Masteroppgave

NTNU
Norges teknisk-naturvitenskapelige universitet
Fakultet for informasjonsteknologi og elektroteknikk
Institutt for matematiske fag

Kenneth Aase

Genetic Group Animal Models in the Genomics Era

Masteroppgave i Fysikk og matematikk
Veileder: Stefanie Muff

Januar 2021

NTNU
Kunnskap for en bedre verden

Kenneth Aase

Genetic Group Animal Models in the Genomics Era

Masteroppgave i Fysikk og matematikk
Veileder: Stefanie Muff
Januar 2021

Norges teknisk-naturvitenskapelige universitet
Fakultet for informasjonsteknologi og elektroteknikk
Institutt for matematiske fag

NTNU

Kunnskap for en bedre verden

Preface

The following master thesis is written for the course TMA4900 at The Norwegian University of Science and Technology (NTNU). It was supervised by Associate Professor Stefanie Muff at the Department of Mathematical Sciences. Professor Henrik Jensen at the Department of Biology provided additional guidance and feedback. Assoc. Prof. Muff and Prof. Jensen are both part of the Centre for Biodiversity Dynamics at NTNU, who provided the data used in the thesis.

Acknowledgements

My main methodological results in this thesis have been dependent on the foundations laid by Simon Rio, Laurence Moreau, Alain Charcosset and Tristan Mary-Huard in their work on the MAGBLUP-RI model.
A huge thank you goes out to Stefanie Muff for her guidance and support as my supervisor. Her honest and detailed feedback has been incredibly valuable to this work, as has her encouragement, positivity and availability at every step along the way. I could not have asked for a better supervisor. I'd also like to thank Henrik Jensen for his explanations of all things biological. One year ago my knowledge of genetics was as good as non-existent, so his involvement and considerations have been very helpful. Finally, I'd like to thank my girlfriend Kat Despain for her support, patience and for her help with proofreading.

Summary

This thesis deals with the use of genetic group animal models in the context of wild animal populations. The animal model is a type of generalized linear mixed model which lets us study a population's genetic parameters, such as the additive genetic variance. Through the use of genetic groups, the animal model can be used to investigate these parameters in genetically differentiated subpopulations. Animal models have traditionally been based on pedigree data, but genome-based approaches are becoming more common. The main focus of this text is an extension of a genome-based genetic groups animal model, which enables its usage on wild animal populations. Our extension involves gametic phasing of genotype data to allow for heterozygous genetic markers, and an expansion of the mathematical framework to allow for an arbitrary number of genetic groups. We contrast the genome-based approach with the traditional pedigree-based approach to animal models and genetic groups, which we also describe in detail. As a practical example, we apply the extended genome-based genetic groups animal model to a metapopulation of house sparrows residing on a system of islands in Northern Norway. For comparison, the equivalent pedigree-based model is also applied to the same data. Both models use a Bayesian framework. The model posteriors obtained from the genome-based model are mostly comparable to their pedigree-based counterparts. We see some limited patterns of disagreement between the two models, but these patterns are typical when comparing pedigree-based and genome-based animal models.

Sammendrag

Denne masteroppgaven tar for seg bruk av dyremodeller med genetiske grupper i studier der vi ser på villdyr-populasjoner. Dyremodellen er en generalisert lineær blandet modell som lar oss undersøke genetiske parametere i en populasjon, for eksempel additiv genetisk varians. Ved hjelp av genetiske grupper kan dyremodellen brukes til å granske disse parametrene i delpopulasjoner som har ulik genetisk struktur. Tradisjonelt sett har dyremodellen basert seg på stamtavledata, men i nyere tid har bruk av genomdata blitt mer vanlig. Hovedfokuset i denne masteroppgaven er en utvidelse av en dyremodell med genombaserte genetiske grupper, som lar oss bruke modellen i ville populasjoner. Utvidelsen vår bygger på gametisk fasing, noe som lar oss inkludere heterozygote genetiske markører, og på en videreutvikling av det matematiske rammeverket, noe som lar oss bruke et villkårlig antall genetiske grupper. Vi setter den genombaserte modellen i kontrast med tradisjonelle stamtavlebaserte dyremodeller og genetiske grupper, som vi også beskriver i detalj. Som et eksempel anvender vi den utvidete genombaserte dyremodellen med genetiske grupper på data fra en metapopulasjon av gråspurver som befinner seg på en øygruppe i Nord-Norge. Til sammenligning anvender vi også en tilsvarende stamtalvebasert modell på det samme datasettet. Begge modellene bruker et bayesiansk rammeverk. A posteriori-fordelingene til modellparametrene fra den genombaserte modellen samsvarer i hovedsak med de tilsvarende fordelingene fra den stamtavlebaserte modellen. Vi ser noen mindre uenigheter mellom de to modellene, men disse er typiske når man sammenligner stamtavlebaserte og genombaserte dyremodeller.

Table of Contents

Chapter 1

Introduction

Within evolutionary biology, the field of population genetics is the study of how genetic variation is distributed within and between populations, and the causes and consequences of such variation (Conner and Hartl 2004). Overall genetic differences between populations are a result of the opposing evolutionary forces of genetic drift and migration. Genetic drift reduces intra-population and increases inter-population genetic variation, while migration has the opposite effects. Furthermore, when the strength and direction of selection on phenotypes (observable biological traits) differs between populations (due to e.g. local conditions), we will see differences in the variation at the gene(s) underlying these adaptive phenotypes. Closely related to population genetics is quantitative genetics (Falconer and Mackay 1996; Lynch and Walsh 1998), which focuses on the study of the genetics, selection and evolution of complex and (usually) continuously varying phenotypes. In quantitative genetics we usually do not investigate the impact of the alleles (variations of a gene) at specific locations in the genome, but instead utilize overall summaries of the individually minor effects of the alleles at many different genes. The focus on continuous traits and the macro-level view of genotypes makes quantitative genetics well-suited to statistical analysis.

Quantitative genetics was originally developed for use in plant and animal breeding, where selection criteria are decided by the breeder (Henderson 1984). A breeder can manipulate the selection to artificially induce a desired change in a phenotype and might use quantitative genetics to study how to perform the selection in the most efficient way. Quantitative genetic theory was later applied in evolutionary ecology, where there are more pitfalls to consider (Charmantier, Garant, and Kruuk 2014). One might run into problems such as sampling issues and a lack of control groups. Questions of interest in ecological quantitative genetics include what causes some wild populations to be better at adapting to environmental changes than others, and the prediction of the rate and direction of future evolutionary change. Answers to these questions are urgently needed in conservation and wildlife management, for example. Quantitative genetics also plays a role in medicine, when polygenic genetic disorders in humans and other animals are studied. In this thesis we will focus on the quantitative genetics of wild animal populations.

One of the main goals of quantitative genetics is to disentangle the environmental and the genetic contributions to a phenotype within a population (Lynch and Walsh 1998; Falconer and Mackay 1996). This issue can be recognized from popular discourse as the question of "nature versus nurture." Additively disentangling the genetic and environmental components of the population phenotypic variance of different traits is of particular interest, as the additive part of the genetic variance has a major evolutionary importance. The additive genetic variance is a determinant of the expected degree of genetic resemblance between parents and their offspring. Thus, the rate of evolutionary change due to selection is determined by the additive genetic variance; the higher the level of additive genetic variance in the population, the faster it is able to respond to a given selection pressure (i.e., the higher the rate of adaptive evolution).

A well-established statistical tool in quantitative genetics is a linear mixed effects model known as "the animal model" (e.g., Kruuk 2004; Wilson et al. 2010). The animal model estimates additive genetic variance by considering the phenotypic values of individuals in a population for which we have information about the relatedness (genetic similarity) between individuals. Measures of relatedness allow the model to (additively) disentangle the degree to which having similar phenotypes correspond to having similar genomes, and thus detect the (additive) effect genes have on the phenotypic trait. Traditionally, relatedness information has been derived from pedigrees (i.e., family trees), which can provide measures of relatedness that are true on expectation. However, realized genetic similarity can often differ greatly from this expectation (Hill and Weir 2011). In addition, pedigrees constructed for wild populations are often error-prone (Keller et al. 2001; Ponzi, Keller, and Muff 2019).

Over the past two decades, the accessibility of genomic data has increased through improving genotyping technology (Meuwissen, Hayes, and Goddard 2016). A myriad of methods now use single nucleotide polymorphisms (SNPs) to derive measures of relatedness (Speed and Balding 2015). SNPs are specific positions in a species' genome where the alleles are especially variable, making these positions more informative. For instance, genome similarity measures can be obtained by comparing the genotypes of two individuals at every SNP (VanRaden 2008). Thus, animal models where relatedness information is extracted from genomic data have become feasible, with accompanying advantages and disadvantages compared to pedigree-based animal models.

