Table of Contents

List of Tables

List of Figures

Chapter 1

Background

1.1 Generalized linear mixed models

A generalized linear mixed model (GLMM) is an extension of the GLM, the generalized linear model (Zuur et al. 2009; Pinheiro and Bates 2006; Galwey 2014; Faraway 2016). While incorporating the linear predictors of a GLM, GLMMs also allow for more random variable terms than merely the residual error. These random variable terms are called *random effects*, whereas the non-random terms are called *fixed effects*. Hence the designation of *mixed* models: they utilize a *mix* of fixed and random effects. Since the random effects do not take some determinate value, we seek to estimate the parameters that determine their distribution rather than the values of the random effects themselves.

Let us formulate a general GLMM in vector notation, with an arbitrary number of fixed and random effects. Letting $\boldsymbol{y}$ be the response vector, which we pass through some link function $\boldsymbol{f}$, the GLMM is given as

$$\boldsymbol{f}(\boldsymbol{y}) = \boldsymbol{\mu} + \mathbf{X}\boldsymbol{\beta} + \mathbf{Z}\boldsymbol{u} + \boldsymbol{\varepsilon}\ , \tag{1.1}$$

where $\boldsymbol{\mu}$ is an intercept vector, $\boldsymbol{\beta}$ is the vector of fixed effects and $\boldsymbol{u}$ is the random effect vector with some given multivariate distribution. The residual vector $\boldsymbol{\varepsilon}$ has distribution $\mathcal{N}(\mathbf{0}, \sigma_\varepsilon^2\mathbf{I})$, and $\boldsymbol{u}$ is usually also assumed to be multivariate normal. $\mathbf{X}$ and $\mathbf{Z}$ are design matrices for fixed and random effects, respectively, and relate the effects to the response appropriately.

As a simple example, take the random intercept model with a single fixed effect (Cohen et al. 2013). In this model we allow the intercept to differ between different groups, with each group intercept taking on a random value. Let y_{ij} be the response for observation j from group i, and f be the link function. If the intercept has mean μ and its stochastic part in group i is the random effect $u_i \sim \mathcal{N}(0, \sigma^2)$, then

$$f(y_{ij}) = \mu + x_{ij}\beta + u_i + \varepsilon_{ij}\ ,$$

where x_{ij} is a covariate corresponding to the fixed effect β and $\varepsilon_{ij} \sim \mathcal{N}(0, \sigma_\varepsilon^2)$ is the residual. Fitting the model would involve estimating μ, β, σ^2 and σ_ε^2.

So what is the purpose of including random effects? Take an example adapted from Galwey (2014, pp. 1–20). Imagine a study with repeated measurements, where several observations are taken from each subject, leading to a natural grouping of the data. This grouping should be taken into account by the model to ensure the independence of residuals, a central assumption of GLMs. One way to avoid this assumption being violated could be a model that instead uses the mean observed value for each subject, but we would naturally prefer to retain the statistical power of including all observations. Another approach would be to include a subject's identity as a fixed categorical covariate, thereby estimating a value that is to be added to the result for observations from a given subject. This method works, but may cost us many degrees of freedom if we have a lot of different subjects. Also, we are often not interested in the effect of each individual subject, but rather the greater population of subjects.

This is where random effects come into play. We can include a random effect $u_i \sim \mathcal{N}(0, \sigma_{\mathrm{u}}^2)$, which is independent and identically distributed (IID) between different subjects i. Fitting the model then involves estimating the variance σ_{u}^2, which says something about the between-subject variance of the larger population. This modeling decision allows us to include all available data, rather than using a mean for each subject, while still only taking up one degree of freedom. It also causes the residual random effect to only describe within-subject variance. Thus, the reason random effects are useful is to explain the response when the data contains a covariance structure between observations. Various forms of covariance structures can be modelled using random effects, not just repeated measurements. We can, for example, include hierarchical and nested structures, by making the random effects covary between observations in other ways (Faraway 2016, p. 195).

Whether a covariate should be considered a fixed effect or a random effect is not always clear, and the rules for making this choice are not universally agreed upon (Gelman 2005; Searle et al. 2006). The determinant of this choice might be either convenience or what aspects of the study system are of interest. One common convention is using fixed effects when all levels of a covariate are present in the data, or when we are interested in the value of the effect itself (Wilson et al. 2010). If not, we would model the covariate as a random effect. That is, if the effects has many levels and/or these levels are a randomly chosen subset of a larger set, or the variation in the greater population is of interest. Under this convention an obvious fixed effect might be the subject's sex, while the subject's identity in a study with repeated measurements is an obviously random effect. In other cases the choice is more ambiguous, such as when modeling the year of measurement for a study running over just a few years.

1.2 The animal model

The animal model (as described by Lynch and Walsh 1998; Kruuk 2004; Wilson et al. 2010; Mrode 2014), is a type of GLMM often applied in the field of quantitative genetics. A characteristic of the model is the inclusion of "genetic values" (also called "breeding values") as random effects to model some trait as a response. Assume N individual animals were measured during a study. An individual i's genetic value a_i denotes the impact of additive genetic effects on the individual's phenotype, i.e. the measured value of the trait. The source of non-independence considered by this random effect is the potential simi-

larity of two individuals' genomes, which can lead to similar impacts on the phenotypes. Closely related individuals are more likely to share genes, potentially causing phenotypes of relatives to be correlated. The issue is quantifying to what degree the variation in trait values can be attributed to an individual's genes.

To tease out this information, we base the covariance structure of the breeding values on relatedness information. Such a structure is obtained by having the genetic value vector $\boldsymbol{g}$ follow the multivariate normal distribution

$$\boldsymbol{g} \sim \mathcal{N}(\mathbf{0}, \sigma^2_{\mathrm{G}}\mathbf{G}) \, ,$$

where $\mathbf{G}$ is the symmetric $N \times N$ genetic relationship matrix (GRM). The entry G_{ij} of $\mathbf{G}$ contains a measure of how similar the genomes of individuals i and j are. For off-diagonal entries we usually have $G_{ij} \in (0, 1)$, where a high value denotes closely related individuals. For diagonal entries we usually have $G_{ii} \geq 1$, where the entries will be greater than 1 when inbreeding (i.e. mating of close relatives) is present. We can write $G_{ii} = 1 + F_i$, where F_i is denoted as individual i's "coefficient of inbreeding," a measure of how inbred i is (Wright 1922). There are many possible choices of GRM $\mathbf{G}$, as we shall explore below.

This covariance structure is scaled by σ^2_{G}, the additive genetic variance, which can be interpreted as the part of the variance in an individual's phenotype caused by additive genetic effects.[1] Thus, animal models are reliant on knowledge of the genome similarity between individuals, encoded by $\mathbf{G}$. From the definition of the distribution of the breeding value vector $\boldsymbol{a}$, it is clear that the breeding values of two animals will only strongly covary if they are closely related, and there is a high additive genetic variance present in the population. It is also clear that the estimated value of σ^2_{G} depends on our choice of $\mathbf{G}$. A simple animal model for phenotype y_i in induvidual i, containing only an intercept μ and random effect breeding values g_i, can be stated as

$$y_i = \mu + g_i + \varepsilon_i \, .$$

1.2.1 Measures of genome similarity

In the context of animal models, $\mathbf{G}$ has customarily been inferred from observed pedigrees (i.e. family trees). Knowing how closely related two individuals are, one can estimate the expected amount of shared genes between the two individuals. Animal models originated in the field of animal and plant breeding, where accurate pedigree records are readily available (Henderson 1984). In wild ecological systems pedigrees are harder to come by, as parentage must be observed in the field or inferred based on genetic marker information (Jones and Ardren 2003). In recent times, an alternative method of directly inferring genome similarity from the observed genotypes of SNP markers has grown in popularity (Gienapp et al. 2017; Speed and Balding 2015; Bérénos et al. 2014). This genomic approach has become a viable option due to improvement in genomic technologies (Meuwissen et al. 2016; Ødegård et al. 2018), as the cost of large-scale genotyping is steadily decreasing, and the knowledge of SNP markers develops.

[1] Non-additive genetic effects are usually neglected in quantitative genetics studies (Kruuk 2004).

Genome similarity inferred from pedigree-based relatedness

We denote the version of the GRM $\mathbf{G}$ that uses pedigree information as $\mathbf{A}$, which is also known as the "genetic relatedness matrix." $\mathbf{A}$ is defined so that its ijth entry $(A)_{ij}$ denotes twice the expected probability ρ_{ij} that an allele (a variation of a gene) picked at random from animal i is identical to, and originates from the same ancestor as, an allele picked at random from animal j (Wright 1922; Weir et al. 2006). This expected probability ρ_{ij} is commonly known as the "coefficient of coancestry" (Lynch and Walsh 1998, p. 135). If $\mathcal{A}$ is the set containing all of i and j's (known) most recent common ancestors, then define

$$(A)_{ij} = 2\rho_{ij} = 2\sum_{k\in\mathcal{A}} \frac{1+F_k}{2^{\phi_{ij}^k}} ,$$

where the inbreeding coefficient F_k is the coefficient of coancestry between k's parents and ϕ_{ij}^k is the number of individuals involved in the path in the pedigree linking i and j through ancestor $k \in \mathcal{A}$, including i and j themselves. By "most recent" common ancestor we mean that none of k's descendants are also common ancestors of i and j. We further consider individuals to be their own ancestors. In the absence of inbreeding, we have the following illustrative examples of coefficients of coancestry.

- $i = j$: here i is its own only most recent common ancestor, so $\mathcal{A} = \{i\}$. Because $\phi_{ii}^i = 1$, we end up with $\rho_{ii} = \frac{1}{2}$.
- i is a parent of j: again i is the only most recent common ancestor, so $\mathcal{A} = \{i\}$. However, $\phi_{ij}^i = 2$, and thus $\rho_{ij} = \frac{1}{2^2} = \frac{1}{4}$.
- i and j are full siblings: we now have two most recent common ancestors, the father s and mother d, giving $\mathcal{A} = \{s, d\}$. For the path through each parent $\phi_{ij}^s = \phi_{ij}^d = 3$, so $\rho_{ij} = \frac{1}{2^3} + \frac{1}{2^3} = \frac{1}{4}$.

When inbreeding *is* present these probabilities will be greater due to i and j sharing more ancestors, which increases the likelihood i and j's alleles originate from the same ancestor.

If we have a pedigree accurately describing the familial relationships in our study population, then $\mathbf{A}$ gives us a measure of relatedness between each individual in the pedigree without requiring direct knowledge about the genotypes at any of their loci. Other advantages include inbreeding begin accounted for explicitly, and the lack of assumptions being made on mating patterns or selection (Kruuk 2004). Furthermore, we do not in general impose any constraints on the shape of the pedigree, but the more well-connected the pedigree, the more informative it will be (Wilson et al. 2010). Methods, such as pedigree-based relatedness, that try to infer genetic relationships based on individual ancestries are commonly called Identity-by-descent (IBD) methods.

A central concept when using the animal model with relatedness inferred from a pedigree is the concept of a "base population," the population for which we estimate genetic parameters. For any pedigree we will inevitably have certain individuals with no known parents; they are the root nodes in the family tree. We label these unknown parents as phantom parents. Note that the phantom parents include not only the earliest cohort in the pedigree (known as the "founder population"), but also includes the phantom parents of later (non-founder) individuals for whom we are missing parentage data. The ensemble

of all phantom parents makes up the base population, about which we make the following assumption: they are entirely unrelated and all share the same genetic parameters and each only has one offspring (Wolak and Reid 2017; Wilson et al. 2010). Any relatedness measures based on pedigrees are relative to its base population (Lynch and Walsh 1998, p. 132), and the breeding values of the base population are assumed to have a baseline mean of zero. Therefore, the animal model estimates σ^2_{G} for individuals in the base population and not the population as a whole. Furthermore, the breeding value of any non-base individual can be interpreted as its genetic deviation from the base population.

Thus, if we have a specific subpopulation for which we wish to measure the genetic parameters, we might choose our pedigree so that the base population will be the subpopulation of interest. This would be done by disregarding the ancestors of members of this subpopulation. Either way the base population will necessarily be somewhat arbitrary, whether it is determined by a deliberate choice of base population or by the constraints of our data collection. Such an arbitrary choice is nonetheless necessary, since the consequence of adding ever more ancestors to some pedigree would be ϕ converging to 1 for individuals far down the pedigree (Speed and Balding 2015). The cut-off must occur at some point.

SNP-based genome similarity measures

A weakness of the coefficient of coancestry is that actual (realized) relatedness between individuals can vary greatly from the expectation denoted by ϕ (Hill and Weir 2011). The probability of choosing two alleles that are identical by descent can be much greater or lower than what is indicated by the pedigree-derived ϕ. Furthermore, errors in observed pedigrees are not uncommon, and might bias the results in unexpected ways (Kruuk 2004). To get a more accurate measure of genome similarity, we might therefore use realized relatedness rather than expected relatedness. This requires a direct comparison of genotypes between individuals, so-called identity-by-state (IBS) methods. However, the genomes of two individuals of the same species are usually very similar; for example, in humans, 1000 Genomes Project Consortium (2015) found that two genomes typically differed at only 0.6% of their base pairs of nucleotides that make up the full genome. Therefore, when comparing genomes, we limit our focus to the loci (specific positions on a chromosome) where the genotypes *do* vary.

