Estimation of Genetic Predisposition

Palle Duun Rohde, Izel Fourie Sørensen & Peter Sørensen

2022-03-18

Contents

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 both by genotype and environment. Therefore, we need a way to infer the genetic predisposition from the phenotypic value. This section introduces the basic concepts of estimating genetic predisposition estimation such as:

- basic principle behind estimating genetic predisposition
- accuracy of estimated genetic predisposition
- use of genetic relationships for estimating genetic predisposition
- the connection between genetic parameters and 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 is the sum of both genetic and environmental factors. In general 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 for estimating genetic predisposition. We will derive a general approach for predicting genetic predisposition for any situation. Even though the procedure is general we will use a simple example to describe it.

2.1 Genetic model

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

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

where μ is the population mean, a is the genetic predisposition (i.e. additive effect), d is the dominance effect, i is the epistasis 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 (dominance and epistasis) are degraded by recombination and are not inherited, even though they may be important for the individual's phenotype. Furthermoe,additive genetic effects explain a majority of the total genetic variance. Therefore we only consider the additive genetic model as the basis for genetic predisposition estimation:

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

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 loci j. We therefore 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 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 unknown parameter using statistical modelling expresses the estimated quantity as a mathematical function of the observed data. The question is how this function should look like and what properties the estimated genetic predisposition should fulfill. Under the assumption of multivariate normality for a and y (which are justified under the central limit theorem and the assumptions of many genetic and environmental factors), the expected value of the genetic predisposition 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:

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

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. 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 the terms $E(a)$, $E(y)$, $Var(y)$, and $Cov(a, y)$ in 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:

a = additive genetic value = true genetic predisposition

$\hat{a} = E(a|y)$ = estimated additive genetic value = estimated genetic predisposition

2.2.1 Estimated genetic predisposition are unbiased

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

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

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

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.

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

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 important to determine 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 accuray 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:

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

Therefore 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 (PEV) of estimated genetic predisposition

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

$$Var(a - \hat{a}) = Var(a)(1 - r^2_{a,\hat{a}}) \tag{6}$$

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

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

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

3 Estimation of genetic predisposition using phenotypic data and pedigree information

Genetic predisposition is estimated using information on phenotypes and genetic relationships for 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 record 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 are 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 are based on rules for the variance of a sum of (normally distributed) random variables:

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

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:

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

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 Estimation of genetic predisposition and accuracy based on phenotypes of close relatives:

It may be possible to use phenotypic records from close relatives, such as, half-sibs, full-sibs, parents and grandparents. Phenotypes collected on close relatives (as compared to distant relatives) provide more information about the genetic predisposition of an individual (as close relatives share more DNA in common). In the following we will provide a general formula for estimating genetic predisposition and their accuracies using phenotypic information on different types of relatives.

General formula for estimating genetic predisposition using different sources of information:

$$\hat{a} = b_{a|y}(y-\mu) \tag{7}$$

where the regression coefficient quantifies the weight (or importances) of the phenotypic information:

$$b_{a|y} = \frac{a' n h^2}{(1+(n-1)r} \tag{8}$$

where a' is the genetic relationship between a specific individual and individuals with phenotypes, n is the number of phenotypic records, h^2 is the trait heritability, and r is correlation between individuals with observations ($r = a^{''}h2 + c^2$, where $a^{''}$ = genetic relationship between individuals with records and target, c^2 = common environmental component).

Thus the importance given to a specific source of information depends on the additive genetic relationship (a') with the individial, the heritability of the trait (h^2), and the amount of information (n), i.e. the number of relatives (progenies or sibs, etc.).

General formula for reliability of estimated genetic predisposition using different sources of information:

$$r_{a,\hat{a}}^2 = \frac{(a')^2 n h^2}{1+(n-1)r} \tag{9}$$

Thus reliability depends on the same factors as the estimated genetic predisposition except for the phenotypic value. Although the reliability depends on the number of records it does not depend on the numerical value of phenotypes. From this formula is it clear that higher reliability (and accuracy) can be achived when:

- genetic relationship to individuals with information (a') is high
- there are many records (n is high)
- heritability (h^2) is high
- correlation between records ($r = a''h^2 + c^2$) is low (little redundancy in observations)

3.3 Genetic relationship used for estimating genetic predisposition

Related individuals share genes and thus resemble each other (have correlated phenotypes, to an extent that depends on additive genetic relationships). Consider a simple parent-offspring example. The offspring get half of the genes from each parent and therefore the genetic predisposition for the offspring is the average of the parents' genetic predisposition plus the Mendelian deviation (the part of the genetic predisposition that is due to random segregation of the genes from each parent):

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

(a = additive genetic value = genetic predisposition

The term $a_{mendelian}$ is necessary, because two fullsibs i and j both having parents $father$ and $mother$ 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 that random contribution of (Mendelian) segregation to genetic predisposition of individuals.