One of the weaknesses of the animal model is that it does not allow subpopulations to have different genetic structures (Quaas 1988). This assumption is sometimes unrealistic, for example when different breeds are crossed in a breeding scenario, or when dealing with geographically structured wild populations with some dispersal between subpopulations. Genetically distinct subpopulations are denoted as "genetic groups," and models that incorporate genetic groups into the pedigree-based animal model exist (Wolak and Reid 2017; Muff et al. 2019). However, equivalent genome-based models were lacking until Rio et al. (2020a) recently proposed a genetic group animal model with a genome-based framework. The model relies on the idea of local ancestry (Geza et al. 2019), which lets us incorporate the fact that different sections of an individual's DNA originate from different genetic groups. However, the model proposed by Rio et al. (2020a) has certain limitations (stemming from its plant breeding origin) that preclude its usage on wild animal data.

This thesis will describe the animal model from a pedigree-based and genome-based

perspective, and present the existing pedigree-based genetic groups model. We then propose an extension of the genome-based genetic groups model, enabling it to be used for wild populations. Our extension involves an expansion of the mathematical framework introduced by Rio et al. (2020a) and utilizes gametic phasing of genotype data. As a proof of concept, we apply the extended genome-based animal model to a quantitative genetics analysis of a system of house sparrows (*Passer domesticus*) and compare our results to a corresponding pedigree-based model similar to the one in Muff et al. (2019). The sparrow population resides on islands in the Helgeland region of Northern Norway and is the subject of a long-running study by the Centre for Biodiversity Dynamics at NTNU (e.g. Jensen et al. 2008), who also provided the data for the analysis. We will operate within a Bayesian framework and will estimate posterior distributions of model parameters using INLA (Rue, Martino, and Chopin 2009). The main goal of the analysis is evaluating the performance of the genome-based genetic groups animal model.

Chapter 2

Background

2.1 Generalized linear mixed models

A generalized linear mixed model (GLMM) is an extension of the GLM, the generalized linear model (Pinheiro and Bates 2006; Zuur et al. 2009; Galwey 2014; Faraway 2016). While incorporating the linear predictors of a GLM, GLMMs also allow for random variable terms. 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 and with an arbitrary number of fixed and random effects. Letting $\boldsymbol{y}$ be the response vector, which we pass through some link function $\boldsymbol{f}(\cdot)$, the GLMM is given as

$$\boldsymbol{f}(\boldsymbol{y}) = \boldsymbol{\mu} + \mathbf{X}\boldsymbol{\beta} + \mathbf{Z}\boldsymbol{\eta}\,, \tag{2.1}$$

where $\boldsymbol{\mu}$ is an intercept vector, $\boldsymbol{\beta}$ is the vector of fixed effects and $\boldsymbol{\eta}$ is the random effect vector with some given multivariate distribution. The random effect $\boldsymbol{\eta}$ is usually 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 linear random intercept model with a single fixed effect (Cohen et al. 2013). In this model we introduce a grouping of the data where each group intercept takes 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 $\eta_i \sim \mathsf{N}\left(0, \sigma_\eta^2\right)$, then

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

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

So what is the purpose of including random effects? Take an example adapted from Galwey (2014, 1-20). Imagine a study with repeated measurements, meaning several observations are taken from each subject, which leads to a natural grouping of the data. This grouping by subject should be taken into account by the model to ensure the independence of residuals, a central assumption of GLMs. One way to avoid the violation of this assumption could be a model instead fit on the mean observed value for each subject, but we would naturally prefer to retain statistical power by including all observations directly. 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. Additionally, we are often not interested in inferences about the effect of each individual subject, but rather the greater population of subjects.

The use of random effects can solve this issue. We can include a random effect $\eta_i \sim \mathsf{N}\left(0, \sigma_\eta^2\right)$, which is independent and identically distributed (IID) between different subjects i. Fitting the model then involves estimating the variance σ_η^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 a summary statistic for each subject, while also causing the residual term present in the linear case 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, 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, Casella, and McCulloch 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 have 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.

2.2 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 known as "breeding values") as random effects to model some phenotypic trait as a response. Assume this trait was measured in N individual animals. An individual i's genetic value g_i denotes the impact of additive genetic effects on the individual's phenotype, that is, on the measured value of the trait. The source of non-independence considered by this random effect is the

potential similarity of two individuals' genomes, which can lead to similar genetic impacts on the phenotypes. For instance, closely related individuals are more likely to share the same alleles at their genes, potentially causing phenotypes of relatives to be correlated. To account for this correlation we must quantify to what degree the variation in trait values can be attributed to an individual's genes.

To tease out this genetic variation, we base the covariance structure of the genetic values on the relatedness between individuals, which we also will denote as their "kinship." Such a structure is obtained by having the vector of genetic values $\boldsymbol{g}$ follow the multivariate normal distribution

$$\boldsymbol{g} \sim \mathsf{N}\left(\mathbf{0}, \sigma^2_{V_\mathrm{A}}\mathbf{K}\right) \ , \tag{2.2}$$

where $\mathbf{K}$ is the symmetric $N \times N$ kinship matrix. The entry K_{ij} of $\mathbf{K}$ contains a measure of how similar the genomes of individuals i and j are. For off-diagonal entries a high value of K_{ij} denotes closely related individuals, where the range of possible values will depend on the choice of $\mathbf{K}$. For diagonal entries we usually have $K_{ii} \geq 1$, where the entries will be greater than 1 when inbreeding (i.e. mating of close relatives) is present. We can write $K_{ii} = 1 + F_i$, 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 kinship measures K_{ij}, as we shall explore below. In the definition of $\boldsymbol{g}$ in equation (2.2) the covariance structure $\mathbf{K}$ is scaled by $\sigma^2_{V_\mathrm{A}}$, the additive genetic variance of the population, which is often denoted simply as V_A in evolutionary ecology literature. The additive genetic variance 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 relatedness between individuals, encoded by $\mathbf{K}$. From the definition of the distribution of the genetic value vector $\boldsymbol{g}$ it is clear that the genetic values of two animals will only strongly covary if their genomes are similar and there is a high additive genetic variance present in the population. It is also clear that the estimated value of $\sigma^2_{V_\mathrm{A}}$ will depend on our choice of $\mathbf{K}$, so going forward we will denote $\sigma^2_{V_\mathrm{A}}$ differently if a specific $\mathbf{K}$ was used to find it. For example, $\sigma^2_{\boldsymbol{\kappa}}$ will be the additive genetic variance implied by the relatedness measure contained in kinship matrix $\boldsymbol{\kappa}$. A simple animal model for the continuous phenotype y_i of individual i, containing only an intercept μ, random effect genetic values g_i and residual term $\varepsilon_i \sim \mathsf{N}\left(0, \sigma^2_\varepsilon\right)$, can be stated as

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

2.2.1 Relatedness measures

In the context of animal models, $\mathbf{K}$ has customarily been inferred from observed pedigrees (i.e., family trees). Knowing from the pedigree how closely related two individuals are, one can estimate the expected amount of alleles at their genes that are shared 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 study 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).

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

More recently, an alternative method of directly inferring relatedness from the observed genotypes of SNP markers has gained popularity (Bérénos et al. 2014; Speed and Balding 2015; Gienapp et al. 2017). This genomic approach has become a viable option due to improvements in genomic technologies (Meuwissen, Hayes, and Goddard 2016; Ødegård et al. 2018), as the cost of large-scale genotyping is steadily decreasing and the identification and mapping of SNP markers in different species is improving (see e.g. Hagen et al. 2020, for house sparrow SNPs). In this section we will consider how to infer relatedness from pedigrees or SNPs and consider the advantages and disadvantages of each approach.

Relatedness inferred from pedigrees

We denote the version of the kinship matrix $\mathbf{K}$ that uses pedigree information as $\mathbf{A}$, which is also known as the "genetic relatedness matrix." For clarity, genetic value vectors based on pedigree-induced kinship $\mathbf{A}$ will be denoted $\boldsymbol{a}$ rather than $\boldsymbol{g}$. The matrix $\mathbf{A}$ is defined such that its ij^{th} entry A_{ij} denotes twice the expected probability ρ_{ij} that an allele 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, Anderson, and Hepler 2006). This expected probability ρ_{ij} is commonly known as the "coefficient of coancestry" (Lynch and Walsh 1998, 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 that 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 the relatedness matrix $\mathbf{A}$ gives us a measure of expected relatedness between each individual in the pedigree, without requiring direct knowledge about the genotypes at

any of their loci (specific DNA positions on a chromosome). Other advantages include explicitly accounting for inbreeding, and the lack of assumptions 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). After all, if the individuals are mostly unrelated, then there is little relatedness information to be gained from the pedigree. Methods, such as pedigree-based relatedness, that try to infer kinships based on individual ancestries are commonly referred to as identity-by-descent (IBD) methods.