A single nucleotide polymorphism, or SNP, is a genetic marker where the second most common nucleotide occurs in a non-trivial proportion of the population. We will only consider diallelic locus, i.e. a specific position on a chromosome that only has two possible alleles. Denote the most common allele the "major allele" and the other (second most common) allele as the "minor allele". Thus, we consider a SNP to be present at a locus if the rate of occurrence of the minor allele, the minor allele frequency (MAF), is sufficiently large (e.g. 1% or 5%) on that locus.

If we have knowledge about the genotypes of M SNPs for each individual in a population of size N, we can define the $N \times M$ genotype matrix $\mathbf{V}$. The entries of this matrix have values $v_{im} \in \{0, 1, 2\}$ and denote the number of copies of the minor allele. Thus when $v_{im} = 0$ individual i's mth SNP is homozygous with two copies of the major allele, when $v_{im} = 1$ the SNP is heterozygous with one copy of each allele, and when $v_{im} = 2$ the SNP is homozygous with two copies of the minor allele. There are many possible def-

initions of SNP-based GRMs, but they all derive from the genotype matrix in some way (Speed and Balding 2015). Many of these definitions also include a SNP m's MAF p_m to weigh the importance of each SNP; two individuals sharing a minor allele with a very low MAF carries more information than sharing a minor allele that is almost just as likely as the major allele.

For one, the relationship matrix presented by VanRaden (2008) is widely used (Crossa et al. 2017; see e.g. Makgahlela et al. 2013; Rio et al. 2020a). This relationship matrix, which we will mark by $\mathbf{G}_{\text{VR}}$, is defined as

$$(G_{\text{VR}})_{ij} = \frac{\sum_{m=1}^{M} (v_{im} - 2p_m)(v_{jm} - 2p_m)}{2\sum_{m=1}^{M} p_m (1 - p_m)} = \frac{(\mathbf{V}_i - 2\boldsymbol{p})(\mathbf{V}_j - 2\boldsymbol{p})^\top}{2\sum_{m=1}^{M} p_m (1 - p_m)} ,$$

where $\mathbf{V}_k$ denotes the kth row of $\mathbf{V}$, and $\boldsymbol{f}$ is the vector of MAFs. In other words,

$$\mathbf{G}_{\text{VR}} = \tilde{\mathbf{V}}\tilde{\mathbf{V}}^\top, \quad \text{where } \left(\tilde{V}\right)_{im} = \frac{v_{im} - 2p_m}{\sqrt{2\sum_{m=1}^{M} p_m (1 - p_m)}} .$$

Another widely used (e.g. Al Abri et al. 2017; Bérénos et al. 2014) of a genomic relationship matrix, which we denote $\mathbf{G}_{\text{GCTA}}$, was given by Yang et al. (2011). The definition of $\mathbf{G}_{\text{GCTA}}$ differs from $\mathbf{G}_{\text{VR}}$ in that each entry is standardized individually, rather than column-wise. The definition is given as

$$(G_{\text{GCTA}})_{ij} = \frac{1}{M}\sum_{m=1}^{M} \frac{(v_{im} - 2p_m)(v_{jm} - 2p_m)}{2p_m (1 - p_m)} ,$$

so again we can write the matrix as product of centered and standardized entries of $\mathbf{V}$, namely

$$\mathbf{G}_{\text{GCTA}} = \hat{\mathbf{V}}\hat{\mathbf{V}}^\top, \quad \text{where } \left(\hat{V}\right)_{im} = \frac{v_{im} - 2p_m}{\sqrt{2Mp_m (1 - p_m)}} .$$

For both above definitions each entry is centered with the MAF, in a way that alleles with a low MAF are weighted more heavily, as discussed above. Further, the definitions standardize their entries in ways causing their diagonals to have a mean value close to 1 (Legarra 2016). In other words, the scaling similar to $\mathbf{A}$; the diagonal entries are 1 if i is outbred, i.e. not inbred. This allows us to, again, denote the inbreeding coefficient as $F_i = (\mathbf{G}_{\text{VR}})_{ii} - 1$.

In addition to $\mathbf{G}_{\text{VR}}$ and $\mathbf{G}_{\text{GCTA}}$, a large number definitions of genomic relationship matrices exists. For instance, (Speed and Balding 2015) suggests a generalization of $\mathbf{G}_{\text{GCTA}}$ where the denominator in $\left(\hat{V}\right)_{im}$ is replaced by $\sqrt{M}\left(2p_m (1 - p_m)\right)^{-\frac{\alpha}{2}}$. We can then treat α as tuning parameter to define any number of genomic relationship matrices, $\mathbf{G}_\alpha$, where $\mathbf{G}_{-1} = \mathbf{G}_{\text{GCTA}}$. An even more general class of GRM estimators is found in Wang et al. (2017). In another approach, Wientjes et al. (2017) defines $\mathbf{G}$ in such a way that can also be used in estimation of between-population genetic correlations. Edwards (2015) constructs two IBD-based relationship matrices that are based on inferring relatedness

from shared segments of DNA on the haplotype level, i.e. the looking at each copy of a chromosome separately.[2]

All this is to say that there is a plethora of genomic relatedness matrices to choose between. Furthermore, the relationship measures will depend on which SNPs/loci are genotyped, the technology used to perform said genotyping and, in the case of haplotype-level methods, the choice of phasing software. Ergo there is no "right choice" of relatedness matrix, the choice should depend on the data at hand and the genetic architecture of the study population at hand (Speed and Balding 2015). Also note that in general the base population (the population for which we estimate the genetic parameters) in these genome-based methods will differ from the pedigree-scenario, where the base population equals the founders of the pedigree. In the single-SNP IBS methods, such as $\mathbf{G}_{\text{VR}}$ and $\mathbf{G}_{\text{GCTA}}$, the base population will be whatever population the allele frequency is derived from (Hayes et al. 2009; Wientjes et al. 2017). Pedigree-free IBD methods, such those of Edwards (2015), have more nebulous base populations; rather than tracing the genes back to the founder of a pedigree, they must be traced back to the point where they first appeared by mutation (Thompson 2013). The difference in base population the various GRMs results in difficulties in comparing results obtained from animal models using different GRMs, since the estimated genetic variances will not refer to the same populations. However, comparison issues can be partially resolved by rescaling the variances to refer to the same base population, as described by Legarra (2016).

(Comparing pedigrees and genomic approaches? To go in introduction)

When known, Pedigrees cheaper, effective genotyping tech did not exist. Genomic approach requires knowledge of SNPs for a given species? also computationally using pedigrees, as will become clear below. especially promising in wild sys where pedigrees can be unreliable, though genomic approach can also be used to improve pedigrees (eg. ådnes cite)

G is non-sparse in genomic case: even unrelated will share a small amount of genes, non zero elements when talking about non-sparsity of G, cite loh et al 2015

Advantage of genomic: Don't make assumption about base population

Improvement from using g rather than A (Al Abri et al. 2017; Bérénos et al. 2014)

Sentence about how this expands research opportunities (Kardos et al. 2016)

In results/discussion: Legarra (2016) talks about the reason for different results in ped/IBS - different base pops? different sources of noise (see notes)

pedigree can be better when few markers are available (Nietlisbach et al. 2017)

1.2.2 Complicating environmental effects

A major use of the animal model is in the estimation of σ^2_{G}, the additive genetic variance in a population (Kruuk 2004; Wilson et al. 2010). In order to correctly estimate σ^2_{G}, other (possibly confounding) sources of covariance must be accounted for. One should therefore include such sources as additional fixed or random effects in the animal model. These

[2]Haplotype-level methods require the extra step of "phasing" the genotype data, more on this in section **??**.

sources of covariance can include simple correlating elements such as time of measurement and individual traits such as sex, but also environmental effects that can falsely be interpreted by the model as additive genetic effects.

As a first example, let us look at the "common environmental effects" (Kruuk and Hadfield 2007). These effects are problematic if individuals residing in the same environment are more likely to have similar genotypes. For instance siblings, who tend to be quite genetically similar, are usually born in and reside in the same environment. Thus, the similarities in phenotype we see in such relatives might actually partially be a product of living in similar environments, rather than due to genetic factors. An animal model that does not account for individuals living in the same environments might therefore overestimate the additive genetic variance present. When repeated measurements are present, one must also consider "permanent environmental effects," namely effects unique to an individual's personal environment. Should repeated measurements be present in the data, it is recommended (Ponzi et al. 2018) to include an ID random effect, as mentioned in Section 1.1. The inclusion of this effect will capture the correlation between measurements from the same individual. The ID effect will also contain the non-additive genetic effects not captured by the breeding values.

Failure to include confounding environmental effects such as the "common environmental effects" might lead to upward bias in additive genetic variance estimates, and it violates the independence of residuals assumption of a GLMM. Their inclusion also facilitates the study of the environmental effects, which might be of interest in and of themselves (Wilson et al. 2010). Similarly, a failure to include individual traits (like sex) as fixed effects might lead to an inflated estimate of the residual variance σ^2_ε.

With the inclusion of such extra effects, the basic animal model with K fixed effects and L random effects in addition to the genetic value g_i and residual ε_{ij} might be stated as follows. Let y_{ij} be the phenotypic measurement j for individual i, and $x^{(k)}_{ij}$ the corresponding measurement of fixed effect $k \in \{1, \dots, K\}$. Usually all random effects are assumed to be independently normally distributed with zero mean, so let $z^{(l)}_{ij} \sim \mathcal{N}(0, \sigma^2_l)$ for $l \in \{1, \dots, L\}$ be the additional random effects. For each $z^{(l)}_{ij}$ we define some covariance structure, and the value of $z^{(l)}_{ij}$ can depend only on i or on both i and j. Then we can write

$$y_{ij} = \mu + \sum_{k=1}^{K} x^{(k)}_{ij}\beta_k + \sum_{l=1}^{L} z^{(l)}_{ij} + g_i + \varepsilon_{ij} \,. \tag{1.2}$$

The matrix form of this model is simply equation (1.1), with $\boldsymbol{f}(\boldsymbol{y}) = \boldsymbol{y}$ and with the random effect vector including the breeding values. Since all random effects are normally distributed with zero mean we get

$$\mathrm{E}(y_{ij}|\boldsymbol{x}_{ij}) = \mu + \sum_{k=1}^{K} x^{(k)}_{ij}\beta_k \quad \text{and} \quad \mathrm{Var}(y_{ij}|\boldsymbol{x}_{ij}) = \sum_{l=1}^{L} \sigma^2_l + \sigma^2_{\mathrm{G}} + \sigma^2_\varepsilon \,.$$

Note that whenever we include a fixed effect, it changes the interpretation of our results for the additive genetic variance. Such a model would give the σ^2_{G} conditioned on the value of the fixed effect. If we, for example, include sex as a categorical fixed effect, we would

estimate the sex-specific σ^2_{G}, that is, the additive genetic variance of an animal *given its sex*.

1.2.3 Genetic groups extension of the animal model

As mentioned, the animal model estimates genetic parameter such as baseline mean genetic values and additive genetic variance in the base population. Thus, the animal model makes an implicit assumption that these genetic parameters are uniform across the base population; it does not allow for subpopulations within the base population to differ genetically. What if this assumption does not hold? Consider the example of a population that has significant immigration from a distant population over the study period (Wolak and Reid 2017). In the pedigree-based GRM these immigrants would be part of the base population, since any measured immigrant will necessarily have unknown parents, whereas in the genomic-based GRM they would be part of the base population if they are used to calculate allele frequencies. If the distant population has systematically different genotypes, then the assumption that the base population lacks any genetic structure would be violated. The violation of this assumption could lead to the estimated mean breeding values and additive genetic variances being biased towards their values among immigrants, rather than in the original study population.

These issues lead us to consider the possibility of partitioning the base population into "genetic groups" (Quaas 1988; Wolak and Reid 2017; Quaas and Pollak 1981). Rather than assuming that the population has breeding values $\boldsymbol{g} \sim \mathcal{N}(\mathbf{0}, \sigma^2_{\mathrm{G}}\mathbf{G})$, each genetic group is allowed a different mean genetic value and possibly a different additive genetic variance (Muff et al. 2019; Rio et al. 2020a). For example, individuals in genetic group r will have mean genetic value γ_r and, if we allow heterogeneous group variances, genetic variance $\sigma^2_{\mathrm{G}_r}$. We will also refer to γ as the "genetic group effect" of group r. The mechanism of this partition differs when working with pedigrees or with genomic data; but in both cases we differentiate between "reference" individuals and "admixed" individuals. Reference individuals are individuals known to belong to a single genetic group, while admixed individuals are allowed partial membership in more than one group. The immigrant problem above could be solved by assigning the known founders of the study population to a "native" genetic group and known immigrants to an "immigrant" genetic groups, thereby incorporating the genetic structure in the base population into the model (as was done by Wolak and Reid 2016; Charmantier et al. 2016).