In this equation the $\frac{1}{2}$ refers to the additive genetic relationship which in this example indicates that the offspring receives half of its genes from its parent. In general the weight given to a specific source of information depends on the 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.

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

3.4 Estimation of genetic predisposition using phenotypic information from multiple sources

Several factors influence which sources of information to use when estimating genetic predisposition for a trait: what information is available, the heritability of the trait, and how and on which individuals the trait can be measured. Therefore in practice it is common to combine information from several sources. As already mentioned, all information available is usually utilized when an individual's genetic predisposition is predicted. The weight given to a specific source of information depends on the additive genetic relationship with the individual, the heritability and the amount of information, i.e. the number of progenies or sibs, etc.

- Phenotypic records on the individual's sibs, half-sibs and full-sibs, are often used in addition to other information, or to give supplementary information, for example on traits that cannot be measured on the individual itself. The accuracy of sib testing depends on the number of sibs that have records. Common-environment effects (e.g. full-sibs raised in the same family) may bias the estimation of genetic predisposition, unless we are able to adjust for them.

- Parental information at different generations (parents, grandparents, etc.) is generally available even before the individual is born, and can provide information very early. However, the genes from each locus of the parents are transmitted at random, so information based on pedigree alone is not very accurate, but can be valuable as additional information. Moreover, the additive genetic relationship, and thus the proportion of common genes between the individual and the pedigree, is halved for every generation backwards (at least 0.5 for a parent, 0.25 for a grandparent, etc.). Finally, there is redundancy in the information provided by different generations of parents. Fore example, if there are accurate estimate of the parent's genetic predisposition, then there is little to gain in using information on grandparents (actually, if the parents true genetic predisposition are known, there would be no additional gain of information from grandparents).

As already mentioned, all information available is usually utilized when an individual's genetic predisposition is predicted. The weight given to a specific source of information depends on the additive genetic relationship with the individual, the heritability and the amount of information, i.e. the number of relatives (progenies, sibs, parents, etc.). In the following sections we will show how genetic predisposition can be calculated when different types of phenotypic information (from different types of relatives) are available.

4 BLUP a general approach for estimation of genetic predisposition using pedigree information

Genetic predisposition can bne estimated using the BLUP (Best Linear Unbiased Prediction) method. BLUP allows for estimation of genetic predisposition using phenotypic information for individuals from a general pedigree (with arbitrary relationships among them). BLUP is based on linear mixed model methodology, and estimates of genetic predisposition can be obtained by solving the mixed model equations. The BLUP method also require a genetic relationship matrix and estimates of variance components (e.g., σ_a^2 and σ_e^2).

The estimation of genetic predisposition based on multiple sources of information must correct for the redundancy between them (e.g., the redundant information provided by parents and grandparents). Moreover, they need to be adjusted for average effects in the populations, "fixed effects".

The BLUP solution to these problems was presented by Charles R. Henderson in several publications (e.g. Henderson1973a and Henderson1975). The key idea behind the solution 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 methodology developed by Henderson is called **BLUP** and the properties of this methodology 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 approaches are widely used for estimation of genetic predisposition. The popularity of BLUP is not only due to the theoretical foundations behind BLUP, 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** 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 the statistical model (specified above) 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:

$$\begin{aligned} a &\sim MVN(0, G) \\ e &\sim MVN(0, R) \\ y &\sim MVN(Xb, V) \end{aligned} \tag{10}$$

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 the BLUP analysis is the estimate the fixed effect b, and random genetic effects, a, in the linear mixed model specified above. This can be done using the BLUE and BLUP equations shown below.

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

$$\hat{b} = (X'V^{-1}X)^{-1}X'V^{-1}y \tag{11}$$

The matrix $(X'V^{-1}X)^{-1}$ denotes the inverse of the matrix $(X'V^{-1}X)$.