A central concept when using the animal model with relatedness inferred from a pedigree is that 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, namely the root nodes in the family tree. We label their unknown parents as "phantom parents." Note that the phantom parents include not only the parents of the earliest cohort in the pedigree (known as the "founder population"), but also the 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 (Wilson et al. 2010; Wolak and Reid 2017). Any relatedness measure based on pedigrees is relative to its base population (Lynch and Walsh 1998, 132), and the genetic values of the base population are assumed to have a baseline mean of zero. Therefore, the pedigree-based animal model estimates σ^2_{A} (i.e. the pedigree-based additive genetic variance) for individuals in the base population and not the population as a whole. Furthermore, the genetic value of any non-base individual can be interpreted as its deviation in genetic merit from the base population. Thus, if we have a specific subpopulation for which we wish to measure the genetic parameters, we might choose to modify our pedigree so that its base population will equal the subpopulation of interest. Such a modification would involve disregarding the ancestors of members of this subpopulation and assuming that all individuals in the respective subpopulation are unrelated. Either way, the base population will necessarily be somewhat arbitrary, whether it is determined by a deliberate choice or by the constraints of our data collection. Such an arbitrary choice is nonetheless necessary, since the consequence of adding more and more ancestors to a pedigree would be ρ_{ij} converging to 1 for individuals far down the pedigree (Speed and Balding 2015). The cut-off must thus occur at some point.

One benefit of the pedigree-based approach is that the unrelatedness assumption in the base population leads $\mathbf{A}$ (and in particular its inverse) to be highly sparse (Henderson 1984). In fact, most pairs of non-base individuals will even not have any common ancestors, since the earliest level of ancestors will be unrelated. This sparseness leads to more effective calculation of σ^2_{A}. A disadvantage of relying on pedigrees is that the results can be sensitive to pedigree errors; one mistake in the pedigree can cause a cascade of errors through the generations and bias the results in unpredictable ways. Since fatherhood can be especially difficult to establish by observation, the error rate in paternal pedigree-links is generally high (Kruuk 2004; Ponzi, Keller, and Muff 2019). Thus, the aforementioned error-cascades are a common and hard-to-detect flaw in pedigree-based methods.

SNP-based relatedness measures

An inherent issue with the coefficient of coancestry is that actual (realized) relatedness between individuals can vary greatly from the expectation denoted by ρ_{ij} (Hill and Weir 2011). The actual probability of choosing two alleles that are identical by descent can be much greater or lower than what is indicated by the pedigree-derived relatedness estimate. Furthermore, as mentioned above, errors in observed pedigrees are not uncommon. We might therefore use realized relatedness rather than expected relatedness in order to get a more accurate measure of genome similarity (Hayes, Visscher, and Goddard 2009). For the direct estimation of the relatedness between two individuals we need 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, the 1000 Genomes Project Consortium (2015) found that two genomes typically differed at only 0.6% of the base pairs of nucleotides that make up the full genome. Therefore, when comparing genomes, we limit our focus to the loci where the genotypes *do* vary within a population.

A single nucleotide polymorphism, or SNP, is a genetic marker where the second most common allele occurs in a non-trivial proportion of the population. We will only consider diallelic loci, i.e. specific positions on a chromosome that only have two possible alleles. Denote the most common allele as 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, 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 "alternate" (usually minor) allele. Thus, when $v_{im} = 0$ individual i's m^{th} SNP is homozygous with two copies of the "reference" (usually 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 alternate allele. SNP-based relatedness matrices, generally called genetic relationship matrices (GRMs), all derive from the genotype matrix in some way (Speed and Balding 2015). Many of these definitions also include SNP m's alternate allele frequency p_m to weigh the importance of each SNP. The rationale behind the weighting is that two individuals sharing a minor allele with a very low allele frequency carries more information than sharing a minor allele that is almost just as likely as the major allele.

One example of this weighing is the GRM presented by VanRaden (2008), which is widely used (Crossa et al. 2017). This GRM, which we will mark by $\mathbf{G}_{\text{VR}}$, has its entries defined as

$$(G_{\text{VR}})_{ij} = \frac{\sum_{m=1}^{M} (v_{im} - 2p_m)(v_{jm} - 2p_m)}{2\sum_{m=1}^{M} p_m (1 - p_m)} = \frac{(\mathbf{V}_i - 2\boldsymbol{p})(\mathbf{V}_j - 2\boldsymbol{p})^\top}{2\sum_{m=1}^{M} p_m (1 - p_m)}, \qquad (2.3)$$

where $\mathbf{V}_k$ denotes the k^{th} row of $\mathbf{V}$, and $\boldsymbol{p}$ is the vector of alternate allele frequencies. In other words,

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

$\mathbf{G}_{\text{VR}}$ is standardized so that its diagonal has a mean value close to 1 (Legarra 2016). In other words, the scaling is similar to $\mathbf{A}$, with the diagonal entries close to 1 if i is outbred, i.e. not inbred. Thus we can again denote the inbreeding coefficient as $F_i = (\mathbf{G}_{\text{VR}})_{ii} - 1$.

A large number of GRMs exist in addition to $\mathbf{G}_{\text{VR}}$. For instance, the GRM introduced by Yang et al. (2011) is also widely used (see e.g. Bérénos et al. 2014; Al Abri et al. 2017). Speed and Balding (2015) suggest a general class of GRMs where a tuning parameter α is introduced, letting us define any number of genomic relationship matrices $\mathbf{G}_\alpha$. An even more general class of kinship estimators was found by Wang, Sverdlov, and Thompson (2017). In another approach, Wientjes et al. (2017) define $\mathbf{K}$ in such a way that can also be used in estimation of between-population genetic correlations.

Edwards (2015) constructs two IBD-based kinship matrices that are not derived from pedigrees. Instead, they rely on inferring relatedness from shared segments of DNA on the haplotype-level, that is, looking at each copy of a chromosome separately. Long regions of shared genes would suggest the existence of recent common ancestors, and thereby indicate the individuals are closely related. Haplotype-level methods such as these require the extra step of "phasing" the genotype data. Gametic phasing of diploid individuals involves inferring for each locus which of the two alleles on a locus is located on which of the two chromosomes (Excoffier, Laval, and Balding 2003). For each locus we thus obtain two haplotypes, one associated with each chromosome, letting us know which alleles in the genome are inherited from the same parent.

All this is to say that we have a plethora of GRMs to choose from. Furthermore, the relatedness 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 method. There is thus no universally correct choice of kinship matrix. Rather, the choice should depend on the data at hand and the genetic architecture of the study population (Speed and Balding 2015).

Note that in general the base population (i.e., the population for which we estimate the genetic parameters) in genome-based methods will differ from the base population in the pedigree-scenario, where the base population equals the set of phantom parents. In the IBS kinship methods with single-SNP comparisons, such as $\mathbf{G}_{\text{VR}}$, the base population will correspond to the population that the allele frequency is derived from (Hayes, Visscher, and Goddard 2009; Wientjes et al. 2017). Thus, single-SNP comparison methods have the potential advantage of letting the entire phenotyped population make up the base population, in contrast to pedigree-based methods. Unlike pedigree-based IBD methods, pedigree-free IBD methods such as those proposed by Edwards (2015) have less clearly defined base populations. In such methods genes must be traced back to the point in time where they first appeared by mutation, rather than tracing the genes back to the founders of a pedigree (Thompson 2013), leading to a base population comprised of disparate genes in various distant ancestors. Overall, a consequence of the discrepancies in base population that are caused by intrinsic differences between kinship estimators is that comparing additive genetic variances obtained from animal models relying on different kinship matrices $\mathbf{K}$ is problematic, since the results apply to different base populations.

Issues with comparison of results pertaining to different base populations can be partially resolved by rescaling the obtained variances to refer to the same base population, as described by Legarra (2016). Suppose we have have two kinship matrices $\mathbf{K}_1$ and $\mathbf{K}_2$

(with possibly different base populations) that have been used to produce two separate additive genetic variance estimates $\hat{\sigma}^2_{\mathrm{K}_1}$ and $\hat{\sigma}^2_{\mathrm{K}_2}$. Further, suppose we are interested in the additive genetic variance of a specific base population $\mathcal{B}$, which is a subset of individuals whose kinships are described by both of our two kinship matrices. Letting $\mathbf{L}_i$ (for $i = 1$ or $i = 2$) be a shrunken version of $\mathbf{K}_i$ which contains only the kinships pertaining to the preferred base population $\mathcal{B}$, we can scale the variance estimates so that

$$\hat{\sigma}^2_{\mathrm{L}_i} = \left(\overline{diag\left(\mathbf{L}_i\right)} - \overline{\mathbf{L}}_i\right)\sigma^2_{\mathrm{K}_i}\ , \tag{2.4}$$

where the first term is the mean of the diagonal of $\mathbf{L}_i$ and second term is the mean value of the entries of $\mathbf{L}_i$. Legarra (2016) then posits that the scaled additive genetic variances $\hat{\sigma}^2_{\mathrm{L}_1}$ and $\hat{\sigma}^2_{\mathrm{L}_2}$ will both refer to our chosen base population $\mathcal{B}$ and can thus be compared directly.