Extending the animal model to include genetic groups not only prevents the aforementioned bias, but also allows us to study new and interesting parameters. In the immigrant example, one could study the differences between the two populations, while in general one could investigate the existence of genetic structure within the base population. For example, one could investigate whether different subsets of the base population have different genetic parameters.

Pedigree-based genetic groups

If we have a pedigree available, it can be used to derive group membership proportions (Wolak and Reid 2017; Schaeffer 1991). Let all phantom parents be included in the reference populations; they are entirely members of a single genetic group. Then, all other

individuals have partial membership in various groups depending on their ancestry. Define q_{ir} to be individual i's membership proportion in genetic group r. If i is a phantom parent, then q_{ir} is 1 for the single group i belongs to. On the other hand, if i is not a phantom parent, we let q_{ir} be equal to the mean of each of i's (possibly phantom) parents' membership proportions in the same group. Thus, group membership is inherited through the generations. This inheritance of group memberships will be true on expectation, considering an individual inherits half of their genetic material from each parent. So, just like $(A)_{ij}$ represents an *expected* probability , q_{ir} represents expected group membership proportion.

At first, only let the genetic groups differ in their mean genetic value. We introduce u_i, an individual's "total additive genetic effects," an effect which can be defined as

$$u_i = \sum_{r=1}^{R} q_{ir}\gamma_r + g_i \ ,$$

where R is the number of genetic groups, while γ_r, q_{ir} and g_i are as defined earlier. Let $\mathbf{Q}$ be an $N \times R$ matrix with entries q_{ir} and let $\boldsymbol{\gamma}$ be a vector of length R containing the genetic group effects. The vector of total additive effects $\mathbf{u}$ then has distribution $\mathcal{N}(\mathbf{Q}\boldsymbol{\gamma}, \sigma^2_{\mathrm{A}}\mathbf{A})$.

This definition causes the mean of the total genetic value u_i to be a weighted average of the means of the different genetic groups, where the weights are i's group membership proportions. One way to implement genetic group effects into the animal model is by estimating γ_r explicitly as a fixed effect for each group r. For identifiability reasons we then add the constraint that one of the groups, say r', has mean total additive genetic effect equal to zero, or we will have an infinite number of solutions. This group will then serve as a reference with $\gamma_{r'} = 0$. The effects γ_r for the other groups will denote deviation in mean total additive genetic effect from the reference group.

We can also have the genetic groups to differ further by allowing heterogeneous additive genetic variance, by separating the genetic value vector $\boldsymbol{g}$ into a sum of "partial genetic values" (Muff et al. 2019). Let $\boldsymbol{g} = \sum_{r=1}^{R} \boldsymbol{g}^{(r)}$, so that

$$u_i = \sum_{r=1}^{R} \left[q_{ir}\gamma_r + g_i^{(r)} \right] \ , \tag{1.3}$$

where $g_i^{(r)}$ is the partial breeding value from group r for individual i. Each partial breeding value corresponds to the contribution from a genetic group r, and has its own $N \times N$ group-specific relatedness matrix $\mathbf{A}_r$ resulting in a group-specific genetic additive genetic variance $\sigma^2_{\mathrm{A}_r}$. One practical interpretation of this partition is that $\boldsymbol{g}^{(r)}$ represents the genetic merit of genes inherited from the base population of group r. Thus, summing these values will once again give the genetic value. We will assume the partial breeding values to be independent because they originate from different base populations. Therefore, we can fit each partial breeding value as a random effect in the animal model. When introducing this decomposition of the random component of $\boldsymbol{u}$, we can write $\boldsymbol{u} \sim \mathcal{N}(\mathbf{Q}\boldsymbol{\gamma}, \sum_{r=1}^{R} \sigma^2_{\mathrm{A}_r}\mathbf{A}_r)$.

As for finding $\mathbf{A}_r$, consider the generalized Cholesky decomposition

$$\mathbf{A} = \mathbf{T}\mathbf{D}\mathbf{T}^\top \ , \tag{1.4}$$

where $\mathbf{T}$ will be an $N \times N$ lower triangular matrix with 1s on the diagonal (Mrode 2014, pp. 23–25). $\mathbf{T}$ encodes for the gene flow between generations. Its ijth entry indicates the

proportion of j's genes that i possesses, so that lower triangular entries are given by

$$t_{ii} = 1 \quad \text{and} \quad t_{ij} = \frac{1}{2} \sum_{p \in \mathcal{P}_i} t_{pj} \ , \ j < i \ ,$$

where $\mathcal{P}_i$ is the set containing each known parent of i. The diagonal entries t_{ii} are trivially 1, since you possess all of your own genes. The non-diagonal entries t_{ij} can be interpreted as follows: The proportion of j's genes that i is expected to inherit equals the mean of the respective proportions of genes that i's parents inherited from j. Computing this mean is straightforward when both of i's parents are known. If at least one parent is unknown, we label these missing parents as phantom parents, like before. Phantom parents are assumed to be entirely unrelated to every individual but their child, and thus they possess none of j's genes. Hence their contribution to the mean would be 0, so we only sum over *known* parents in the above expression. A group-specific version of $\mathbf{T}$ can be defined in a way that retains these properties within a given group. For group r define $\mathbf{T}_r$ such that column j of $\mathbf{T}$ is multiplied by q_{jr}, i.e. $\mathbf{T}_r$ has entries

$$t^{(r)}_{ii} = q_{jr} \quad \text{and} \quad t^{(r)}_{ij} = t_{ij} q_{jr} \ , \ j < i \ .$$

Then $t^{(r)}_{ij}$ denotes the proportion of j's genes *within* group r that i possesses.

Meanwhile, the $\mathbf{D}$ in equation (1.4) is an $N \times N$ diagonal matrix that scales the variance in genetic values according to the number of unknown parents and how inbred said parents are. The matrix is defined such that

$$d_{ii} = 1 - \frac{1}{4} \sum_{p \in \mathcal{P}_i} (1 + F_p) \ ,$$

where F_p is the coefficient of inbreeding as defined previously. Note that d_{ii} is smaller when more parents are known. Thus, there is more variance in i's breeding value the fewer of i's parents are known, which is intuitive as we then have less relatedness information for i. We can also see from this expression that an individual's breeding value will have less variance if its parents are severely inbred; the inbreeding has caused there to be less diversity in the genes i can inherit. To get a group-specific $\mathbf{D}_r$, modify the definition of d_{ii} so that

$$d^{(r)}_{ii} = 1 - \frac{q_{ir}}{4} \sum_{p \in \mathcal{P}_i} (1 + F_p) \ .$$

This definition of $\mathbf{D}_r$ is an approximation, as an exact expression would also use group-specific inbreeding coefficients $F^{(r)}_p$ in the definition of $d^{(r)}_{ii}$. The approximation makes the model more computationally feasible, without having a critical impact on the results (Muff et al. 2019). With $\mathbf{T}_r$ and $\mathbf{D}_r$ available, we can compute the group-specific genetic relatedness matrices using the expression

$$\mathbf{A}_r = \mathbf{T}_r \mathbf{D}_r \mathbf{T}_r^\top \ .$$

- Explain segregation variance

- Segregation variance negligible because of inf. model. Segregation variance = "variance caused by differences in allele combinations, average allelic effect, and linkage disequilibrium at and about loci underlying the phenotype in the mixing breeds". In wild populations this variance can be assumed as a consequence of the infinitesimal model (Muff et al. 2019). If not neglected there would be even more random effects to estimate, making the model even more computationally cumbersome.

Genome-based genetic group animal model

In the genomic setting, we cannot trace the inheritance of expected partial group membership q_{ir} through the generations. Thus, we need some other way to determine group membership proportions for admixed individuals. Rio et al. (2020a) suggest two genome-based genetic group models that solve this issue using the "local ancestry" of each individual allele. An allele's local ancestry indicates which group that specific allele has descended from. In a wild population this information is not readily available, and we must infer local ancestry from the genotype data. Fortunately, many methods that perform this inference have been developed (Geza et al. 2019; Padhukasahasram 2014). Out of the two aforementioned models, we shall focus on the model which uses an animal-model formulation, labeled "MAGBLUP-RI" by Rio et al. (2020a). However, this model needs some extensions to be used in wild systems, rather than the plant breeding setup it was developed for.

Firstly, in the plant breeding context subject can be assumed to be homozygous on (almost) every locus, i.e. each locus has two copies of the same allele (Chase 1952). Such individuals are typically produced through heavy inbreeding, via many generations of enforced breeding between close relatives (Beck et al. 2000). As animals in wild populations breed freely without human intervention, these populations have a fair amount of heterozygotes (loci with the two different alleles), even in populations where inbreeding occurs (is this true? reference?). Secondly, the controlled breeding setup allows us to easily restrict breeding to only two genetic groups. Rio et al. (2020a) assume only two genetic groups, which simplifies the analysis of the segregation variance. When in a wild system, there is the potential for any number of genetic groups to interact, which justifies the need to extend the model to work in the case of more groups. In Chapter 2, we will present an extension of the MAGBLUP-RI model in which allows for heterozygosity and 3 (or R????) genetic groups.

- explain haplotypes and phased data here?

Full genetic group animal model (update this)

So, through the use of genetic group effects γ_r and partial genetic values $g_i^{(r)}$, we can now treat $\boldsymbol{u}$ as a genetic value vector, where each individual's mean breeding value and additive genetic variance depends on its group membership proportions. Using the notation from equation (1.2), with g_i replaced by the definition of u_i in equation (1.3), we can state the genetic groups animal model with group-specific mean breeding value and additive genetic

variance as

$$y_{ij} = \mu + \sum_{k=1}^{K} x_{ij}^{(k)} \beta_k + \sum_{r=1}^{R} \left(q_{ir} \gamma_r + g_i^{(r)} \right) + \sum_{l=1}^{L} z_{ij}^{(l)} + \varepsilon_{ij} \ ,$$

where the partial breeding value vectors $\boldsymbol{g^{(r)}}$ are distributed as $\mathcal{N}(0, \sigma^2_{\mathrm{G}_r} \mathbf{G}_r)$.

1.3 Bayesian inference

In this analysis we will adopt a Bayesian framework for statistical inference (Givens and Hoeting 2012, pp. 11–13). This approach involves considering all model parameters to be stochastic variables, rather than having some fixed unknown value. For the animal model this assumption would mean that all fixed effects (including genetic group effects g_r) and the variances of all random effects are stochastic variables.

As part of the Bayesian approach, the model parameter vector $\boldsymbol{\psi}$ is given some prior distribution $f(\boldsymbol{\psi})$, indicating a priori knowledge or belief about the parameters. Let $\boldsymbol{x}$ be a data vector containing all observations, and $\mathcal{L}(\boldsymbol{\psi}|\boldsymbol{x})$ be the likelihood function for the model, indicating how well values of $\boldsymbol{\psi}$ fit the data. Using Bayes' theorem, we can then update our prior distribution to incorporate the information we have learned from the data. Thus, the updated distribution $f(\boldsymbol{\psi}|\boldsymbol{x})$ for $\boldsymbol{\psi}$ given $\boldsymbol{x}$, called the posterior distribution, is found to be

$$f(\boldsymbol{\psi}|\boldsymbol{x}) = c\mathcal{L}(\boldsymbol{\psi}|\boldsymbol{x}) f(\boldsymbol{\psi}) \ ,$$

where c is a normalizing constant, i.e.

$$c^{-1} = \int_{-\infty}^{\infty} \mathcal{L}(\boldsymbol{\psi}|\boldsymbol{x}) f(\boldsymbol{\psi}) \, \mathrm{d}\boldsymbol{\psi} \ ,$$

making $f(\boldsymbol{\psi}|\boldsymbol{x})$ a proper distribution. Having a full posterior distribution for a parameter, rather than a point estimate, gives us more information to work with. Uncertainty estimates are already included in the shape and wideness of the posterior. If we are interested in point estimates we can, for example, consider the posterior mode or posterior mean. As an alternative to the confidence intervals obtained in frequentist statistics, we can simply examine the posterior distribution's quantiles, which in the Bayesian context are called credible intervals (CI). A commonly considered CI is the highest posterior density credible interval (HPD CI), which is the narrowest possible credible interval containing $(1 - \alpha)\%$ of the probability weight.

The major challenge in Bayesian statistics is that finding c is often hard, as the above integral usually does not have a closed form solution. Finding the posterior distributions therefore often involves heavy computations, i.e. in numerical integration of (1.3). In some special cases we can pick so-called conjugate priors, which ensure the posterior distribution is of the same family as the prior, thus giving a closed-form expression for the posterior distribution. However, conjugate priors usually do not exist. It is often preferred to choose uninformative priors, that is, priors that do not hold much information. Uninformative priors let the data "speak for itself," but it can be nontrivial to define these priors to actually be uninformative. To investigate the impact of the choice of prior, one

can estimate $f(\psi|\boldsymbol{x})$ when different priors are chosen to see how the posterior changes, a so-called prior sensitivity analysis.