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

$$\hat{a} = GV^{-1}(y - X\hat{b}) \tag{12}$$

which is similar to the expression shown earlier for the expected value of the genetic predisposition conditional on the observed phenotype y:

$$E(a|y) = Cov(a, y)[Var(y)]^{-1}(y - E(y)) \tag{13}$$

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

4.3 Mixed Model Equations

The solutions shown in (25) for $\hat{a}$ and in (24) for $\hat{b}$ are not suitable for practical purposes. Both solutions contain the inverse V^{-1} of matrix V. The matrix V corresponds to the variance-covariance matrix of all observations y. The inverse matrix V^{-1} is not easy to compute. Furthermore, procedures to invert general matrices are computationally expensive and are prone to rounding errors. In one of his many papers, Henderson has shown that the results for $\hat{a}$ and $\hat{b}$ are the same when solving the following system of equations simultaneously:

$$\begin{bmatrix} X^T R^{-1} X & X^T R^{-1} Z \\ Z^T R^{-1} X & Z^T R^{-1} Z + G^{-1} \end{bmatrix} \begin{bmatrix} \hat{b} \\ \hat{a} \end{bmatrix} = \begin{bmatrix} X^T R^{-1} y \\ Z^T R^{-1} y \end{bmatrix} \tag{14}$$

The above shown equations are called **mixed model equations** (MME). They no longer contain the inverse V^{-1} and hence these MME are much simpler to solve. Instead, the MME contain 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 equations shown in (14).

In order to solve the mixed model equations (or BLUP equations), the additive genetic relationship matrix A and estimates of the variance components (i.e. σ_a^2 and σ_e^2) are required. The additive genetic relationship matrix A can be computed using a recursive method from a pedigree of the individuals in the study population. The variance components can be estimated using the REML method based on phenotypes and genetic relationships for individuals in the study population.

4.4 BLUP genetic predisposition are useful for ranking and decision making in precision helath

BLUP estimates of genetic predisposition (EVBs), especially from the linear mixed model including all relationships, are useful tools 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 that contribute 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. This means that individuals can also be compared across environments, age classes, assuming the data is connected
- Several traits can be analyzed simultanously

It should, however, be noted that the genetic evaluation 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. It should also not be forgotten that BLUP assumes that the genetic parameters used are the true ones. In practice that 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 phenotypic data and genomic information

In recent years much attention has been given to genomic information due to the dramatic development in genotyping technologies. Today dense genetic maps are available for based on DNA markers in the form of single nucleotide polymorphisms (SNP) and they enable us to divide the entire genome into thousands of relatively small chromosome segments.

5.1 Genomic markers

The different locations in the genome that are considered in genomic selection are called **markers**. When looking at the complete set of markers making up the genomic information in a population, the so-called **Single Nucleotide Polymorphisms** (SNPs) have been shown to be the most useful types of markers. These SNPs correspond to differences of single bases at a given position in the genome. Based on empirical analyses of very many SNP-loci, almost all SNP just take two different states. Furthermore it is important that these SNPs are more or less evenly spread over the complete genome. Some SNPs may be located in coding regions and some my be placed in regions of unknown function.

5.2 Quantitative Trait Loci and linkage disequilibrium

The loci that are relevant for the trait are called **Quantitative Trait Loci** (QTL). Any given SNP-Marker can only be informative for a given QTL, if a certain **linkage disequilibrium** between the QTL and the marker locus exists. The idea behind linkage disequilibrium is that a certain positive QTL-allele evolved in a certain genetic neighborhood of a number of SNP loci. As a result of that the positive QTL-allele is very often inherited with the same SNP-allele. Over the generations, recombination between the QTL and the neighboring SNP-loci can happen and thereby weaken the statistical association between the positive QTL-allele and the given SNP-allele. This recombination effect is smaller when the QTL and the SNP-loci are physically closer together on the chromosome. The non-random association between QTL and SNP-markers is called linkage disequilibrium.

The marker locus is called M and the QTL is called Q, then the LD can be measured by

$$D = p(M_1Q_1) * p(M_2Q_2) - p(M_1Q_2) * p(M_2Q_1) \tag{15}$$

where $p(M_xQ_y)$ corresponds to the frequency of the combination of marker allele M_x and QTL allele Q_y. Very often the LD measure shown in (15) is re-scaled to the interval between 0 and 1 which leads to

$$r^2 = \frac{D^2}{p(M_1) * p(M_2) * p(Q_1) * p(Q_2)} \tag{16}$$

In (16) r^2 describes the proportion of the variance at the QTL which is explained by the marker M. Hence the LD must be high such that the marker can explain a large part of the variance at the QTL. For humans more than $500,000$ SNP markers are required to get a sufficient coverage of the complete genome.

6 BLUP a general approach for estimation of genetic predisposition using genomic information

Estimation of genetic predispositions is conceptually simple to calculate from genomic information. 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.