GRMs will generally be dense, that is, have very few elements equal to zero. An example of this denseness is clear from the definition (2.3). Even unrelated individuals will share alleles at a small amount of SNPs, and the use of p_m causes entries to almost never equal zero. Denseness in the GRMs is the major disadvantage of genome-derived relatedness compared to pedigree-derived relatedness, as it leads to higher computational cost and thus slower calculations. However, the computational cost is outweighed by improvements in the accuracy gained from using GRMs rather than $\mathbf{A}$ (Bérénos et al. 2014; Al Abri et al. 2017). Furthermore, genomic methods are not prone to the fickle biases induced by pedigree errors, though genomic data can also be used to validate and correct for mistakes in pedigrees (Flanagan and Jones 2019). On the other hand, the efficacy of using GRMs is reliant on the number of genotyped SNPs being sufficiently high. In fact, pedigrees can perform better when only a few genetic markers are available (Nietlisbach et al. 2017). Conversely, Bérénos et al. (2014) found that their additive genetic variance estimates stabilized at around 20 000 genotyped SNPs and that adding more markers did not lead to more accurate estimates. Thus, adding SNPs after a certain point does not improve results.

2.2.2 Complicating environmental effects

A major use of the animal model is in the estimation of $\sigma^2_{V_A}$ – the additive genetic variance in a population (Kruuk 2004; Wilson et al. 2010). In order to correctly estimate this parameter, we must account for other (possibly confounding) sources of covariance. Such covariance sources should therefore be included additional fixed or random effects in the animal model. These 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 similarities. 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 to include an ID random effect (Ponzi et al. 2018), as mentioned in Section 2.1. 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 that are not captured by genetic values (Wilson et al. 2010).

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 for a continuous trait y with B 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 $x_{ij}^{(b)}$ the corresponding measurement of fixed effect $b \in \{1, \ldots, B\}$. Let the additional random effects $z_{ij}^{(l)}$ have a normal distribution with zero mean and some covariance structure $\boldsymbol{\Sigma}_l$, so that $\boldsymbol{z}^{(l)} \sim \mathsf{N}\left(\mathbf{0}, \sigma_l^2 \boldsymbol{\Sigma}_l\right)$ for $l \in \{1, \ldots, L\}$ are the vectors of additional random effects. We let each random effect be independent of the other random effects and the genetic value, i.e. $\boldsymbol{z}^{(l)} \perp \boldsymbol{z}^{(l')} \perp g_i$ for $l \neq l'$. Then we can write

$$y_{ij} = \mu + \sum_{b=1}^{B} x_{ij}^{(b)} \beta_b + \sum_{l=1}^{L} z_{ij}^{(l)} + g_i + \varepsilon_{ij} \, . \tag{2.5}$$

The matrix form of this model is simply equation (2.1), with $\boldsymbol{f}(\boldsymbol{y}) = \boldsymbol{y}$ for a continuous $\boldsymbol{y}$ and with the genetic value vector $\boldsymbol{g}$ and residual effect vector $\boldsymbol{\varepsilon}$ included in $\boldsymbol{\eta}$. Since all random effects in equation (2.5) are normally distributed with zero mean, we can write

$$\mathsf{E}(y_{ij} | \boldsymbol{x}_{ij}) = \mu + \sum_{b=1}^{B} x_{ij}^{(b)} \beta_b \quad \text{and} \quad \mathsf{Var}(y_{ij} | \boldsymbol{x}_{ij}) = \sum_{l=1}^{L} \sigma_l^2 + \sigma^2_{V_\mathrm{A}} + \sigma^2_\varepsilon \, .$$

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 $\sigma^2_{V_\mathrm{A}}$ 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 $\sigma^2_{V_\mathrm{A}}$, that is, the additive genetic variance of a population of animals *given their sexes*.

2.2.3 Genetic groups extension of the animal model

As mentioned, the estimates of genetic parameters such as baseline mean genetic values and additive genetic variance produced by animal model apply to the base population. Thus, the animal model makes an implicit assumption that these genetic parameters are uniform across the entire 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 is violated. The violation of this assumption could lead the estimated mean genetic values and additive genetic variances to be biased towards their values among immigrants rather than the original study population.

These issues lead us to consider the possibility of partitioning the base population into *genetic groups* (Quaas and Pollak 1981; Quaas 1988; Wolak and Reid 2017). Rather than assuming that the population has genetic values $\boldsymbol{g} \sim \mathsf{N}\left(\mathbf{0}, \sigma^2_{V_\mathrm{A}}\mathbf{K}\right)$, 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, which we will also refer to as the "genetic group effect" of group r. The mechanism of partitioning the study population will differ when working with pedigrees or with genomic data. We will be differentiating between "purebred" individuals and "admixed" individuals. Purebred 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 1 and known immigrants to an "immigrant" genetic group 2, thereby incorporating the genetic structure in the base population into the model (as was done by Wolak and Reid 2016 and 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.

For admixed individuals it is not straightforward to split the genetic variance into group-specific genetic variances, as there is an additional source of variance that must be accounted for, namely the segregation variance (Slatkin and Lande 1994). This variance manifests due to group differences in allele effects and the level of linkage disequilibrium (LD; correlation between genotypes at different loci). Segregation variances can grow non-trivially large when considering admixed individuals in plant or animal breeding scenarios, when purebreds are crossed to form admixed individuals (see e.g. Rio et al. 2020a), or when the number of loci deciding the phenotypes is very low (Muff et al. 2019). We denote the segregation variance between groups r and r' as $\sigma^2_{\mathrm{S}_{rr'}}$. Since a segregation variance occurs between all combinations of groups, $R(R-1)$ segregation variances must be estimated in the presence of R genetic groups, quickly making the model much more computationally cumbersome as R increases. Thus, models that include segregation variances (such as Lo, Fernando, and Grossman 1993, Cantet and Fernando 1995 and García-Cortés and Toro 2006) require a lot more statistical power to fit. Luckily, the segregation variance will be small when using the infinitesimal model, that is, under the assumption that complex phenotypes are determined by very small contributions from genes at a large number

of loci. This assumption is very common in study of wild systems (Wilson et al. 2010). Thus, we can usually ignore segregation variance in such studies.

Pedigree-based genetic groups

If we have a pedigree available, it can be used to derive expected group membership proportions by tracing all matings and applying the usual Mendelian rules of inheritance (Schaeffer 1991; Wolak and Reid 2017). Each phantom parent must be assigned as a purebred in a single genetic group, which will depend on the criteria by which we define our groups (e.g. immigrant vs. non-immigrant). Thus, the base population will be partitioned into individuals belonging purely to different groups. Each partitioned part of the base population can then be considered the base population of a single genetic group. Define $q_{ir} \in [0, 1]$ as the membership proportion of individual i in genetic group r, so that

$$\sum_{r=1}^{R} q_{ir} = 1 \, .$$

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} equal the mean of each of i's (possibly phantom) parents' membership proportions in r. Thus, group membership is inherited through the generations, and all non-phantom individuals can have partial membership in various groups depending on their ancestry. This inheritance of group memberships will be true on expectation, considering an individual inherits half of their genetic material from each parent. So, in the same way that A_{ij} represents an *expected* probability, q_{ir} represents an expected group membership proportion.

To begin with, only let the genetic groups differ in their mean genetic value. We introduce u_i, an individual's "total additive genetic value," which can be defined as

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

where R is the number of genetic groups, and a_i is an entry in the pedigree-based genetic value vector $\boldsymbol{a}$, while the genetic group effects γ_r and group membership proportions q_{ir} are as defined previously. The above definition of the total genetic value u_i causes its mean to be a weighted average of the means of the different genetic groups, where the weights are i's group membership proportions. 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 $\boldsymbol{u}$ then has distribution $\mathsf{N}\left(\mathbf{Q}\boldsymbol{\gamma}, \sigma^2_{\mathrm{A}}\mathbf{A}\right)$.

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 baseline with $\gamma_{r'} = 0$. The effects γ_r for the other groups will denote deviation in mean total additive genetic effect from the baseline group.