1.3.1 Bayesian Inference using INLA

Chapter 2

Theory: Extension of a genomic genetic groups model

Extension of the MAGLUP-RI model in Rio et al. (2020a) to include heterozygosity with co-dominance (rather than only homozygous lines) and genetic groups (rather than 2).

- Go through extension, making it readable, leaving many calculations in appendix
- Use notation making it easily comparable to pedigree version
- Be explicit about assumptions:
  - 50-50 co-dominance for all heterozygous loci
  - Treating between two haplotypes on same locus the same as if they were on different locus, except that they share the same allele frequency
  - No explicit modelling of LD, but assuming all allele have an effect (direct as a QTL or through LD with a QTL)

2.1 Definitions

We now let the genomic total genetic value U_i of individual i be a sum of contributions from each allele. The allele contributions β_{mr}^{ref} and β_{mr}^{alt} of the reference and alternate allele, respectively, located on locus $m \in \{1, \ldots, M\}$ will depend on which genetic group $r \in \{1, \ldots, R\}$ the allele is descended from. In other words, we model allele effects to be group-specific, and we further model these contributions to be deterministic.

- Existence of such group-specific allele effect is shown in Technow et al. (2012) (LD with QTL differs between groups) and Rio et al. (2020b)

To indicate an allele's local ancestry (which genetic group it is descended from), we will use the random indicator variable $\Lambda_{imr}^{(h)}$, where $h \in \{1, 2\}$ indicates which of the two

copies of the chromosome strand the allele is located on. Practically, we will not difference between $h = 1$ or $h = 2$, by assuming that the two strands in a diploid organism are interchangeable in their allele contributions. The possible outcomes of $\Lambda_{imr}^{(h)}$ are

$$\Lambda_{imr}^{(h)} = \begin{cases} 1, & \text{allele is descended from group } r, \\ 0, & \text{otherwise.} \end{cases}$$

Give $\mathbf{\Lambda}_{im}^{(h)} = \left[\Lambda_{im1}^{(h)}, \Lambda_{im2}^{(h)}, \ldots, \Lambda_{imR}^{(h)}\right]$ a categorical distribution, i.e. a multinomial distribution with only one trial. Thus, exactly one of the entries of $\mathbf{\Lambda}_{im}^{(h)}$ equals 1 and the other entries equal 0. Define $\pi_{ir} = \Pr\left(\Lambda_{imr}^{(h)} = 1\right)$, where π_{ir} can be interpreted as i's true group membership proportion in group r, so that $\sum_{r=1}^{R} \pi_{ir} = 1$.

- From the properties of a multinomial distribution $\mathrm{E}\left(\Lambda_{imr}^{(h)}\right) = \pi_{ir}$ and $\mathrm{Var}\left(\Lambda_{imr}^{(h)}\right) = \pi_{ir}\left(1 - \pi_{ir}\right)$.
- $W_{imr}^{(h)}$: Number of copies of genotype alt on the hth allele of locus m of individual i, given that said allele was descended from group r. We let $W_{imr}^{(h)} \sim$ Bernoulli $(p = f_{mr})$, where f_{mr} is the frequency of genotype alt on alleles at locus m in group r. Thus $\mathrm{E}(W_{imr}) = f_{mr}$ and $\mathrm{Var}(W_{imr}) = f_{mr}(1 - f_{mr})$.
- $\Gamma_{ij}^{(r)} = \mathrm{Corr}\left(W_{imr}, W_{jmr}\right)$: The conditional kinship between individuals i and j on their shared group p ancestries
- $\theta_{ij}^{(r)} = \mathrm{E}\left(\Lambda_{imr}^{(h)}\Lambda_{jmr}^{(h')}\right)$: proportion of shared group ancestry. Value of h irrelevant since Λ_{imr}^{1} and Λ_{imr}^{2} have the same distribution
- $\mathrm{Cov}(\Lambda_{imr}^{(h)}, \Lambda_{jmr}^{(h')}) = \Delta_{ij}^{(r)} = \theta_{ij}^{(r)} - \pi_i^{(r)}\pi_j^{(r)} \forall m$: Within-group allele ancestry covariance. Again h does not matter
- $\gamma_r = \sum_{m=1}^{M} \gamma_{mr} = \sum_{m=1}^{M} \left[\beta_{mr}^{\text{ref}} + f_{mr}\left(\beta_{mr}^{\text{alt}} - \beta_{mr}^{\text{ref}}\right)\right]$: expected allele effect across group r.

2.2 Model for genetic value

Assume that:

- $\mathrm{Corr}\left(W_{imr}^{(h)}, W_{jm'r'}^{(h')}\right) = 0, \quad \forall m, m' \neq m, \forall r, r', \forall i, j, \forall h, h'$. That is, no within-group LD or between-group LD, within or between individuals.
- $\Lambda_{imr}^{(h)} \perp W_{jm'r'}^{(h')}, \quad \forall i, m, r, j, m', r', h, h'$.
- $\mathrm{Cov}\left(\Lambda_{imr}^{(h)}, \Lambda_{jm'r'}^{(h')}\right) = 0$ for $i \neq j$, $m \neq m'$, $r \neq r'$ and $\forall h, h'$.

Definition of genetic value

$$U_i = \sum_{r=1}^{R}\sum_{m=1}^{M}\frac{1}{2}\sum_{h=1}^{2}\Lambda_{imr}^{(h)}\left[\beta_{mr}^{\text{ref}} + W_{imr}^{(h)}\left(\beta_{mr}^{\text{alt}} - \beta_{mr}^{\text{ref}}\right)\right] \tag{2.1}$$

2.3 Mean genetic value

Then

$$\mathrm{E}\left(U_i\right) = \sum_{r=1}^{R}\sum_{m=1}^{M} \mathrm{E}\left\{\Lambda_{imr}^{(h)}\left[\beta_{mr}^{\text{ref}} + W_{imr}^{(h)}\left(\beta_{mr}^{\text{alt}} - \beta_{mr}^{\text{ref}}\right)\right]\right\} \tag{2.2}$$

$$= \sum_{r=1}^{R}\sum_{m=1}^{M}\left[\pi_{ir}\beta_{mr}^{\text{ref}} + \mathrm{E}\left(W_{imr}^{(h)}\Lambda_{imr}^{(h)}\right)\left(\beta_{mr}^{\text{alt}} - \beta_{mr}^{\text{ref}}\right)\right] \tag{2.3}$$

$$= \sum_{r=1}^{R}\sum_{m=1}^{M}\left[\pi_{ir}\beta_{mr}^{\text{ref}} + \mathrm{E}\left(W_{imr}^{(h)}\right)\mathrm{E}\left(\Lambda_{imr}^{(h)}\right)\left(\beta_{mr}^{\text{alt}} - \beta_{mr}^{\text{ref}}\right)\right] \tag{2.4}$$

$$= \sum_{r=1}^{R}\sum_{m=1}^{M}\left[\pi_{ir}\left(\beta_{mr}^{\text{ref}} + f_{mr}\left(\beta_{mr}^{\text{alt}} - \beta_{mr}^{\text{ref}}\right)\right)\right] \tag{2.5}$$

$$= \sum_{r=1}^{R}\sum_{m=1}^{M}\pi_{ir}\gamma_{mr} = \sum_{r=1}^{R}\pi_{ir}\gamma_r\ . \tag{2.6}$$

Thus, the mean genetic value of individual i is a weighted sum of the group means, where the weights are the realized group membership proportions.

2.4 Equivalent model for genetic value

Define mean centered versions of the random variables:

$$\bar{\Lambda}_{imr}^{(h)} = \Lambda_{imr}^{(h)} - \pi_{ir} \quad \text{and} \quad \bar{W}_{imr}^{(h)} = \Lambda_{imr}^{(h)}\left(W_{imr}^{(h)} - f_{mr}\right)\ , \tag{2.7}$$

so that $\mathrm{E}\left(\bar{\Lambda}_{imr}^{(h)}\right) = 0$ and $\mathrm{E}\left(\bar{W}_{imr}^{(h)}\right) = 0$. Then

$$U_i = \sum_{r=1}^{R}\pi_{ir}\gamma_r + \sum_{m=1}^{M}\left[\sum_{r=1}^{R-1}\frac{\bar{\Lambda}_{imr}^{(1)} + \bar{\Lambda}_{imr}^{(2)}}{2}\left(\gamma_{mr} - \gamma_{mR}\right) + \sum_{r=1}^{R}\frac{\bar{W}_{imr}^{(1)} + \bar{W}_{imr}^{(2)}}{2}\left(\beta_{mr}^{\text{alt}} - \beta_{mr}^{\text{ref}}\right)\right]$$

is an equivalent form of U_i (see secion 2.6.1). Note that for two alleles on the same locus we have within-group centered local ancestry covariance

$$\mathrm{Cov}(\bar{\Lambda}_{imr}^{(h)}\ ,\ \bar{\Lambda}_{jmr}^{(h')}) = \Delta_{ij}^{(r)}\ ,$$

(regardless of the values of h and h', see section 2.6.2) and within-group centered genotype covariance

$$\mathrm{Cov}\left(\bar{W}_{imr}^{(h)}, \bar{W}_{jmr}^{(h')}\right) = \theta_{ij}^{(r)}\Gamma_{ij}^{(r)}\, f_{mr}\left(1 - f_{mr}\right)$$

(see section 2.6.3). Furthermore, we have

$$\mathrm{Cov}\left(\bar{\Lambda}_{imr}^{(h)}, \bar{\Lambda}_{jm'r}^{(h')}\right) = \cdots = \frac{\Delta_{ij}^{(r)}}{M-1}, \quad m \neq m'\ ,$$

which was shown in Rio et al 2020 S1, and is basically identical for this extension. Finally, for $R > 2$ we need $\mathrm{Cov}\left(\bar{\Lambda}^{(h)}_{imr}, \bar{\Lambda}^{(h)}_{jmr'}\right)$, the between-group covariance of local ancestry on a locus m. The result below (see section 2.6.5) is for $R = 3$ is

$$\mathrm{Cov}(\bar{\Lambda}^{(h)}_{imr}, \bar{\Lambda}^{(h)}_{jmr'}) = \frac{\Delta^{(r'')}_{ij} - \Delta^{(r)}_{ij} - \Delta^{(r')}_{ij}}{2}, \quad r'' \in \{1,2,3\}\backslash\{r,r'\}\,.$$

We also show (in section 2.6.4) that

$$\mathrm{Cov}\left(\bar{W}^{(h)}_{imr}, \bar{W}^{(h')}_{jm'r}\right) = 0, \quad m \neq m'\,,$$

i.e. no between-individual covariance between genotypes on different loci.

2.5 Covariance between genetic values

In summary, we have

$$U_i = \sum_{r=1}^{R} \pi_{ir}\gamma_r + \sum_{m=1}^{M}\left[\sum_{r=1}^{R-1} \frac{\bar{\Lambda}^{(1)}_{imr} + \bar{\Lambda}^{(2)}_{imr}}{2}\left(\gamma_{mr} - \gamma_{mR}\right) + \sum_{r=1}^{R} \frac{\bar{W}^{(1)}_{imr} + \bar{W}^{(2)}_{imr}}{2}\left(\beta^{\text{alt}}_{mr} - \beta^{\text{ref}}_{mr}\right)\right]$$

and the nonzero covariances (all other covariances are 0)

- $\mathrm{Cov}\left(\bar{W}^{(h)}_{imr}, \bar{W}^{(h')}_{jmr}\right) = \theta^{(r)}_{ij}\Gamma^{(r)}_{ij} f_{mr}\left(1 - f_{mr}\right)$
- $\mathrm{Cov}\left(\bar{\Lambda}^{(h)}_{imr}, \bar{\Lambda}^{(h')}_{jmr}\right) = \Delta^{(r)}_{ij}$
- $\mathrm{Cov}\left(\bar{\Lambda}^{(h)}_{imr}, \bar{\Lambda}^{(h')}_{jm'r}\right) = -\frac{\Delta^{(r)}_{ij}}{M-1}, \quad (m \neq m')$
- $\mathrm{Cov}\left(\bar{\Lambda}^{(h)}_{imr}, \bar{\Lambda}^{(h)}_{jmr'}\right) =$
  - $R = 2$: $-\Delta^{(r)}_{ij}$, (though this is irrelevant)
  - $R = 3$: $\frac{\Delta^{r''}_{ij} - \Delta^{r}_{ij} - \Delta^{r'}_{ij}}{2}, \quad r'' \in \{1,2,3\}\backslash\{r,r'\}$
  - $R \geq 4$: ???