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 estimated based on their **genomic relationship** to individuals in the reference population. Both approaches allow for estimation of genetic predisposition for individuals without phenotypes and close relationships. This is one of the main advantages the genomic prediction. As soon as DNA is available for an individual, its marker genotypes can be determined and the genetic predisposition can be estimated. Furthermore, genetic predisposition estimated using genomic information is generally more accurate than genetic predisposition estimated based only on 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:

$$y = Xb + Ms + e \tag{17}$$

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 above the marker and residual effects (s and e) and the phenotypes (y) are considered to be random variables which follow a multivariate normal (MVN) distribution. In general terms the distributions of these random variables are:

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

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:

$$\sigma_s^2 = \frac{\sigma_a^2}{\sum_{j=1}^{m} 2p_j(1-p_j)} \tag{18}$$

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 positive QTL-allele.

6.1.1 Estimation of marker effects in MBLUP

Estimates of the fixed effect 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:

$$\hat{b} = (X'V^{-1}X)^{-1}X'V^{-1}y \tag{19}$$

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

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

The BLUP equation for the estimate of the marker effects consists of three parts; The term, $y - X\hat{b}$, shows that the observed phenotypic values are corrected for the fixed effects represented by $X\hat{b}$. The covariance between the true marker effects (s) and phenotypes (y) is $Cov(s, y) = SM' = M'\sigma_s^2$. The inverse of the phenotypic covariance matrix is $[Var(y)]^{-1} = V^{-1}$. 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 (17) have the following structure:

$$\begin{bmatrix} X^TR^{-1}X & X^TR^{-1}M \\ M^TR^{-1}X & M^TR^{-1}M + S^{-1} \end{bmatrix} \begin{bmatrix} \hat{b} \\ \hat{s} \end{bmatrix} = \begin{bmatrix} X^TR^{-1}y \\ M^TR^{-1}y \end{bmatrix} \tag{21}$$

6.1.2 Estimation of genetic predisposition in MBLUP

The estimates of the marker effects $\hat{s}$ in (21) 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:

$$\hat{a} = \sum_{j=1}^{m} M_j\hat{s}_j \tag{22}$$

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 can be encoded in different ways. The results from the genotyping laboratory represents the nucleotides found at a given genome position. To be used in the linear model the nucleotides (genotypes) at each position (marker locus) must be encoded numerically. 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 a general GBLUP model is:

$$y = Xb + Za + e \tag{23}$$

where

y : is the vector of observed values of the trait,
b : is a vector of fixed effects,
a : is a vector of random genomic 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.
Z : is a known design matrix that relates the elements of a to their corresponding element in y.

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 which follow a multivariate normal (MVN) distribution. In general terms the distributions of these random variables are:

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

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 the fixed effect b, and random genomic effects, a, 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:

$$\hat{b} = (X'V^{-1}X)^{-1}X'V^{-1}y \tag{24}$$

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

$$\hat{a} = \tilde{G}Z'V^{-1}(y - X\hat{b}) \tag{25}$$

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

$$\begin{bmatrix} X^TR^{-1}X & X^TR^{-1}Z \\ Z^TR^{-1}X & Z^TR^{-1}Z + \tilde{G}^{-1} \end{bmatrix} \begin{bmatrix} \hat{b} \\ \hat{a} \end{bmatrix} = \begin{bmatrix} X^TR^{-1}y \\ Z^TR^{-1}y \end{bmatrix} \tag{26}$$

From (25) we can see that the GBLUP estimation procedure looks very similar to estimation of genetic predisposition based on pedigree information (PBLUP). In GLUP 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 test population in GBLUP

The estimates of the genetic predisposition $\hat{a}_1$ for individuals with phenotypes can be used to estimate genetic predisposition $\hat{a}_2$ for individuals with only genomic information. 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:

$$\hat{a}_2 = \tilde{G}_{12}\tilde{G}_{11}^{-1}\hat{a}_1 \tag{27}$$

The equation given in (27) consists of three parts; The term, $\hat{a}_1$, represent the genetic predisposition for individuals with phenotypes in the reference population. 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 $Cov(a_1, a_2) = \tilde{G}_{12} = G_{12}\sigma_a^2$. The inverse of the genomic covariance matrix for individuals with phenotypes in the training population is $\tilde{G}_{11}^{-1} = \sigma_a^{-2}G_{11}^{-1}$.

6.2.3 Genomic Relationship Matrix G

The additive genomic relationship matrix G is constructed using all genomic markers as follows:

$$G = \frac{WW^T}{\sum_{i=1}^{m} 2p_i(1-p_i)} \tag{28}$$

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, which contains the genotypes coded as 0, 1 or 2 counting the number of minor allele. The centering of the allele counts and 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

Furthermore, 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. The accuracy of genetic predisposition estimates is trait specific and depends on the heritability and the number of phenotypic records. In general the accuracy of genetic predisposition increases when the size of the reference 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.