We can also have the genetic groups to differ further by allowing heterogeneous additive genetic variance, through separating the genetic value vector $\boldsymbol{a}$ into a sum of "partial

genetic values" (Muff et al. 2019). Ignoring segregation variances for the reasons outlined previously, let $\boldsymbol{a} = \sum_{r=1}^{R} \boldsymbol{a}^{(r)}$, where $\boldsymbol{a}^{(r)}$ is the vector of partial genetic values with the individual-specific partial genetic values $a_i^{(r)}$ as its entries. Updating the definition of u_i, we can say

$$u_i = \sum_{r=1}^{R} \left[q_{ir}\gamma_r + a_i^{(r)} \right] . \tag{2.6}$$

Each partial genetic 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 $\boldsymbol{a}^{(r)}$ 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 genetic values to be independent because they originate from different base populations. Therefore, we can fit each partial genetic value as a random effect in the animal model. When introducing this decomposition of the random component of $\boldsymbol{u}$, we can write

$$\boldsymbol{u} \sim \mathsf{N}\left(\mathbf{Q}\boldsymbol{\gamma}, \sum_{r=1}^{R} \sigma^2_{\mathrm{A}_r} \mathbf{A}_r \right) . \tag{2.7}$$

When it comes to finding $\mathbf{A}_r$, consider the generalized Cholesky decomposition

$$\mathbf{A} = \mathbf{T}\mathbf{D}\mathbf{T}^{\top} , \tag{2.8}$$

where $\mathbf{T}$ will be an $N \times N$ lower triangular matrix with 1s on the diagonal and $\mathbf{D}$ is an $N \times N$ diagonal matrix (Mrode 2014, 23-25). $\mathbf{T}$ encodes for the gene flow between generations, so that its ij^{th} entry indicates the proportion of j's genes that i is expected to possess. The lower triangular entries of $\mathbf{T}$ 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. However, 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 all individuals but their descendants. Thus, they have inherited none of j's genes. Hence their contribution to the mean would be 0, which is why 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

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

Then $t_{ij}^{(r)}$ denotes the expected proportion of j's genes *within* group r that i possesses.

Meanwhile, the $\mathbf{D}$ in equation (2.8) scales the Mendelian sampling variance in genetic values according to the number of unknown parents and how inbred said parents are. The matrix is defined such that

$$d_{ii} = 1 - \frac{1}{4} \sum_{p \in \mathcal{P}_i} (1 + F_p) \ , \tag{2.9}$$

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 genetic value the fewer of i's parents are known, which is intuitive as we then have less relatedness information for i, which causes larger uncertainty in the actual genetic value. We can also see from this expression that an individual's genetic value will have less variance if its parents are severely inbred, which results in less diversity in the genes i can inherit. To get a group-specific $\mathbf{D}_r$, we modify definition (2.9) of d_{ii} so that

$$d_{ii}^{(r)} = 1 - \left(\frac{1}{|\mathcal{P}_i|} \sum_{p \in \mathcal{P}_i} q_{pr} \right) \left(\frac{1}{4} \sum_{p \in \mathcal{P}_i} (1 + F_p) \right) \ ,$$

where $|\mathcal{P}_i|$ is the number of known parents of i. In other words, we scale the second term in the definition of d_{ii} by the mean group membership proportion among known parents. This definition of $\mathbf{D}_r$ is an approximation, as an exact expression would also use group-specific inbreeding coefficients $F_p^{(r)}$ in the definition of $d_{ii}^{(r)}$. 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

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

So, through the use of genetic group effects γ_r and partial genetic values $a_i^{(r)}$, we can treat $\boldsymbol{u}$ as a genetic value vector where each individual's mean genetic value and additive genetic variance depends on its group membership proportions. Using the notation from equation (2.5), with a_i replaced by the definition of u_i in equation (2.6), we can state the genetic groups animal model with group-specific mean genetic value and additive genetic variance as

$$y_{ij} = \mu + \sum_{b=1}^{B} x_{ij}^{(b)} \beta_b + \sum_{r=1}^{R} \left(q_{ir} \gamma_r + g_i^{(r)} \right) + \sum_{l=1}^{L} z_{ij}^{(l)} + \varepsilon_{ij} \ , \tag{2.10}$$

where the partial genetic value vectors $\boldsymbol{a}^{(r)}$ are distributed as $\mathsf{N}\left(\mathbf{0}, \sigma^2_{\mathrm{A}_r} \mathbf{A}_r\right)$.

Genome-based genetic groups

In the genomic setting, we cannot trace the inheritance of expected partial group membership q_{ir} through the generations via knowledge of the pedigree. We therefore need some other way to determine group membership proportions for admixed individuals. Strandén and Mäntysaari (2013) suggest a genetic groups model, which was applied in

Makgahlela et al. (2013). Though the model is derived based on pedigrees, the authors claim genome-based genetic relationship matrices can be used in place of pedigree-based genetic relationship matrices. However, this model involves an approximation based on an assumption that the relatedness between an individuals' parents is zero. In other words, no inbreeding is present, which is not realistic in wild populations. Weir and Goudet (2017) present a hierarchical model which incorporates both relatedness and population structure (i.e. genetic groups), but does not use an animal model formulation.

Rio et al. (2020a) propose a genome-based genetic group animal model denoted as MAGBLUP-RI (multigroup admixed genomic best linear unbiased prediction random individual), that solves the issue of group membership proportions by using the *local ancestry* of each individual allele. An allele's local ancestry indicates which group the allele has descended from. MAGBLUP-RI involves defining the total genetic value U_i of individual i as a sum of contributions to the phenotype from each genotyped loci, where the contribution depends on the local ancestry of that locus. All loci are assumed to be homozygous, that is, they have two copies of the same allele. Let β_{mr}^{ref} or β_{mr}^{alt} be the contribution of locus m specific to group $r \in \{1, 2\}$, if locus m is homozygous with two reference or alternate alleles, respectively. Thus, we define the total genetic value

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

where Λ_{imr} is a random variable indicating whether the local ancestry of i's m^{th} locus is group r and W_{im} is a random variable indicating which allele is homozygously present at m. Using our notation for genotypes from Section 2.2.1, $W_{im} = 0$ indicates $v_{im} = 0$ and $W_{im} = 1$ indicates $v_{im} = 2$, while $v_{im} \neq 1$ due to the homozygosity assumption. We will give more details on the model in Section 3.1, but for now note that the main MAGBLUP-RI results are the group-specific GRMs with ij^{th} entries

$$\frac{\sum_{m=1}^{M} \lambda_{imr} \left(w_{im} - \hat{p}_{mr} \right) \lambda_{jmr} \left(w_{jm} - \hat{p}_{mr} \right)}{\sum_{m=1}^{M} \lambda_{imr} \lambda_{jmr} \hat{p}_{mr} \left(1 - \hat{p}_{mr} \right)} \times \hat{\theta}_{ij}^{(r)} = \widehat{\Gamma}_{ij}^{(r)} \times \hat{\theta}_{ij}^{(r)} \tag{2.11}$$

and a segregation covariance matrix $\boldsymbol{\Delta}$ with ij^{th} entries

$$\Delta_{ij} = \hat{\theta}_{ij}^{(1)} - \hat{\pi}_{i1} \hat{\pi}_{j1} \, . \tag{2.12}$$

The lowercase variables λ and w denote realizations of the random indicators Λ and W, respectively, while $\hat{p}_{mr}$ is the estimated alternate allele frequency within group r, $\hat{\pi}_{ir}$ is i's estimated group membership proportion in group r and $\hat{\theta}_{ij}^{(r)}$ is the estimate of i and j's shared group membership in group r. Note that the factor $\widehat{\Gamma}_{ij}^{(r)}$ in expression (2.11) is a modified version of the GRM $\mathbf{G}_{\text{VR}}$ proposed by VanRaden (2008), which we defined in equation (2.3). Firstly, the modification involves multiplying all terms in both the numerator sum and denominator sum in $\mathbf{G}_{\text{VR}}$ by $\lambda_{imr}\lambda_{jmr}$. Thus, genotypes only contribute to the relatedness estimate if they share local ancestry. Secondly, we no longer scale allele frequency centering by 2 since w can only take values 0 or 1, not 2. Finally, these group-specific relatednesses $\widehat{\Gamma}_{ij}^{(r)}$ are scaled by $\hat{\theta}_{ij}^{(r)}$, the shared group membership of

the individuals. The scaling is performed so that the impact on the group-specific additive genetic variance from a pair of individuals only comes from the proportion of their genes that originate from the same group.

In order to use the MAGBLUP-RI model to analyze wild study systems rather than artificial breeding setups, we have to introduce some modeling extensions. First, in the plant or animal breeding context, an individual can be assumed to be homozygous on (almost) every locus, that is, each locus has two copies of the same allele (Chase 1952). Such individuals are typically produced via systematically enforced breeding attempts between close relatives, resulting in extreme inbreeding (Beck et al. 2000). Thus, Rio et al. (2020a) assume homozygosity at every locus, which is also why MAGBLUP-RI only considers the local ancestry of each *locus*, not each *allele*. As animals in wild populations usually breed freely without human intervention, these populations have a high amount of heterozygous loci (that is, loci with at least two different alleles), even in populations where inbreeding occurs, unless the population is small enough that genetic drift becomes a major factor (Conner and Hartl 2004). Second, in a controlled breeding setup it is easy to restrict breeding to merely two genetic groups. Rio et al. (2020a) therefore assume the existence of only two genetic groups, which simplifies the analysis of the segregation variance. On the other hand, there is the potential for an arbitrary number of genetic groups to be present in a wild system, which justifies the need to extend the model to work in the case of more groups. In Section 3.1, we will present an extension of the MAGBLUP-RI model which allows for heterozygosity and an arbitrary number of genetic groups.

