estimate genetic predisposition using genomic information. The first approach is referred to as MBLUP. In this approach, marker effects are estimated from observed phenotypic and genomic marker data recorded in the reference (or training) population. Genetic predisposition is estimated from the marker effects and genomic marker data for individuals in the test population. The second approach is referred to as GBLUP. In this approach, genetic predisposition is estimated from observed phenotypic and genomic marker data for individuals in the reference (or training) population. Genetic predisposition for individuals in the test population is estimated based on their **genomic relationship** to individuals in a reference population. Both approaches allow for estimation of genetic predisposition for individuals without phenotypes and close relationships. This is one of the main advantages with genomic prediction. As soon as DNA is available for an individual, its marker genotypes can be determined and their genetic predisposition can be estimated. Furthermore, genetic predisposition estimated from genomic information is generally more accurate than when estimated from pedigree information.

6.1 A linear mixed model for estimating marker effects (MBLUP)

The linear mixed model for estimating marker effects contains the observation vector for the trait(s) of interest (y), the fixed effects (b) which explain systematic differences in y, the random marker effects (s), and the random residual effects (e). A matrix formulation of a general linear mixed model for estimating marker effects is:

$$y = Xb + Ms + e, \tag{18}$$

where

y : is the vector of observed values of the trait,

X : is a known design matrix that relates the elements of b to their corresponding element in y.

b : is a vector of fixed effects,

M : is a known design matrix that relates the elements of s to their corresponding element in y.

s : is a vector of random marker effects,

e : is a vector of random residual effects.

In the linear mixed model specified above, the marker and residual effects (s and e) and the phenotypes (y) are considered to be random variables following a multivariate normal (MVN) distribution:

$$\begin{aligned} s &\sim MVN(0, S) \\ e &\sim MVN(0, R) \\ y &\sim MVN(Xb, V), \end{aligned}$$

where $S = I_s\sigma_s^2$ is a square matrix of (co)variances among marker effects (usually assumed independent), and $R = I\sigma_e^2$ is a square matrix of residual (co)variances among residuals (also assumed independent, most of the times), and $V = MM'\sigma_s^2 + I\sigma_e^2$ is the overall phenotypic covariance matrix.

The marker variance σ_s^2 is defined as:

$$\sigma_s^2 = \frac{\sigma_a^2}{\sum_{j=1}^{m} 2p_j(1-p_j)}, \tag{19}$$

where σ_a^2 is the total additive genetic variance, m is the number of markers, and p_j is the frequency of the marker-allele that is associated with the effect QTL-allele.

6.1.1 Estimation of marker effects in MBLUP

Estimates of the fixed effects, b, and random marker effects, s, in the linear mixed model specified above can be done using the BLUE and BLUP equations shown below.

The best linear unbiased estimator (BLUE) of the fixed effects, $\hat{b}$, is:

$$\hat{b} = (X'V^{-1}X)^{-1}X'V^{-1}y. \tag{20}$$

The best linear unbiased prediction (BLUP) of the marker effects, $\hat{s}$, is:

$$\hat{s} = SM'V^{-1}(y - X\hat{b}). \tag{21}$$

The BLUP equation for estimating marker effects consists of three parts: (1) the term $y - X\hat{b}$ shows that the observed phenotypic values are corrected for the fixed effects represented by $X\hat{b}$, (2) the covariance between the true marker effects (s) and phenotypes (y) described as $Cov(s, y) = SM' = M'\sigma_s^2$ and (3) the inverse of the phenotypic covariance matrix described as $[Var(y)]^{-1} = V^{-1}$. Alternatively, estimates of the (fixed and random) effects in the model can be obtained by solving the mixed model equations. The mixed-model equations for the model given in (18) have the following structure:

$$\begin{bmatrix} X^TR^{-1}X & X^TR^{-1}M \\ M^TR^{-1}X & M^TR^{-1}M + S^{-1} \end{bmatrix} \begin{bmatrix} \hat{b} \\ \hat{s} \end{bmatrix} = \begin{bmatrix} X^TR^{-1}y \\ M^TR^{-1}y \end{bmatrix}. \tag{22}$$

6.1.2 Estimation of genetic predisposition in MBLUP

The estimates of the marker effects $\hat{s}$ in (22) can be used to estimate genetic predisposition $\hat{a}$ for any individual with genomic information (*i.e.*, genotypes for the same set of markers used in the reference population) by:

$$\hat{a} = \sum_{j=1}^{m} M_j\hat{s}_j, \tag{23}$$

where M_j corresponds to the vector of observed marker genotypes of an individual.