We define the group-specific genetic variances, and intergroup segregation variances the same way as Rio et al 2020. Group-specific genetic variances:

$$\sigma^2_{G_r} = \sum_{m=1}^{M} f_{mr}\left(1 - f_{mr}\right)\left(\beta^{\text{alt}}_{mr} - \beta^{\text{ref}}_{mr}\right)^2\,,$$

and intergroup segregation variances between groups r and r':

$$\sigma^2_{S,rr'} = \left[\frac{M}{M-1}\sum_{m=1}^{M}\left(\gamma_{mr} - \gamma_{mr'}\right)^2 - \frac{1}{M-1}\left(\sum_{m=1}^{M}\left(\gamma_{mr} - \gamma_{mr'}\right)\right)^2\right]$$

Note that when M is large and/or the difference in group mean (such as in the infinitesimal model) is small, then

$$\sigma^2_{S,rr'} \approx \sum_{m=1}^{M} \left(\gamma_{mr} - \gamma_{mr'}\right)^2 \ .$$

We will assume this approximation can be made here (wild system), so $\mathrm{Cov}(\bar{\Lambda}^{(h)}_{imr}, \bar{\Lambda}^{(h')}_{jm'r}) = 0$ also. We can then derive the covariance between genetic values of different individuals as follows.

$$\begin{aligned}
&\mathrm{Cov}(U_i, G_j \mid \boldsymbol{\pi}_i, \boldsymbol{\pi}_j, \boldsymbol{\theta}_{ij}, \boldsymbol{\Gamma}_{ij}) \\
&= \mathrm{Cov}\left(\sum_{m=1}^{M}\sum_{r=1}^{R-1} \frac{\bar{\Lambda}^{(1)}_{imr} + \bar{\Lambda}^{(2)}_{imr}}{2}\left(\gamma_{mr} - \gamma_{mR}\right), \sum_{m'=1}^{M}\sum_{r'=1}^{R-1} \frac{\bar{\Lambda}^{(1)}_{jmr} + \bar{\Lambda}^{(2)}_{jmr}}{2}\left(\gamma_{mr} - \gamma_{mR}\right)\right) \\
&+ \mathrm{Cov}\left(\sum_{m=1}^{M}\sum_{r=1}^{R} \frac{\bar{W}^{(1)}_{imr} + \bar{W}^{(2)}_{imr}}{2}\left(\beta^{\text{alt}}_{mr} - \beta^{\text{ref}}_{mr}\right), \sum_{m'=1}^{M}\sum_{r'=1}^{R} \frac{\bar{W}^{(1)}_{jmr} + \bar{W}^{(2)}_{jmr}}{2}\left(\beta^{\text{alt}}_{mr} - \beta^{\text{ref}}_{mr}\right)\right) \\
&= \sum_{m=1}^{M}\sum_{r=1}^{R-1} \frac{\mathrm{Cov}\left(\bar{\Lambda}^{(1)}_{imr} + \bar{\Lambda}^{(2)}_{imr}, \bar{\Lambda}^{(1)}_{jmr} + \bar{\Lambda}^{(2)}_{jmr}\right)}{4}\left(\gamma_{mr} - \gamma_{mR}\right)^2 \\
&+ \sum_{m=1}^{M}\sum_{r=1}^{R-1}\sum_{r'\neq r}^{R-1} \frac{\mathrm{Cov}\left(\bar{\Lambda}^{(1)}_{imr} + \bar{\Lambda}^{(2)}_{imr}, \bar{\Lambda}^{(1)}_{jmr'} + \bar{\Lambda}^{(2)}_{jmr'}\right)}{4}\left(\gamma_{mr} - \gamma_{mR}\right)\left(\gamma_{mr'} - \gamma_{mR}\right) \\
&+ \sum_{m=1}^{M}\sum_{r=1}^{R} \frac{\mathrm{Cov}\left(\bar{W}^{(1)}_{imr} + \bar{W}^{(2)}_{imr}, \bar{W}^{(1)}_{jmr} + \bar{W}^{(2)}_{jmr}\right)}{4}\left(\beta^{\text{alt}}_{mr} - \beta^{\text{ref}}_{mr}\right)^2
\end{aligned}$$

Now consider the case $R = 3$.

$$
\begin{aligned}
&\mathrm{Cov}(U_i, G_j \mid \boldsymbol{\pi}_i, \boldsymbol{\pi}_j, \boldsymbol{\theta}_{ij}, \boldsymbol{\Gamma}_{ij}) \\
&= \sum_{m=1}^{M} \sum_{r=1}^{2} \frac{4\Delta_{ij}^{(r)}}{4} \left(\gamma_{mr} - \gamma_{mR}\right)^2 \\
&+ \sum_{m=1}^{M} \frac{4\frac{\Delta_{ij}^{(3)} - \Delta_{ij}^{(1)} - \Delta_{ij}^{(2)}}{2}}{4} \left(\gamma_{m1} - \gamma_{m3}\right)\left(\gamma_{m2} - \gamma_{m3}\right) \cdot 2 \\
&+ \sum_{r=1}^{3} \sum_{m=1}^{M} \frac{4\theta_{ij}^{(r)} \Gamma_{ij}^{(r)} f_{mr}(1 - f_{mr})}{4} \left(\beta_{mr}^{\text{alt}} - \beta_{mr}^{\text{ref}}\right)^2 \\
&= \sum_{r=1}^{2} \Delta_{ij}^{(r)} \left[\sum_{m=1}^{M} \left(\gamma_{mr} - \gamma_{mR}\right)^2\right] \\
&+ \left(\Delta_{ij}^{(3)} - \Delta_{ij}^{(1)} - \Delta_{ij}^{(2)}\right) \sum_{m=1}^{M} \left(\gamma_{m1}\gamma_{m2} - \gamma_{m1}\gamma_{m3} - \gamma_{m2}\gamma_{m3} + \gamma_{m3}^2\right) \\
&+ \sum_{r=1}^{3} \theta_{ij}^{(r)} \Gamma_{ij}^{(r)} \left[\sum_{m=1}^{M} f_{mr}(1 - f_{mr}) \left(\beta_{mr}^{\text{alt}} - \beta_{mr}^{\text{ref}}\right)^2\right] \\
&= \Delta_{ij}^{(1)} \sigma_{S,13}^2 + \Delta_{ij}^{(2)} \sigma_{S,23}^2 \\
&- \frac{1}{2}\left(\Delta_{ij}^{(3)} - \Delta_{ij}^{(1)} - \Delta_{ij}^{(2)}\right)\left(\sigma_{S,13}^2 + \sigma_{S,23}^2 - \sigma_{S,12}^2\right) \text{ (see section 2.6.6)} \\
&+ \sum_{r=1}^{3} \theta_{ij}^{(r)} \Gamma_{ij}^{(r)} \sigma_{G_r}^2 \\
&= \sum_{r=1}^{2} \sum_{r'=r+1}^{3} \frac{\Delta_{ij}^{(r)} + \Delta_{ij}^{(r')} - \Delta_{ij}^{(r'')}}{2} \sigma_{S,rr'}^2 + \sum_{r=1}^{3} \theta_{ij}^{(r)} \Gamma_{ij}^{(r)} \sigma_{G_r}^2 \,,
\end{aligned}
$$

where $r'' \in \{1, 2, 3\} \setminus \{r, r'\}$. This can be used as a covariance structure in a GLMM. To estimate $\Gamma_{ij}^{(r)}$ with the group-specific version of the Van Raden matrix (which also accounts for which of two the alleles we are on, which is possible because we have phased the data):

$$
(\mathbf{G}_r)_{ij} = \frac{\sum_{m=1}^{M} \sum_{h=1}^{2} a_{imr}^h \left(w_{im}^h - \tilde{f}_{mr}\right) a_{jmr}^h \left(w_{jm}^h - \tilde{f}_{mr}\right)}{\sum_{m=1}^{M} \sum_{h=1}^{2} a_{imr}^h a_{jmr}^h \tilde{f}_{mr}\left(1 - \tilde{f}_{mr}\right)},
$$

where a_{imr}^h and w_{im}^h are observed local ancestries and genotype, respectively. $\tilde{f}_{mr}$ is the observed within-group allele frequency, estimated by $\tilde{f}_{mr} = \frac{\sum_{i=1}^{N} \sum_{h=1}^{2} a_{imr}^h w_{im}^h}{\sum_{i=1}^{N} \sum_{h=1}^{2} a_{imr}^h}$, where N is the number of individuals. We estimate the other covariance matrices by

$$
(\boldsymbol{\theta}_r)_{ij} = \frac{1}{2M} \sum_{m=1}^{M} \sum_{h=1}^{2} a_{imr}^h a_{jmr}^h, \text{ (interpret as shared group membership)}
$$

and

$$(\boldsymbol{\Delta}_r)_{ij} = (\boldsymbol{\theta}_r)_{ij} - \tilde{\pi}_{ir}\tilde{\pi}_{jr} \ ,$$

where the π_{ir}-estimate $\tilde{\pi}_{ir} = \frac{1}{2M}\sum_{m=1}^{M}\sum_{h=1}^{2} a_{imr}^{h}$. Thus we can fit the animal model

$$\boldsymbol{y} = \mathbf{X}\boldsymbol{\gamma} + \sum_{r=1}^{2}\sum_{r'=r+1}^{3} \boldsymbol{g}_{S,rr'} + \sum_{r=1}^{3} \boldsymbol{g}_r + \boldsymbol{\varepsilon} \ ,$$

where $\boldsymbol{g}_{S,rr'} \sim \mathcal{N}(0, \frac{\boldsymbol{\Delta}_r + \boldsymbol{\Delta}_{r'} - \boldsymbol{\Delta}_{r''}}{2}\sigma^2_{S,rr'})$, $\boldsymbol{g}_r \sim \mathcal{N}(0, (\boldsymbol{\theta}_r \circ \mathbf{G}_r)\,\sigma^2_{G_r})$ and $\boldsymbol{\varepsilon} \sim \mathcal{N}(0, \mathbf{I}\sigma^2_{\varepsilon})$, where "$\circ$" is the Hadamard product, i.e. element-wise multiplication.

2.6 Miscellaneous calculations

2.6.1 Derivation of equivalent model

$$
\begin{aligned}
U_i &= \sum_{m=1}^{M}\sum_{r=1}^{R}\frac{1}{2}\sum_{h=1}^{2}\left[\Lambda_{imr}^{(h)}\left(\beta_{mr}^{\text{ref}}+W_{imr}^{(h)}\left(\beta_{mr}^{\text{alt}}-\beta_{mr}^{\text{ref}}\right)\right)\right] \\
&= \sum_{m=1}^{M}\sum_{r=1}^{R}\frac{1}{2}\sum_{h=1}^{2}\left[\Lambda_{imr}^{(h)}\left(\beta_{mr}^{\text{ref}}+\left(\frac{\bar{W}_{imr}^{(h)}}{\Lambda_{imr}^{(h)}}+f_{mr}\right)\left(\beta_{mr}^{\text{alt}}-\beta_{mr}^{\text{ref}}\right)\right)\right] \\
&= \sum_{m=1}^{M}\sum_{r=1}^{R}\frac{1}{2}\sum_{h=1}^{2}\left[\Lambda_{imr}^{(h)}\beta_{mr}^{\text{ref}}+\left(\bar{W}_{imr}^{(h)}+\Lambda_{imr}^{(h)}f_{mr}\right)\left(\beta_{mr}^{\text{alt}}-\beta_{mr}^{\text{ref}}\right)\right] \\
&= \sum_{m=1}^{M}\sum_{r=1}^{R}\frac{1}{2}\sum_{h=1}^{2}\left[\Lambda_{imr}^{(h)}\left(\beta_{mr}^{\text{ref}}+f_{mr}\left(\beta_{mr}^{\text{alt}}-\beta_{mr}^{\text{ref}}\right)\right)+\bar{W}_{imr}^{(h)}\left(\beta_{mr}^{\text{alt}}-\beta_{mr}^{\text{ref}}\right)\right] \\
&= \sum_{m=1}^{M}\sum_{r=1}^{R}\frac{1}{2}\sum_{h=1}^{2}\left[\Lambda_{imr}^{(h)}\gamma_{mr}+\bar{W}_{imr}^{(h)}\left(\beta_{mr}^{\text{alt}}-\beta_{mr}^{\text{ref}}\right)\right] \\
&= \sum_{m=1}^{M}\frac{1}{2}\sum_{h=1}^{2}\left[\sum_{r=1}^{R-1}\left(\bar{\Lambda}_{imr}^{(h)}+\pi_{ir}\right)\gamma_{mr}+\Lambda_{imR}^{(h)}\gamma_{mR}+\sum_{r=1}^{R}\bar{W}_{imr}^{(h)}\left(\beta_{mr}^{\text{alt}}-\beta_{mr}^{\text{ref}}\right)\right] \\
&= \sum_{m=1}^{M}\frac{1}{2}\sum_{h=1}^{2}\left[\sum_{r=1}^{R-1}\left(\bar{\Lambda}_{imr}^{(h)}+\pi_{ir}\right)\gamma_{mr}+\left(1-\sum_{r=1}^{R-1}\Lambda_{imr}^{(h)}\right)\gamma_{mR}+\sum_{r=1}^{R}\bar{W}_{imr}^{(h)}\left(\beta_{mr}^{\text{alt}}-\beta_{mr}^{\text{ref}}\right)\right] \\
&= \sum_{m=1}^{M}\frac{1}{2}\sum_{h=1}^{2}\left[\sum_{r=1}^{R-1}\left(\bar{\Lambda}_{imr}^{(h)}+\pi_{ir}\right)\gamma_{mr}+\left(1-\sum_{r=1}^{R-1}\left(\bar{\Lambda}_{imr}^{(h)}+\pi_{ir}\right)\right)\gamma_{mR}+\sum_{r=1}^{R}\bar{W}_{imr}^{(h)}\left(\beta_{mr}^{\text{alt}}-\beta_{mr}^{\text{ref}}\right)\right] \\
&= \sum_{m=1}^{M}\frac{1}{2}\sum_{h=1}^{2}\left[\sum_{r=1}^{R-1}\left(\bar{\Lambda}_{imr}^{(h)}+\pi_{ir}\right)\gamma_{mr}+\left(\pi_{iR}-\sum_{r=1}^{R-1}\bar{\Lambda}_{imr}^{(h)}\right)\gamma_{mR}+\sum_{r=1}^{R}\bar{W}_{imr}^{(h)}\left(\beta_{mr}^{\text{alt}}-\beta_{mr}^{\text{ref}}\right)\right] \\
&= \sum_{m=1}^{M}\frac{1}{2}\sum_{h=1}^{2}\left[\sum_{r=1}^{R-1}\bar{\Lambda}_{imr}^{(h)}\left(\gamma_{mr}-\gamma_{mR}\right)+\sum_{r=1}^{R}\pi_{ir}\gamma_{mr}+\sum_{r=1}^{R}\bar{W}_{imr}^{(h)}\left(\beta_{mr}^{\text{alt}}-\beta_{mr}^{\text{ref}}\right)\right] \\
&= \sum_{m=1}^{M}\left[\sum_{r=1}^{R-1}\frac{\bar{\Lambda}_{imr}^{(1)}+\bar{\Lambda}_{imr}^{(2)}}{2}\left(\gamma_{mr}-\gamma_{mR}\right)+\sum_{r=1}^{R}\pi_{ir}\gamma_{mr}+\sum_{r=1}^{R}\frac{\bar{W}_{imr}^{(1)}+\bar{W}_{imr}^{(2)}}{2}\left(\beta_{mr}^{\text{alt}}-\beta_{mr}^{\text{ref}}\right)\right] \\
&= \sum_{r=1}^{R}\pi_{ir}\gamma_{r}+\sum_{m=1}^{M}\left[\sum_{r=1}^{R-1}\frac{\bar{\Lambda}_{imr}^{(1)}+\bar{\Lambda}_{imr}^{(2)}}{2}\left(\gamma_{mr}-\gamma_{mR}\right)+\sum_{r=1}^{R}\frac{\bar{W}_{imr}^{(1)}+\bar{W}_{imr}^{(2)}}{2}\left(\beta_{mr}^{\text{alt}}-\beta_{mr}^{\text{ref}}\right)\right]
\end{aligned}
$$