In a wild population local ancestry information is not readily available, and must therefore be inferred from the genotype data. Fortunately, many methods that perform this inference have been developed (Padhukasahasram 2014; Geza et al. 2019). These methods generally rely on the genotyped population having been partitioned so that each individual is designated as either purebred in a specific group or as admixed. Purebred individuals in a group are used as a reference for what the genomes of individuals from that group usually look like. The local ancestry inference methods then use these reference genomes to assign tracts of each admixed individual's genome as descended from a specific group. Thus, the local ancestries of the alleles of a purebred individual are all the same (a single group), while the local ancestries of the alleles within an admixed individual can vary across its genome.

2.3 Bayesian inference

In this analysis we will adopt a Bayesian framework for statistical inference (Givens and Hoeting 2012, 11-13). The Bayesian approach considers all model parameters as 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 treated as random 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 $\boldsymbol{x}$, 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. He and Hodges (2008) recommend using posterior modes for variance components such as the additive genetic variance in the animal model, because their posteriors are often skewed. 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, for example in numerical integration of (2.3). 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. To investigate the impact of the choice of prior, one can estimate $f(\boldsymbol{\psi}|\boldsymbol{x})$ when different priors are chosen to see how the posterior changes, a so-called prior sensitivity analysis.

Here we will use the Integrated Nested Laplace Approximation (INLA) technique to estimate the posteriors in the Bayesian model (Rue, Martino, and Chopin 2009). In short, INLA allows for fast Bayesian inference in latent Gaussian models by computing accurate approximations to the model posteriors. The class of latent Gaussian models includes a large number of models, such as the animal model (Steinsland and Jensen 2010).

Chapter 3

Methods

We will apply two different types of genetic groups animal models to measurements of various phenotypes of house sparrows living on islands in Northern Norway. One of the model setups will be the pedigree-based genetic groups model established in Section 2.2.3 and the other will be our genome-based genetic groups model, extended from the MAGBLUP-RI model introduced by Rio et al. (2020a). In implementing the models on a real house sparrow data set, our genetic groups will be allowed to differ in both their mean genetic values and in their additive genetic variance, and the genetic group structure will be based on which subset of islands the sparrows originate from. The primary focus will be a test of the genome-based genetic groups model's validity by checking if it provides reasonable results (i.e. results comparable to the pedigree-based model). To this end we try the models on three different phenotypes. Biological analysis of the results will be secondary, as Muff et al. (2019) already performed such an analysis of two of the phenotypes using the pedigree-based model. All `R` code and calls to other software used to generate results in the project is compiled in Appendix B.

3.1 Extension of MAGBLUP-RI

What follows is an extended version of MAGBLUP-RI (Rio et al. 2020a), as discussed in Section 2.2.3. The extension allows us to include heterozygosity without dominance effects rather than only homozygous lines, and $R > 2$ genetic groups. To take heterozygosity into account we will consider the two allele haplotypes at each locus *separately*, and thus expand MAGBLUP-RI to be a haplotype-level method. We thus split the genotype W_{im} with local ancestry Λ_{imr} into two haplotypes $W_{im}^{(1)}$ and $W_{im}^{(2)}$ with local ancestries $\Lambda_{imr}^{(1)}$ and $\Lambda_{imr}^{(2)}$, respectively. The extension from 2 to an arbitrary number R genetic groups is nontrivial, especially when segregation variances are involved. In this section we start by defining all necessary notation, then derive the covariance matrices between total genetic value vectors. Finally, we give estimators for all relevant parameters and a full statement of the genome-based genetic groups animal model.

3.1.1 Definitions

Again let the "genomic total additive genetic value" U_i of individual i be a sum of genetic contributions from each allele. In Section 2.2.3 we gave the MAGBLUP-RI definition of β_{mr}^{ref} and β_{mr}^{ref} as the contributions to the phenotype from a homozygous locus m originating from group r, with two reference or alternate alleles, respectively. We now assume these contributions to be equally distributed between the two alleles at the locus, so that a reference or alternate allele located on locus $m \in \{1, \dots, M\}$ and originating from genetic group $r \in \{1, \dots, R\}$ will have deterministic contributions $\frac{1}{2}\beta_{mr}^{\text{ref}}$ or $\frac{1}{2}\beta_{mr}^{\text{alt}}$ to the phenotype, respectively. We can think of these contributions as the "allele effects" of each allele. A homozygous locus with two reference alleles will still have a total contribution β_{mr}^{ref}, but the contribution from a heterozygous locus will equal the mean of the possible homozygous contributions. In other words, we assume no dominance effects, since the contributions from all heterozygous loci lie exactly between those loci's two possible homozygous contributions. Further note that we (like Rio et al. (2020a)) model allele effects to be group-specific, that is, two copies of the same allele can have a different effect if they originate from a different genetic group. If, for example, one of the haplotypes at locus m is a reference allele descended from group r, and the other haplotype is an alternate allele descended from group r', then the contribution from locus m is $\frac{1}{2}\left(\beta_{mr}^{\text{ref}} + \beta_{mr'}^{\text{alt}}\right)$.

To indicate an allele's local ancestry (i.e. 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 strands in a diploid organism the allele is located on. Practically, we will not differentiate between $h = 1$ or $h = 2$. We will assume that the local ancestries of these two alleles, each originating from one of the two chromosome strands in a diploid organism, are interchangeable when it comes to their likelihood of descent from a particular group. In other words, we assume that $\Lambda_{imr}^{(1)}$ and $\Lambda_{imr}^{(2)}$ are IID. $\Lambda_{imr}^{(h)}$ has the possible outcomes

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

The random vector $\mathbf{\Lambda}_{im}^{(h)} = \left[\Lambda_{im1}^{(h)}, \Lambda_{im2}^{(h)}, \dots, \Lambda_{imR}^{(h)}\right]$ has a categorical distribution, that is, a multinomial distribution with only one trial. Thus, exactly one of the entries of $\mathbf{\Lambda}_{im}^{(h)}$ equals 1, while the other entries equal 0. We parameterize the distribution with the probabilities $\pi_{ir} = \mathsf{P}\left(\Lambda_{imr}^{(h)} = 1\right)$ for each group r, where $\sum_{r=1}^{R} \pi_{ir} = 1$. The probability π_{ir} can be interpreted as i's true group membership proportion in group r. Note the difference in interpretation between π_{ir} and the previously used q_{ir}: the latter is i's IBD-derived *expected* group membership proportion, given that the pedigree is correct, while the former is its *actual* group membership proportion.

An important part of analyzing the covariance structure between total genetic values is the degree to which local ancestry tracts of two individuals' genomes overlap. We therefore define some parameters that measure just this. We denote by $\theta_{ij}^{(rr')}$ the proportion of alleles where i's allele belongs to group r and j's allele belongs to group r', which can also be written as $\mathsf{E}\left(\Lambda_{imr}^{(h)}\Lambda_{jmr'}^{(h)}\right)$. $\theta_{ij}^{rr'}$ thus measures the genome-overlap between i's

r-descended alleles and j's r'-descended alleles. In general $\theta_{ij}^{(rr')} \neq \theta_{ij}^{(r'r)}$ for $r \neq r'$, and $\theta_{ij}^{(rr')}$ does not depend on locus m and DNA strand h since the distribution of $\Lambda_{jmr}^{(h')}$ does not depend on m or h. In particular $\theta_{ij}^{(rr)}$, which we denote $\theta_{ij}^{(r)}$ for simplicity, is the shared group ancestry between individuals i and j in group r. An illustrative example of these overlapping genome-regions is given in Figure 3.1, with $R = 3$ and $\theta_{ij}^{(13)}$, $\theta_{ij}^{(2)}$, $\theta_{ij}^{(31)}$ $\theta_{ij}^{(32)}$ all equal to 0. Take extra note of the difference between $\theta_{ij}^{(12)}$ and $\theta_{ij}^{(21)}$. We can now define the covariance of two individuals' allele ancestries at a locus m using π and θ. Let $\mathsf{Cov}\left(\Lambda_{imr}^{(h)}, \Lambda_{jmr'}^{(h')}\right)$ be denoted by $\Delta_{ij}^{(rr')}$ (or simply $\Delta_{ij}^{(r)}$ if $r = r'$), which by the definition of covariance equals $\theta_{ij}^{(rr')} - \pi_{ir}\pi_{jr'}$. Note that this covariance does not depend on the values of h and h', since the designations $h = 1$ and $h = 2$ are arbitrary when considering two different individuals – they are only relevant within a given individual.