6.1.3 Encoding of the marker genotype matrix M

The elements in the matrix M [perhaps rather: "The elements in the M matrix" - perhaps it should be repeated that the M matrix is the marker genotype matrix? - similar to Genomic Relationship Matrix - see below] can be encoded in different ways. The results from the genotyping laboratory represents the nucleotides found at a given genome position. The nucleotides (genotypes) at each position (marker locus) must be encoded numerically for it to be used in the linear model. Let us assume that at a given SNP-position, the bases G or C are observed and G corresponds to the allele with the positive effect on our trait of interest. Based on the two observed alleles, the possible genotypes are GG, GC or CC. One possible code for this SNP in the matrix M might be the number of G-alleles, which corresponds to 2, 1 and 0. Alternatively, it is also possible to use the codes 1, 0 and -1 instead, which corresponds to the factors with which a is multiplied to get the genotypic values in the single locus model.

6.2 A linear mixed model for estimating genetic predisposition (GBLUP)

The linear mixed model for estimating genetic predisposition (GBLUP) contains the observation vector for the trait(s) of interest (y), the fixed effects b that explain systematic differences in y, and the random genomic effects a and random residual effects e. A matrix formulation of the GBLUP model is:

$$y = Xb + Za + e, \tag{24}$$

where

$$\begin{aligned} y &: \text{is the vector of observed values of the trait,} \\ b &: \text{is a vector of fixed effects,} \\ a &: \text{is a vector of random genomic effects,} \\ e &: \text{is a vector of random residual effects,} \\ X &: \text{is a known design matrix that relates the elements of } b \text{ to their corresponding element in } y, \\ Z &: \text{is a known design matrix that relates the elements of } a \text{ to their corresponding element in } y. \end{aligned}$$

In the linear mixed model (specified above) the genomic and residual effects (a and e) and the phenotypes (y) are considered to be random variables that follow a multivariate normal (MVN) distribution:

$$\begin{aligned} a &\sim MVN(0, \tilde{G}) \\ e &\sim MVN(0, R) \\ y &\sim MVN(Xb, V), \end{aligned}$$

where $\tilde{G} = G\sigma_a^2$, and $R = I\sigma_e^2$ are square matrices of genomic and residual (co)variances among the individuals, respectively, and $V = G\sigma_a^2 + I\sigma_e^2$ is the overall phenotypic covariance matrix. The genomic relationship matrix (G) is estimated from genomic marker data instead of pedigree information.

6.2.1 Estimation of genomic effects in GBLUP

Estimates of fixed effects (b) and random genomic effects (a) in the linear mixed model specified above are obtained using the BLUE and BLUP equations shown below.

The best linear unbiased estimator (BLUE) of the fixed effects ($\hat{b}$) is:

$$\hat{b} = (X'V^{-1}X)^{-1}X'V^{-1}y. \tag{25}$$

The best linear unbiased prediction (BLUP) of the genomic effects ($\hat{a}$) is:

$$\hat{a} = \tilde{G}Z'V^{-1}(y - X\hat{b}). \tag{26}$$

The BLUP equation for the estimate of the genomic effects consists of three parts: (1) the term $y - X\hat{b}$ shows that the observed phenotypic values are corrected for the fixed effects as represented by $X\hat{b}$, (2) the covariance between the true genomic effects (a) and phenotypes (y) described as $Cov(a, y) = \tilde{G}Z' = GZ'\sigma_a^2$ and (3) the inverse of the phenotypic covariance matrix described as $[Var(y)]^{-1} = V^{-1}$. Alternatively, estimates of the (fixed and random)[why is "fixed and random" in brackets?] effects in the model can be obtained by solving the mixed model equations which have the following structure:

$$\begin{bmatrix} X^TR^{-1}X & X^TR^{-1}Z \\ Z^TR^{-1}X & Z^TR^{-1}Z + \tilde{G}^{-1} \end{bmatrix} \begin{bmatrix} \hat{b} \\ \hat{a} \end{bmatrix} = \begin{bmatrix} X^TR^{-1}y \\ Z^TR^{-1}y \end{bmatrix}. \tag{27}$$

From (26) we can see that the GBLUP estimation procedure looks very similar to estimation of genetic predisposition based on pedigree information (PBLUP). In GLUP [should this be GBLUP?] the covariances between genetic predisposition is based on the genomic relationship matrix G which is computed from genomic markers whereas in PBLUP it is based on the numerator relationship matrix A computed from pedigree information.