2.6.2 Same-locus within-group centered local ancestry covariance

$$
\text{Cov}(\bar{\Lambda}_{imr}^{(h)},\,\bar{\Lambda}_{jmr}^{(h')})=\text{Cov}(\bar{\Lambda}_{imr}^{(h)}-\pi_{ir},\,\bar{\Lambda}_{jmr}^{(h')}-\pi_{jr})=\text{Cov}(\Lambda_{imr}^{(h)},\,\Lambda_{jmr}^{(h')})=\Delta_{ij}^{(r)},
$$

2.6.3 Same-locus within-group centered genotype covariance

$$
\begin{aligned}
\mathrm{Cov}\left(\bar{W}_{imr}^{(h)}, \bar{W}_{jmr}^{(h')}\right) &= \mathrm{Cov}\left(\Lambda_{imr}^{(h)}\left(W_{imr}^{(h)} - f_{mr}\right)\ ,\ \Lambda_{jmr}^{(h')}\left(W_{jmr}^{(h')} - f_{mr}\right)\right) \\
&= \mathrm{Cov}\left(\Lambda_{imr}^{(h)} W_{imr}^{(h)}, \Lambda_{jmr}^{(h')} W_{jmr}^{(h')}\right) \\
&- f_{mr}\mathrm{Cov}\left(\Lambda_{imr}^{(h)} W_{imr}^{(h)}\ ,\ \Lambda_{jmr}^{(h')}\right) - f_{mr}\mathrm{Cov}\left(\Lambda_{imr}^{(h)}, \Lambda_{jmr}^{(h')} W_{jmr}^{(h')}\right) \\
&+ f_{mr}^2\mathrm{Cov}\left(\Lambda_{imr}^{(h)}, \Lambda_{jmr}^{(h')}\right) \\
&= \mathrm{E}\left(\Lambda_{imr}^{(h)} W_{imr}^{(h)} \Lambda_{jmr}^{(h')} W_{jmr}^{(h')}\right) - \mathrm{E}\left(\Lambda_{imr}^{(h)} W_{imr}^{(h)}\right)\mathrm{E}\left(\Lambda_{jmr}^{(h')} W_{jmr}^{(h')}\right) \\
&- f_{mr}[\mathrm{E}\left(\Lambda_{imr}^{(h)} W_{imr}^{(h)} \Lambda_{jmr}^{(h')}\right) - \mathrm{E}\left(\Lambda_{imr}^{(h)} W_{imr}^{(h)}\right)\mathrm{E}\left(\Lambda_{jmr}^{(h')}\right) \\
&\quad + \mathrm{E}\left(\Lambda_{imr}^{(h)} \Lambda_{jmr}^{(h')} W_{jmr}^{(h')}\right) - \mathrm{E}\left(\Lambda_{imr}^{(h)}\right)\mathrm{E}\left(\Lambda_{jmr}^{(h')} W_{jmr}^{(h')}\right)] \\
&+ f_{mr}^2\left(\theta_{ij}^{(r)} - \pi_{ir}\pi_{jr}\right) \\
&= \mathrm{E}\left(\Lambda_{imr}^{(h)} \Lambda_{jmr}^{(h')}\right)\mathrm{E}\left(W_{imr}^{(h)} W_{jmr}^{(h')}\right) - \mathrm{E}\left(\Lambda_{imr}^{(h)}\right)\mathrm{E}\left(W_{imr}^{(h)}\right)\mathrm{E}\left(\Lambda_{jmr}^{(h')}\right)\mathrm{E}\left(W_{jmr}^{(h')}\right) \\
&- f_{mr}[\mathrm{E}\left(\Lambda_{imr}^{(h)} \Lambda_{jmr}^{(h')}\right)\mathrm{E}\left(W_{imr}^{(h)}\right) - \mathrm{E}\left(\Lambda_{imr}^{(h)}\right)\mathrm{E}\left(\Lambda_{jmr}^{(h')}\right)\mathrm{E}\left(W_{jmr}^{(h')}\right) \\
&\quad + \mathrm{E}\left(\Lambda_{imr}^{(h)} \Lambda_{jmr}^{(h')}\right)\mathrm{E}\left(W_{jmr}^{(h')}\right) - \mathrm{E}\left(\Lambda_{imr}^{(h)}\right)\mathrm{E}\left(\Lambda_{jmr}^{(h')}\right)\mathrm{E}\left(W_{jmr}^{(h')}\right)] \\
&+ f_{mr}^2\left(\theta_{ij}^{(r)} - \pi_{ir}\pi_{jr}\right) \\
&= \theta_{ij}^{(r)}\left(\mathrm{Cov}\left(W_{imr}^{(h)}, W_{jmr}^{(h')}\right) + \mathrm{E}\left(W_{imr}^{(h)}\right)\mathrm{E}\left(W_{jmr}^{(h')}\right)\right) - f_{mr}^2\pi_{ir}\pi_{jr} \\
&- f_{mr}\left(2\theta_{ij}^{(r)} f_{mr} - 2f_{mr}\pi_{ir}\pi_{jr}\right) \\
&+ f_{mr}^2\left(\theta_{ij}^{(r)} - \pi_{ir}\pi_{jr}\right) \\
&= \theta_{ij}^{(r)}\left(\Gamma_{ij}^{(r)} \cdot f_{mr}\left(1 - f_{mr}\right) + f_{mr}^2\right) - f_{mr}^2\pi_{ir}\pi_{jr} \\
&- 2f_{mr}^2\left(\theta_{ij}^{(r)} - \pi_{ir}\pi_{jr}\right) \\
&+ f_{mr}^2\left(\theta_{ij}^{(r)} - \pi_{ir}\pi_{jr}\right) \\
&= \theta_{ij}^{(r)}\Gamma_{ij}^{(r)} f_{mr}\left(1 - f_{mr}\right)\ .
\end{aligned}
$$

2.6.4 Between-individual covariance between genotypes on different loci

$$
\begin{aligned}
\mathrm{Cov}\left(\bar{W}_{imr}^{(h)}, \bar{W}_{jm'r}^{(h')}\right) &= \mathrm{Cov}\left(\Lambda_{imr}^{(h)}\left(W_{imr}^{(h)} - f_{mr}\right)\ ,\ \Lambda_{jm'r}^{(h')}\left(W_{jm'r}^{(h')} - f_{m'r}\right)\right) \\
&= \mathrm{Cov}\left(\Lambda_{imr}^{(h)} W_{imr}^{(h)}\ ,\ \Lambda_{jm'r}^{(h')} W_{jm'r}^{(h')}\right) \\
&- f_{m'r}\mathrm{Cov}\left(\Lambda_{imr}^{(h)} W_{imr}^{(h)}\ ,\ \Lambda_{jm'r}^{(h')}\right) - f_{mr}\mathrm{Cov}\left(\Lambda_{imr}^{(h)}\ ,\ \Lambda_{jm'r}^{(h')} W_{jm'r}^{(h')}\right) \\
&+ f_{mr} f_{m'r}\mathrm{Cov}\left(\Lambda_{imr}^{(h)}, \Lambda_{jm'r}^{(h')}\right) \\
&= \mathrm{E}\left(\Lambda_{imr}^{(h)} W_{imr}^{(h)} \Lambda_{jm'r}^{(h')} W_{jm'r}^{(h')}\right) - \mathrm{E}\left(\Lambda_{imr}^{(h)} W_{imr}^{(h)}\right)\mathrm{E}\left(\Lambda_{jm'r}^{(h')} W_{jm'r}^{(h')}\right) \\
&- f_{m'r}\left[\mathrm{E}\left(\Lambda_{imr}^{(h)} W_{imr}^{(h)} \Lambda_{jmr}^{(h')}\right) - \mathrm{E}\left(\Lambda_{imr}^{(h)} W_{imr}^{(h)}\right)\mathrm{E}\left(\Lambda_{jmr}^{(h')}\right)\right] \\
&- f_{mr}\left[\mathrm{E}\left(\Lambda_{imr}^{(h)} \Lambda_{jm'r}^{(h')} W_{jm'r}^{(h')}\right) - \mathrm{E}\left(\Lambda_{imr}^{(h)}\right)\mathrm{E}\left(\Lambda_{jm'r}^{(h')} W_{jm'r}^{(h')}\right)\right] \\
&+ f_{mr} f_{m'r}\left(-\frac{\theta_{ij}^{(r)}}{M-1}\right) \\
&= \mathrm{E}\left(\Lambda_{imr}^{(h)} \Lambda_{jm'r}^{(h')}\right)\mathrm{E}\left(W_{imr}^{(h)} W_{jm'r}^{(h')}\right) - \mathrm{E}\left(\Lambda_{imr}^{(h)}\right)\mathrm{E}\left(W_{imr}^{(h)}\right)\mathrm{E}\left(\Lambda_{jm'r}^{(h')}\right)\mathrm{E}\left(W_{jm'r}^{(h')}\right. \\
&- f_{m'r}\left[\mathrm{E}\left(\Lambda_{imr}^{(h)} \Lambda_{jm'r}^{(h')}\right)\mathrm{E}\left(W_{imr}^{(h)}\right) - \mathrm{E}\left(\Lambda_{imr}^{(h)}\right)\mathrm{E}\left(\Lambda_{jm'r}^{(h')}\right)\mathrm{E}\left(W_{imr}^{(h')}\right)\right] \\
&- f_{mr}\left[\mathrm{E}\left(\Lambda_{imr}^{(h)} \Lambda_{jm'r}^{(h')}\right)\mathrm{E}\left(W_{jm'r}^{(h')}\right) - \mathrm{E}\left(\Lambda_{imr}^{(h)}\right)\mathrm{E}\left(\Lambda_{jm'r}^{(h')}\right)\mathrm{E}\left(W_{jm'r}^{(h')}\right)\right] \\
&+ f_{mr} f_{m'r}\left(-\frac{\theta_{ij}^{(r)}}{M-1}\right) \\
&= \left(\mathrm{Cov}\left(\Lambda_{imr}^{(h)}, \Lambda_{jm'r}^{(h')}\right) + \pi_{ir}\pi_{jr}\right)\left(\mathrm{Cov}\left(W_{imr}^{(h)}, W_{jm'r}^{(h')}\right) + f_{mr} f_{m'r}\right) \\
&- f_{mr} f_{m'r}\pi_{ir}\pi_{jr} \\
&- f_{m'r}\left[\left(\mathrm{Cov}\left(\Lambda_{imr}^{(h)}, \Lambda_{jm'r}^{(h')}\right) + \pi_{ir}\pi_{jr}\right) f_{mr} - \pi_{ir}\pi_{jr} f_{mr}\right] \\
&- f_{mr}\left[\left(\mathrm{Cov}\left(\Lambda_{imr}^{(h)}, \Lambda_{jm'r}^{(h')}\right) + \pi_{ir}\pi_{jr}\right) f_{m'r} - \pi_{ir}\pi_{jr} f_{m'r}\right] \\
&+ f_{mr} f_{m'r}\left(-\frac{\theta_{ij}^{(r)}}{M-1}\right) \\
&= \left(-\frac{\theta_{ij}^{(r)}}{M-1} + \pi_{ir}\pi_{jr}\right)\cdot\left(0 + f_{mr} f_{m'r}\right) \\
&+ f_{mr} f_{m'r}\left(-\pi_{ir}\pi_{jr} + \frac{\theta_{ij}^{(r)}}{M-1} + \frac{\theta_{ij}^{(r)}}{M-1} - \frac{\theta_{ij}^{(r)}}{M-1}\right) \\
&= 0
\end{aligned}
$$