Haplotypes are given by the random variable $W_{im}^{(h)}$, which indicates the presence of the alternate allele at the m^{th} locus m on chromosome strand h in individual i's genome. Thus, a genotype v_{im} (as defined in Section 2.2.1) will equal the sum of the haplotypes $W_{im}^{(1)}$ and $W_{im}^{(2)}$. We specify the distribution of haplotype $W_{im}^{(h)}$ conditional on its group membership, letting

$$W_{im}^{(h)} \;\Big|\; \left(\Lambda_{imr}^{(h)} = 1\right) \sim \text{Bernoulli}(p_{mr}) \,.$$

The Bernoulli-parameter p_{mr} can be interpreted as the group-specific allele frequency of the alternate allele at locus m in group r, that is, how common the alternate allele is within that group. In other words, the genetic groups differ both in their allele effects and in their allele frequencies. Further note that we, similar to local ancestries, treat the two haplotypes on m with indices $h = 1$ and $h = 2$ separately, merely letting them share the same distribution. Let $\Gamma_{ij}^{(r)}$ denote the within-group genetic similarity of two individuals $i \neq j$, that is, the conditional correlation

$$\Gamma_{ij}^{(r)} = \mathsf{Corr}\left(W_{im}^{(h)},\, W_{jm}^{(h')} \;\Big|\; \Lambda_{imr}^{(h)} = 1,\, \Lambda_{jmr}^{(h')} = 1\right) \,,$$

regardless of whether $h = h'$ or not. Since it measures a genetic similarity of haplotypes in a group, $\Gamma_{ij}^{(r)}$ can be considered a relatedness conditional on group membership.

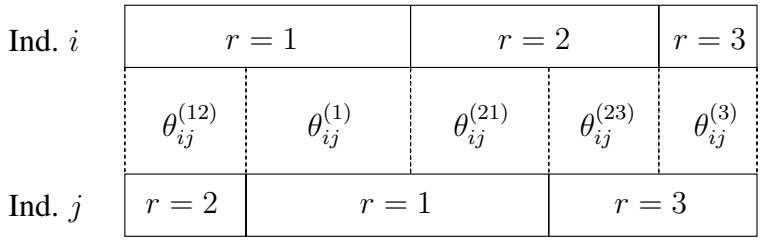


Figure 3.1: An example of a possible set of overlapping local ancestry regions in two individuals' genomes. In this particular example we have $R = 3$ and $\theta_{ij}^{(13)}$, $\theta_{ij}^{(2)}$, $\theta_{ij}^{(31)}$ $\theta_{ij}^{(32)}$ all equal to zero. Furthermore, all local ancestry regions are contiguous, which will not be the case in general.

We now define the genomic total genetic value U_i of individual i as the sum over the contributions of both alleles (with $h = 1$ or $h = 2$) at all loci m and from all groups r,

$$U_i = \sum_{m=1}^{M} \sum_{h=1}^{2} \sum_{r=1}^{R} \frac{1}{2} \Lambda_{imr}^{(h)} \left[\beta_{mr}^{\text{ref}} + W_{im}^{(h)} \left(\beta_{mr}^{\text{alt}} - \beta_{mr}^{\text{ref}} \right) \right] . \tag{3.1}$$

From the definition it is clear that locus m contributes the mean of the effects of its two alleles, which are dependent on the alleles' local ancestry and their haplotypes. The local ancestry $\Lambda_{imr}^{(r)}$ decides which group contributes a certain allele's effect, and $W_{im}^{(h)}$ decides whether said contribution comes from a reference or alternate allele. We can use this definition of total genetic value to define the genetic group effect γ_r of group r. Let the genetic group effect equal the sum of expected locus contributions γ_{mr} if all alleles belonged to that group, that is, $\Lambda_{imr}^{(h)} = 1$ for all m and both h, so that

$$\gamma_r = \sum_{m=1}^{M} \gamma_{mr} = \sum_{m=1}^{M} \sum_{h=1}^{2} \frac{1}{2} \left[\beta_{mr}^{\text{ref}} + p_{mr} \left(\beta_{mr}^{\text{alt}} - \beta_{mr}^{\text{ref}} \right) \right] .$$

Using this result we can see (Appendix A.1) that the expected total genetic value of individual i is

$$\mathsf{E}\left(U_i\right) = \sum_{r=1}^{R} \pi_{ir} \gamma_r \, .$$

Thus the mean total genetic value of individual i is a weighted sum of the group means, where the weights are the group membership proportions, similar to the pedigree-derived genetic groups model of equation (2.7).

By mean-centering the random variables and defining a group-specific version of $W_{im}^{(h)}$, that is,

$$\widetilde{\Lambda}_{imr}^{(h)} = \Lambda_{imr}^{(h)} - \pi_{ir} \quad \text{and} \quad \widetilde{W}_{imr}^{(h)} = \Lambda_{imr}^{(h)} \left(W_{im}^{(h)} - p_{mr} \right)$$

so that $\mathsf{E}\left(\widetilde{\Lambda}_{imr}^{(h)}\right) = 0$ and $\mathsf{E}\left(\widetilde{W}_{imr}^{(h)}\right) = 0$, we can rewrite (see Appendix A.2) the definition of U_i to an equivalent and useful form, namely

$$U_i = \mathsf{E}\left(U_i\right) + \sum_{m=1}^{M} \sum_{h=1}^{2} \frac{1}{2} \left[\sum_{r=1}^{R-1} \widetilde{\Lambda}_{imr}^{(h)} \left(\gamma_{mr} - \gamma_{mR} \right) + \sum_{r=1}^{R} \widetilde{W}_{imr}^{(h)} \left(\beta_{mr}^{\text{alt}} - \beta_{mr}^{\text{ref}} \right) \right] . \tag{3.2}$$

Since we in equation (3.2) split out the mean value of U_i, the random part of the total genetic value is captured by the second term. The alternative form (3.2) of the genetic value will be our starting point when determining the covariance structure of the total genetic values.

Lastly, we define the group-specific genetic variances and between-group segregation variances similarly to Rio et al. (2020a), except we use $2M$ in place of M since we have split genotypes into haplotypes. The genetic variance of group r is given as

$$\sigma_{\text{G}_r}^2 = \sum_{m=1}^{M} p_{mr} \left(1 - p_{mr}\right) \left(\beta_{mr}^{\text{alt}} - \beta_{mr}^{\text{ref}} \right)^2 ,$$

and the segregation variance between groups r and r' is given as

$$\sigma^2_{\mathrm{S}_{rr'}} = \frac{2M}{2M-1}\sum_{m=1}^{M}\left(\gamma_{mr}-\gamma_{mr'}\right)^2 - \frac{1}{2M-1}\left(\gamma_r-\gamma_{r'}\right)^2 \ . \tag{3.3}$$

The genetic variance thus depends on the haplotype variance $p_{mr}\left(1-p_{mr}\right)$ and squared differences in allele effect between the two alleles at a locus. Note that our definition of segregation variance differs from the definition given by Lynch and Walsh (1998, 227), whose definition resembles the first term in equation (3.3). As discussed in Section 2.2.3, there are several cases where the segregation variance is very small. If the difference between group contributions at each loci is small, the first term in equation (3.3) goes to zero, whereas the second term will be negligible if the number of loci M is large. Both of these conditions are assumed to hold under the infinitesimal model, which highlights why segregation variances are commonly neglected in wild systems. We shall nonetheless derive the full theoretical model, including all segregation variances, for the sake of completeness.

3.1.2 Covariance between total genetic values

We want to derive the covariance between the total genetic values U_i and U_j of two different individuals i and j, so that we can use the total genetic value vector $\boldsymbol{U}$ as a random effect in a genome-based genetic groups animal model. To this end, we will now explore the components of the covariance structure of total genetic values between two individuals, and the assumptions we make about said structure. The alternate expression (3.2) for U_i contains two types random variables, the centered local ancestries $\widetilde{\Lambda}$ and centered haplotypes $\widetilde{W}$. We thus need to consider covariances between haplotypes, covariances between local ancestries and covariances between haplotypes and local ancestries.