6.2.2 Estimation of genetic predisposition for individuals in the test population in GBLUP ["in" GBLUP or "with" GBLUP or "using" GBLUP?]

The estimates of the genetic predisposition $\hat{a}_1$ for individuals with phenotypes [should this be: "for individuals with phenotypic information" - all individuals have phenotypes?] can be used to estimate genetic predisposition $\hat{a}_2$ for individuals with only genomic information [for individuals that only have genomic information available OR "for those individuals where only genomic information is available"]. Under the assumption of multivariate normality for the true genetic predisposition ($a \sim MVN(0, G\sigma_a^2)$), the expected value of the estimated genetic predisposition for individuals without phenotypes (a_2) conditional on the genetic predisposition for individuals with phenotypes ($\hat{a}_1$) can be written as:

$$\hat{a}_2 = \tilde{G}_{12}\tilde{G}_{11}^{-1}\hat{a}_1. \tag{28}$$

The equation given in (28) consists of three parts: [check if you agree - I did the same for the previous "consists of three parts" descriptions] (1) the term $\hat{a}_1$ represents the genetic predisposition for individuals with phenotypes in the reference population, (2) the covariance between the true genetic predisposition for invididuals without phenotypes (a_2), for individuals in the test population, and individuals with phenotypes (a_1) in the training population is given as $Cov(a_1, a_2) = \tilde{G}_{12} = G_{12}\sigma_a^2$ and (3) the inverse of the genomic covariance matrix for individuals with phenotypes in the training population is given as [I added "given as" - but perhaps it could be "described as"?] $\tilde{G}_{11}^{-1} = \sigma_a^{-2}G_{11}^{-1}$.

6.2.3 Genomic Relationship Matrix G

The additive genomic relationship matrix, G, is constructed using all genomic markers (or potentially pruned for LD) as follows:

$$G = \frac{WW^T}{\sum_{i=1}^{m} 2p_i(1 - p_i)}, \tag{29}$$

where W is the centered and scaled genotype matrix, and m is the total number of markers. Each column vector of W was calculated as follows: $w_i = M_i - 2p_i - 0.5$, where p_i is the minor allele frequency of the i'th genomic marker and M_i is the i'th column vector of the allele count matrix (M) [earlier the marker genotype matrix was called M - it should be consistent] which contains the genotypes coded as 0, 1 or 2 counting the number of minor alleles. The centering of the allele counts and the scaling factor, $\sum_{i=1}^{m} 2p_i(1-p_i)$, ensures that the genomic relationship matrix (G) has similar properties as the numerator relationship matrix (A).

The main difference between the two types of genetic relationship matrices A and G, is that A is based on the concept of identity by descent (sharing of the same alleles, transmitted from common ancestors), whereas G is based on the concept of identity by state (sharing of the same alleles, regardless of their origin).

6.3 Accuracy of genetic predisposition

[Genetic predisposition estimated using genomic information is generally more accurate than genetic predisposition estimated based only on pedigree information. One of the reasons for this is that the genomic relationship matrix more efficient use of phenotypic information for all individuals (based on degree of allele sharing) in the estimation procedure.] Perhaps this could be rewritten as follows: Genetic predisposition that is estimated by using genomic information is generally more accurate than genetic predisposition estimated by using only pedigree information. One of the reasons for this is the genomic relationship matrix's more efficient use of phenotypic information for all individuals (based on degree of allele sharing) in the estimation procedure.

The accuracy of genetic predisposition estimates is trait specific and depends on the heritability and the number of phenotypic records. [The accuracy of genetic predisposition increases when the size of the training population increases, when the reference population represents as much of the relevant genetic variation in the population as possible, when individuals in the test population are closely related to the reference population, when genetic diversity in the population is low (*i.e.*, low effective population size) and with better statistical models.] Perhaps this could be rewritten as follows: The accuracy of genetic predisposition increases when (1) the size of the training population increases, (2) the reference population represents as much of the relevant genetic variation in the population as possible, (3) individuals in the test population are closely related to the reference population, (4) the genetic diversity in the population is low (*i.e.*, low effective population size) and (5) better statistical models are applied.