2.6.5 Between-group covariance of local ancestry on a locus

The result below for $R = 3$ is with groups A, B, C.

$$\begin{aligned} c_{AB} &= \mathrm{Cov}(\bar{\Lambda}^{(h)}_{imA}, \bar{\Lambda}^{(h')}_{jmB}) = \mathrm{Cov}(\Lambda^{(h)}_{imA}, \Lambda^{(h')}_{jmB}) = \mathrm{Cov}(\Lambda^{(h)}_{imA}, 1 - \Lambda^{(h')}_{jmA} - \Lambda^{(h')}_{jmC}) \\ &= -\Delta^{A}_{ij} - c_{AC} \ . \end{aligned}$$

Similarily:

$$\begin{aligned} c_{AB} &= \mathrm{Cov}(\bar{\Lambda}^{(h)}_{imA}, \bar{\Lambda}^{(h')}_{jmB}) = \mathrm{Cov}(\Lambda^{(h)}_{imA}, \Lambda^{(h')}_{jmB}) = \mathrm{Cov}(1 - \Lambda^{(h)}_{imB} - \Lambda^{(h)}_{imC}, \Lambda^{(h')}_{jmB}) \\ &= -\Delta^{B}_{ij} - c_{BC} \ . \end{aligned}$$

And

$$c_{AC} = -\Delta^{C}_{ij} - c_{BC} \ .$$

Thus

$$\begin{aligned} \Delta^{B}_{ij} + c_{BC} = \Delta^{A}_{ij} + c_{AC} = \Delta^{A}_{ij} - \Delta^{C}_{ij} - c_{BC} \implies c_{BC} &= \frac{\Delta^{A}_{ij} - \Delta^{B}_{ij} - \Delta^{C}_{ij}}{2} \\ \implies \mathrm{Cov}(\bar{\Lambda}^{(h)}_{imr}, \bar{\Lambda}^{(h)}_{jmr'}) &= \frac{\Delta^{(r'')}_{ij} - \Delta^{(r)}_{ij} - \Delta^{(r')}_{ij}}{2}, \quad r'' \in \{1,2,3\} \backslash \{r, r'\} \end{aligned}$$

2.6.6 Rewriting a γ expression

Consider the following:

$$\begin{aligned} &\sum_{m=1}^{M} \left[\gamma_{m1}\gamma_{m2} - \gamma_{m1}\gamma_{m3} - \gamma_{m2}\gamma_{m3} + \gamma^2_{m3}\right] \\ &= \sum_{m=1}^{M} \left[-\frac{1}{2}\left(\gamma_{m1} - \gamma_{m2}\right)^2 + \frac{1}{2}\gamma^2_{m1} + \frac{1}{2}\gamma^2_{m2} - \gamma_{m1}\gamma_{m3} - \gamma_{m2}\gamma_{m3} + \gamma^2_{m3}\right] \\ &= -\frac{1}{2}\sum_{m=1}^{M} \left[\left(\gamma_{m1} - \gamma_{m3}\right)^2 + \left(\gamma_{m2} - \gamma_{m3}\right)^2 - \left(\gamma_{m1} - \gamma_{m2}\right)^2\right] \\ &= -\frac{1}{2}\left[\sigma^2_{S,13} + \sigma^2_{S,23} - \sigma^2_{S,12}\right] \end{aligned}$$

Chapter 3

Methods (Outline)

3.1 Data description

The methods were applied to a data set of house sparrows from the Helgeland study population. (reuse some of the description from project thesis)

3.2 Creating GRMs

The available data is a "dosage" file on the `PLINK 1.9` (Chang et al. 2015) `.raw` format, containing allele counts for 3116 sampled individuals on roughly $180k$ SNP markers. Also available is morphological phenotype data for 1984 out of the 3116 genotyped individuals. We load genomic data into `R` with the `loadRAW` function from the `BGData` package (Grueneberg and de los Campos 2019). This creates the create genotype matrix $\mathbf{V}$, containing entries $V_{ij} \in \{0, 1, 2\}$. V_{ij} denotes the number of copies of the counted allele on individual i's jth SNP marker. These matrices are too large to manipulate in-memory, so we use the file-backed matrix system from `BGData`.

`BGData` has the function `getG` to generate different GRMs, using paralell computation. The function is essentially useful to compute $\mathbf{V}\mathbf{V}^\top$, while allowing custom centering and standardization for rows of $\mathbf{V}$, so that different GRMs can be created. Such as VR and GCTA, or any other, such as the generalization mentioned in Speed and Balding (2015).

The non-genetic groups GRMs have simple definitions and we can only need the allele frequencies, which can also easily be found with `BGData`. The extension of the Rio model requires addition data, however: the local ancestry of each allele. This will also be a very large data set, a $3116 \times 360k$ matrix indicating the local ancestry of each allele. We again use the file-backed matrices to handle this data. Group-specific genomic matrices are computed according to the definition in the extension of the Rio et al. (2020a) model (see pdf, soon background section)

3.3 Local Ancestry inference

- Local ancestry of an allele is which genetic group that specific allele is descended from.

- Requires a definition of the reference populations, i.e. which animals are purebreeds and which are admixed (have partial group memberships). In a wild system with dispersal such as this one, the choice is somewhat arbitrary.

- There are many available methods that infer local ancestry (Geza et al. 2019; Padhukasahasram 2014). Two were tried: EILA and Loter

- One option is the R package EILA (Yang et al. 2013). Does not allow missing data. However, in this case it does not converge for most chromosomes, only converges on chromosomes with just a few SNPs. Does not seem to handle 3 groups well.

- The other option used was the Python package Loter (Dias-Alves et al. 2018). Requires phased and imputed data. The command-line implentation has worked well for two different choices of reference populations so far. Including handling 3 groups.

- The following choice of reference populations was made for this analysis:

  - The reference populations are defined such that animals that are purebred according to the pedigree. The reference populations had 1336 (`inner`), 286 (`outer`) and 106 (`other`) individuals, leaving 1388 admixed individuals. This took a very long time to run: loter computation terminated after 1.5 weeks. Correlation between the group membership proportions from pedigree (group membership is inherited) vs. local ancestry (group membership is proportion of individual's alleles descended from that group) is 0.90 for inner group, 0.91 for outer group, 0.78 for other group. Results below are for this choice of reference population.

  - How to define reference population in the absence of pedigree information? Eg. BONE (Kuismin et al. 2020), combined with some cutoff value for when we are sufficiently sure an individual is from that group.

  - A second attempt at a reference population: use only natal data (how was this data found?) to define reference populations, unknown natal island = admixed. Uses no pedigree data. Currently running.

- Thought: is `other` group even necessary in the genomic case? Very few alleles are assigned to have local ancestry from this group. And in the genomic case we do not have the constrain that founders must be purebred, so the founders of the other group could be considered admixed if we wanted to, which would simplify model greatly.

3.4 Phasing: Inferring Haplotypes from SNP Data

- Was done using `Beagle 5.1` (Browning et al. 2018) with default settings.
- Done seperately for each reference population and admixed individuals
- Omitted some SNPs: The SNPs "not assigned to a particular chromosome or linkage group" and the SNPs "assigned to a linkage group but not to a specific chromosome"
- (Should also have possibly omitted the sex chromosome SNPs? Henrik mentioned this in email)
- Produces phased genotype data, i.e. haplotypes and also imputes missing data (replaces missing data with inferred values)

3.5 Model fit

(Can reuse parts from project thesis)

The animal model results were achieved using `INLA` (Rue et al. 2009), as it can be shown that the animal model can be stated as an LGM (true for genomic? ref?). Runtimes are less than five minutes using the shrunken GRMs, i.e. GRMs were only columns and rows for phenotyped individuals were kept. All models were rerun twice (using the `inla.rerun()` function) to improve stability and increase confidence in the results.

Penalized complexity priors were used (Simpson et al. 2017). For all models the variance components had penalized complexity priors $\text{PC}(1, 0.05)$, except the model for mass, which used a $\text{PC}(2, 0.05)$ prior (the model diverged otherwise).

Fixed effects included an overall intercept, month of measurement, sex, age at measurement, inbreeding coefficients as computed by Niskanen et al. (2020). The genetic group models also included group mean effects produced by the respective methods (pedigree-based or local ancestry-based), and the `inner` group was taken to be the reference group with group mean 0.

Random effects include (group-specific) (additive?) genetic variance

3.6 Resampling

If included: From Sorensen et al. (2001). (Mostly reuse section from project)

3.7 Side project: Genetic assignment using BONE

Kuismin et al. (2020). No progress so far

Chapter 4

Results and Discussion (Preliminary)

Genetic group model with wing length as response

	Fixed effects					
	sex	FGRM	month	age	outer	other
Genomic	-2.77;-2.77	-1.09;-1.09	-0.19;-0.19	0.46;0.46	-0.50;-0.50	-0.37;-0.37
	(-2.90, -2.64)	(-2.53, 0.34)	(-0.22, -0.15)	(0.43, 0.50)	(-0.74, -0.26)	(-0.77, 0.03)
Pedigree	-2.75;-2.75	-1.89;-1.89	-0.18;-0.18	0.46;0.46	-0.59;-0.59	-0.40;-0.40
	(-2.89, -2.60)	(-3.37, -0.41)	(-0.22, -0.15)	(0.43, 0.50)	(-0.80, -0.39)	(-0.72, -0.09)

Table 4.1: Mode;mean and 0.95 CI

Preliminary thoughts on overall results:

- Pedigree results are wrong compared to muff et al 2019
- Genomic group-specific variances seem reasonable.
- More variance because of considering haplotypes instead of genotypes?
- How does co-dominance assumption in Rio extension affect result? More gen. var. and less ID var?

Genetic group model with wing length as response

	Variances					
	$\hat{\sigma}^2_{\varepsilon}$	$\hat{\sigma}^2_{\text{year}}$	$\hat{\sigma}^2_{\text{ID}}$	$\hat{\sigma}^2_{\text{inner}}$	$\hat{\sigma}^2_{\text{outer}}$	$\hat{\sigma}^2_{\text{other}}$
Genomic	0.98;0.98	0.06;0.07	0.08;0.11	1.91;1.92	2.21;2.23	1.70;1.73
	(0.93, 1.04)	(0.02, 0.17)	(0.02, 0.33)	(1.67, 2.17)	(1.71, 2.85)	(1.01, 2.67)
Pedigree	0.98;0.98	0.07;0.08	0.50;0.51	2.87;2.88	4.43;4.51	2.10;2.20
	(0.93, 1.03)	(0.02, 0.20)	(0.33, 0.77)	(2.38, 3.42)	(3.18, 6.23)	(1.12, 3.85)

Table 4.2: Mode;mean and 0.95 CI

Genetic group model with mass as response

	Fixed effects					
	sex	FGRM	month	age	outer	other
Genomic	0.47;0.47	-1.06;-1.06	-0.29;-0.29	0.08;0.08	-0.55;-0.55	-0.17;-0.17
	(0.30, 0.65)	(-2.94, 0.81)	(-0.35, -0.23)	(0.02, 0.14)	(-0.87, -0.22)	(-0.68, 0.33)
Pedigree	0.45;0.45	-1.39;-1.39	-0.28;-0.28	0.08;0.08	-0.49;-0.49	-0.34;-0.34
	(0.27, 0.63)	(-3.25, 0.46)	(-0.34, -0.22)	(0.02, 0.13)	(-0.78, -0.21)	(-0.77, 0.09)

Table 4.3: Mode;mean and 0.95 CI

Genetic group model with mass length as response

	Variances					
	$\hat{\sigma}^2_{\varepsilon}$	$\hat{\sigma}^2_{\text{year}}$	$\hat{\sigma}^2_{\text{ID}}$	$\hat{\sigma}^2_{\text{inner}}$	$\hat{\sigma}^2_{\text{outer}}$	$\hat{\sigma}^2_{\text{other}}$
Genomic	2.88;2.88	0.04;0.05	0.62;0.64	1.78;1.80	2.48;2.53	0.80;0.90
	(2.72, 3.04)	(0.01, 0.13)	(0.34, 1.05)	(1.40, 2.32)	(1.68, 3.60)	(0.20, 2.15)
Pedigree	2.87;2.87	0.05;0.06	1.03;1.04	2.54;2.57	3.83;3.94	0.83;1.03
	(2.71, 3.02)	(0.01, 0.16)	(0.74, 1.42)	(1.94, 3.34)	(2.39, 6.11)	(0.12, 3.07)

Table 4.4: Mode;mean and 0.95 CI

Genetic group model with tarsus length as response

	Fixed effects					
	sex	FGRM	month	age	outer	other
Genomic	-0.08;-0.08	-0.68;-0.68	0.03;0.03	0.00;0.00	-0.02;-0.02	0.09;0.09
	(-0.15, -0.02)	(-1.41, 0.05)	(0.02, 0.04)	(-0.01, 0.01)	(-0.12, 0.09)	(-0.13, 0.30)
Pedigree	-0.09;-0.09	-0.77;-0.77	0.03;0.03	-0.00;-0.00	-0.02;-0.02	-0.02;-0.02
	(-0.16, -0.02)	(-1.50, -0.05)	(0.02, 0.04)	(-0.01, 0.01)	(-0.12, 0.08)	(-0.20, 0.15)

Table 4.5: Mode;mean and 0.95 CI

Genetic group model with tarsus length as response

	Variances of random effects					
	$\hat{\sigma}^2_{\varepsilon}$	$\hat{\sigma}^2_{\text{year}}$	$\hat{\sigma}^2_{\text{ID}}$	$\hat{\sigma}^2_{\text{inner}}$	$\hat{\sigma}^2_{\text{outer}}$	$\hat{\sigma}^2_{\text{other}}$
Genomic	0.02;0.02	0.01;0.02	0.25;0.26	0.38;0.38	0.23;0.23	0.40;0.41
	(0.02, 0.02)	(0.01, 0.04)	(0.20, 0.33)	(0.30, 0.48)	(0.14, 0.35)	(0.21, 0.70)
Pedigree	0.02;0.02	0.01;0.02	0.37;0.37	0.44;0.45	0.24;0.27	0.58;0.60
	(0.02, 0.02)	(0.01, 0.04)	(0.31, 0.43)	(0.34, 0.60)	(0.09, 0.59)	(0.27, 1.11)

Table 4.6: Mode;mean and 0.95 CI

Bibliography

1000 Genomes Project Consortium (2015). "A global reference for human genetic variation". In: *Nature* 526.7571, pp. 68–74.

Al Abri, Mohammed A et al. (2017). "Application of genomic estimation methods of inbreeding and population structure in an Arabian Horse Herd". In: *Journal of Heredity* 108.4, pp. 361–368.

Beck, Jon A. et al. (2000). "Genealogies of mouse inbred strains". In: *Nature genetics* 24.1, pp. 23–25.

Bérénos, Camillo et al. (2014). "Estimating quantitative genetic parameters in wild populations: a comparison of pedigree and genomic approaches". In: *Molecular ecology* 23.14, pp. 3434–3451.

Browning, Brian L. et al. (2018). "A one-penny imputed genome from next-generation reference panels". In: *The American Journal of Human Genetics* 103.3, pp. 338–348.

Chang, Christopher C. et al. (2015). "Second-generation PLINK: rising to the challenge of larger and richer datasets". In: *Gigascience* 4.1, s13742–015.

Charmantier, Anne et al. (2016). "Mediterranean blue tits as a case study of local adaptation". In: *Evolutionary Applications* 9.1, pp. 135–152.

Chase, Sherret S. (1952). "Production of Homozygous Diploids of Maize from Monoploids 1". In: *Agronomy Journal* 44.5, pp. 263–267.

Cohen, Jacob et al. (2013). *Applied multiple regression/correlation analysis for the behavioral sciences*. 3rd ed. Routledge.

Crossa, José et al. (2017). "Genomic selection in plant breeding: methods, models, and perspectives". In: *Trends in plant science* 22.11, pp. 961–975.

Dias-Alves, Thomas et al. (2018). "Loter: A software package to infer local ancestry for a wide range of species". In: *Molecular biology and evolution* 35.9, pp. 2318–2326.

Edwards, David (2015). "Two molecular measures of relatedness based on haplotype sharing". In: *BMC bioinformatics* 16.1, p. 383.

Faraway, Julian J. (2016). *Extending the linear model with R: generalized linear, mixed effects and nonparametric regression models*. 2nd ed. CRC press.

Galwey, Nicholas W. (2014). *Introduction to mixed modelling: beyond regression and analysis of variance*. 3rd ed. John Wiley & Sons.

Gelman, Andrew (2005). “Analysis of variance—why it is more important than ever”. In: *The annals of statistics* 33.1, pp. 1–53.
Geza, Ephifania et al. (2019). “A comprehensive survey of models for dissecting local ancestry deconvolution in human genome”. In: *Briefings in bioinformatics* 20.5, pp. 1709–1724.
Gienapp, Phillip et al. (2017). “Genomic quantitative genetics to study evolution in the wild”. In: *Trends in Ecology & Evolution* 32.12, pp. 897–908.
Givens, Goef H. and Jennifer A. Hoeting (2012). *Computational Statistics*. 2nd ed. John Wiley & Sons.
Grueneberg, Alexander and Gustavo de los Campos (2019). “BGData-A Suite of R Packages for Genomic Analysis with Big Data”. In: *G3: Genes, Genomes, Genetics* 9.5, pp. 1377–1383.
Hayes, Ben John et al. (2009). “Increased accuracy of artificial selection by using the realized relationship matrix”. In: *Genetics research* 91.1, pp. 47–60.
Henderson, C.R. (1984). *Applications of linear models in animal breeding*. University of Guelph Press.
Hill, W.G. and Bruce S. Weir (2011). “Variation in actual relationship as a consequence of Mendelian sampling and linkage”. In: *Genetics research* 93.1, pp. 47–64.
Jones, Adam G. and William R. Ardren (2003). “Methods of parentage analysis in natural populations”. In: *Molecular ecology* 12.10, pp. 2511–2523.
Kardos, Marty et al. (2016). “Genomics advances the study of inbreeding depression in the wild”. In: *Evolutionary applications* 9.10, pp. 1205–1218.
Kruuk, Loeske E. B. (2004). “Estimating genetic parameters in natural populations using the ‘animal model’”. In: *Philosophical Transactions of the Royal Society of London. Series B: Biological Sciences* 359.1446, pp. 873–890.
Kruuk, Loeske E. B. and Jarrod D. Hadfield (2007). “How to separate genetic and environmental causes of similarity between relatives”. In: *Journal of evolutionary biology* 20.5, pp. 1890–1903.
Kuismin, Markku et al. (2020). “Genetic assignment of individuals to source populations using network estimation tools”. In: *Methods in Ecology and Evolution* 11.2, pp. 333–344.
Legarra, Andres (2016). “Comparing estimates of genetic variance across different relationship models”. In: *Theoretical population biology* 107, pp. 26–30.
Lynch, Michael and Bruce Walsh (1998). *Genetics and analysis of quantitative traits*. Vol. 1. Sinauer Sunderland, MA.
Makgahlela, Mahlako Linah et al. (2013). “Across breed multi-trait random regression genomic predictions in the N ordic R ed dairy cattle”. In: *Journal of animal breeding and genetics* 130.1, pp. 10–19.
Meuwissen, Theo et al. (2016). “Genomic selection: A paradigm shift in animal breeding”. In: *Animal frontiers* 6.1, pp. 6–14.
Mrode, Raphael A. (2014). *Linear models for the prediction of animal breeding values*. 3rd ed. Cabi.
Muff, Stefanie et al. (2019). “Animal models with group-specific additive genetic variances: extending genetic group models”. In: *Genetics Selection Evolution* 51.1, p. 7.

Nietlisbach, Pirmin et al. (2017). “Pedigree-based inbreeding coefficient explains more variation in fitness than heterozygosity at 160 microsatellites in a wild bird population”. In: *Proceedings of the Royal Society B: Biological Sciences* 284.1850, p. 20162763.

Niskanen, Alina K. et al. (2020). “Consistent scaling of inbreeding depression in space and time in a house sparrow metapopulation”. In: *Proceedings of the National Academy of Sciences*.

Padhukasahasram, Badri (2014). “Inferring ancestry from population genomic data and its applications”. In: *Frontiers in genetics* 5, p. 204.

Pinheiro, José and Douglas Bates (2006). *Mixed-effects models in S and S-PLUS*. Springer Science & Business Media.

Ponzi, Erica et al. (2018). “Heritability, selection, and the response to selection in the presence of phenotypic measurement error: effects, cures, and the role of repeated measurements”. In: *Evolution* 72.10, pp. 1992–2004.

Quaas, R. L. (1988). “Additive genetic model with groups and relationships”. In: *Journal of Dairy Science* 71.5, pp. 1338–1345.

Quaas, R. L. and EJ Pollak (1981). “Modified equations for sire models with groups”. In: *Journal of Dairy Science* 64.9, pp. 1868–1872.

Rio, Simon et al. (Sept. 2020a). “Accounting for group-specific allele effects and admixture in genomic predictions: theory and experimental evaluation in maize”. In: *Genetics* 216.1, pp. 27–41.

Rio, Simon et al. (Mar. 2020b). “Disentangling group specific QTL allele effects from genetic background epistasis using admixed individuals in GWAS: an application to maize flowering”. In: *PLoS genetics* 16.3, e1008241.

Rue, Håvard et al. (2009). “Approximate Bayesian inference for latent Gaussian models by using integrated nested Laplace approximations”. In: *Journal of the royal statistical society: Series b (statistical methodology)* 71.2, pp. 319–392.

Schaeffer, L. R. (1991). “CR Henderson: Contributions to predicting genetic merit”. In: *Journal of dairy science* 74.11, pp. 4052–4066.

Searle, Shayle R. et al. (2006). *Variance components*. 1st ed. Vol. 391. John Wiley & Sons.

Simpson, Daniel et al. (2017). “Penalising model component complexity: A principled, practical approach to constructing priors”. In: *Statistical science* 32.1, pp. 1–28.

Sorensen, Daniel et al. (2001). “Inferring the trajectory of genetic variance in the course of artificial selection”. In: *Genetics Research* 77.1, pp. 83–94.

Speed, Doug and David J Balding (2015). “Relatedness in the post-genomic era: is it still useful?” In: *Nature Reviews Genetics* 16.1, pp. 33–44.

Technow, Frank et al. (2012). “Genomic prediction of hybrid performance in maize with models incorporating dominance and population specific marker effects”. In: *Theoretical and Applied Genetics* 125.6, pp. 1181–1194.

Thompson, Elizabeth A (2013). “Identity by descent: variation in meiosis, across genomes, and in populations”. In: *Genetics* 194.2, pp. 301–326.

VanRaden, Paul M. (2008). “Efficient methods to compute genomic predictions”. In: *Journal of dairy science* 91.11, pp. 4414–4423.

Wang, Bowen et al. (2017). “Efficient estimation of realized kinship from single nucleotide polymorphism genotypes”. In: *Genetics* 205.3, pp. 1063–1078.

Weir, Bruce S. et al. (2006). “Genetic relatedness analysis: modern data and new challenges”. In: *Nature Reviews Genetics* 7.10, pp. 771–780.

Wientjes, Yvonne CJ et al. (2017). “Multi-population genomic relationships for estimating current genetic variances within and genetic correlations between populations”. In: *Genetics* 207.2, pp. 503–515.

Wilson, Alastair J. et al. (2010). “An ecologist’s guide to the animal model”. In: *Journal of Animal Ecology* 79.1, pp. 13–26.

Wolak, Matthew E. and Jane M. Reid (2016). “Is pairing with a relative heritable? Estimating female and male genetic contributions to the degree of biparental inbreeding in song sparrows (Melospiza melodia)”. In: *The American Naturalist* 187.6, pp. 736–752.

— (2017). “Accounting for genetic differences among unknown parents in microevolutionary studies: how to include genetic groups in quantitative genetic animal models”. In: *Journal of Animal Ecology* 86.1, pp. 7–20.

Wright, Sewall (1922). “Coefficients of inbreeding and relationship”. In: *The American Naturalist* 56.645, pp. 330–338.

Yang, James J. et al. (2013). “Efficient inference of local ancestry”. In: *Bioinformatics* 29.21, pp. 2750–2756.

Yang, Jian et al. (2011). “GCTA: a tool for genome-wide complex trait analysis”. In: *The American Journal of Human Genetics* 88.1, pp. 76–82.

Zuur, Alain et al. (2009). *Mixed effects models and extensions in ecology with R*. Springer Science & Business Media.

Ødegård, Jørgen et al. (2018). “Large-scale genomic prediction using singular value decomposition of the genotype matrix”. In: *Genetics Selection Evolution* 50.1, pp. 1–12.