Starting with haplotype covariances, first note that we assume an absence of LD. In other words, $\mathsf{Corr}\left(W^{(h)}_{im}, W^{(h')}_{jm'}\right) = 0$ when $m \neq m'$ regardless of the values of the other super/subscripts. As shown in Appendix A.3.1, the assumption of no LD leads to no correlation for centered haplotypes; $\mathsf{Cov}\left(\widetilde{W}^{(h)}_{imr}, \widetilde{W}^{(h')}_{jm'r}\right) = 0$ for $m \neq m', \forall\ h, h'$. Further note that centered haplotypes do not covary across groups, since we in (Appendix A.3.2) find that $\mathsf{Cov}\left(\widetilde{W}^{(h)}_{imr}, \widetilde{W}^{(h')}_{imr'}\right) = 0$ for $r \neq r', \forall\ h, h'$. In fact, the only non-zero covariance between the centered haplotypes of different individuals is the within-group covariance

$$\mathsf{Cov}\left(\widetilde{W}^{(h)}_{imr}, \widetilde{W}^{(h')}_{jmr}\right) = \theta^{(r)}_{ij}\Gamma^{(r)}_{ij}p_{mr}\left(1-p_{mr}\right) \ , \tag{3.4}$$

(as always, regardless of h and h') as derived in Appendix A.3.3. Recall that $\theta^{(r)}_{ij}$ denotes the shared group membership of i and j in group r, while $\Gamma^{(r)}_{ij}$ is the group-specific relatedness between i and j.

Next, consider local ancestry covariances, which only matter if we include segregation variances, as is obvious from expression (3.2) for total genetic value and the definition of segregation variance (3.3). Since $\widetilde{\Lambda}$ is just a centered version of Λ, we know per definition

that

$$\mathsf{Cov}\left(\widetilde{\Lambda}^{(h)}_{imr}\ ,\ \widetilde{\Lambda}^{(h')}_{jmr'}\right) = \Delta^{(rr')}_{ij}\ \forall\ h, h'\ , \tag{3.5}$$

and we further show in Appendix A.4 that

$$\mathsf{Cov}\left(\widetilde{\Lambda}^{(h)}_{imr}, \widetilde{\Lambda}^{(h')}_{jm'r'}\right) = -\frac{\Delta^{(rr')}_{ij}}{2M-1}, \quad m \neq m',\ \forall\ h, h', r, r'\ . \tag{3.6}$$

Recall that M is the total number of loci, so the between-locus allele ancestry covariance will be closer to zero the more loci we are considering. One way to interpret this result is that the fewer loci we are considering, the more extreme the covariances in local ancestry between loci will be. Finally, as for covariances between haplotypes and local ancestry, we show in Appendix A.5 that

$$\mathsf{Cov}\left(\widetilde{\Lambda}^{(h)}_{imr}, \widetilde{W}^{(h')}_{jm'r'}\right) = 0 \tag{3.7}$$

for all super and subscripts.

Since the only non-zero covariances are equations (3.4) through (3.6), the covariance between total genetic values of different individuals can now be derived. The full derivation is given in Appendix A.6 and yields the following result. Letting $\mathcal{R} = \{1, \ldots, R\}$ for an arbitrary number of groups R, we obtain the following covariance structure

$$\begin{aligned}\mathsf{Cov}(U_i,\ U_j \mid \boldsymbol{\pi}_i, \boldsymbol{\pi}_j, \boldsymbol{\theta}_{ij}, \boldsymbol{\Gamma}_{ij}) &= \sum_{r=1}^{R} \theta^{(r)}_{ij}\Gamma^{(r)}_{ij}\sigma^2_{\mathsf{G}_r} \\ &+ \frac{1}{2}\sum_{r=1}^{R-1}\sum_{r'=r+1}^{R}\left(\Delta^{(r)}_{ij} + \Delta^{(r')}_{ij} - \sum_{r'', r^* \in \mathcal{R}\setminus\{r,r'\}} \Delta^{(r''r^*)}_{ij}\right)\sigma^2_{\mathsf{S}_{rr'}}.\end{aligned} \tag{3.8}$$

The covariance structure for group-specific additive genetic variances thus involves the group-conditional relatedness $\Gamma^{(r)}_{ij}$ scaled by the shared group membership proportion $\theta^{(r)}_{ij}$ (first term), while the structure for the segregation variances (second term) involves taking the sum of the within-group local ancestry covariances of the two groups in question, and subtracting all local-ancestry covariances within or between the other groups. For example, if $R = 4$, then the coefficient of $\sigma^2_{\mathsf{S}_{13}}$ would be

$$\frac{1}{2}\left(\Delta^{(1)}_{ij} + \Delta^{(3)}_{ij} - \Delta^{(2)}_{ij} - \Delta^{(24)}_{ij} - \Delta^{(42)}_{ij} - \Delta^{(4)}_{ij}\right)\ .$$

Since we will use $R = 3$ in our analysis, we note that in particular we have the following result:

$$\begin{aligned}\mathsf{Cov}(U_i, U_j \mid \boldsymbol{\pi}_i, \boldsymbol{\pi}_j, \boldsymbol{\theta}_{ij}, \boldsymbol{\Gamma}_{ij}) &= \sum_{r=1}^{3}\theta^{(r)}_{ij}\Gamma^{(r)}_{ij}\sigma^2_{G_r} \\ &+ \frac{1}{2}\sum_{r=1}^{2}\sum_{r'=r+1}^{3}\left(\Delta^{(r)}_{ij} + \Delta^{(r')}_{ij} - \Delta^{(r'')}_{ij}\right)\sigma^2_{S,rr'}\ ,\end{aligned}$$

where $r'' \in \{1, 2, 3\} \setminus \{r, r'\}, r \neq r'$.

3.1.3 Genome-based genetic group animal model

The covariance structure in equation (3.8) can be used in a GLMM to define a genome-based genetic groups animal model. In this section we show how a data set[1] of realized local ancestries $\lambda_{imr}^{(h)}$ and haplotypes $w_{im}^{(h)}$ for individual $i \in \{1, \ldots, N\}$, loci $m \in \{1, \ldots, M\}$, genetic groups $r \in \{1, \ldots, R\}$ and DNA strand $h \in \{1, 2\}$ can be used to estimate the model parameters p_{mr}, π_{ir}, $\theta_{ij}^{(r)}$, $\Delta_{ij}^{(r)}$ and $\Gamma_{ij}^{(r)}$. We define the parameter estimators similarly to Rio et al. (2020a), but using haplotypes rather than genotypes. Let the group allele-frequency p_{mr} at locus m in group r and true group membership proportion π_{ir} for individual i be estimated by the observed group allele-frequency

$$\widehat{p}_{mr} = \frac{\sum_{i=1}^{N} \sum_{h=1}^{2} \lambda_{imr}^{(h)} w_{im}^{(h)}}{\sum_{i=1}^{N} \sum_{h=1}^{2} \lambda_{imr}^{(h)}} , \tag{3.9}$$

and by the observed group membership proportion

$$\widehat{\pi}_{ir} = \frac{1}{2M} \sum_{m=1}^{M} \sum_{h=1}^{2} \lambda_{imr}^{(h)} , \tag{3.10}$$

respectively. The proportion of overlapping group memberships of groups r and r', $\theta_{ij}^{(rr')}$, can also be estimated by its observed variant, so

$$\widehat{\theta}_{ij}^{(rr')} = \frac{1}{4M} \sum_{m=1}^{M} \sum_{h=1}^{2} \sum_{h'=1}^{2} \lambda_{imr}^{(h)} \lambda_{jmr'}^{(h')} . \tag{3.11}$$

Recall that the designations $h = 1$ and $h = 2$ are arbitrary and will not correspond between different individuals. Thus, both alleles at locus m in one individual must be compared to both alleles on locus m in another individual, necessitating the double-sum over h and h' in the estimator for $\widehat{\theta}_{ij}^{(rr')}$. Moreover, the estimators for $\theta_{ij}^{(rr')}$ and π_{ir} imply the following definition of a $\Delta_{ij}^{(rr')}$-estimator:

$$\widehat{\Delta}_{ij}^{(rr')} = \widehat{\theta}_{ij}^{(rr')} - \widehat{\pi}_{ir} \widehat{\pi}_{jr'} . \tag{3.12}$$

When it comes to the group-conditional haplotype-correlations $\Gamma_{ij}^{(r)}$, that is, the group-specific relatedness, we recall the GRM given by Rio et al. (2020a), which we defined in expression (2.11). We further modify this GRM by summing over individual alleles (haplotypes) rather than genotypes. As in equation (3.11), we must compare both alleles at locus m in individual i with both alleles at locus m in individual j. Our modified estimator for $\Gamma_{ij}^{(r)}$ is therefore given by

$$\widehat{\Gamma}_{ij}^{(r)} = \frac{\sum_{m=1}^{M} \sum_{h=1}^{2} \sum_{h'=1}^{2} \lambda_{imr}^{(h)} \left(w_{im}^{(h)} - \widehat{p}_{mr} \right) \lambda_{jmr}^{(h')} \left(w_{jm}^{(h')} - \widehat{p}_{mr} \right)}{\frac{1}{2} \sum_{m=1}^{M} \sum_{h=1}^{2} \sum_{h'=1}^{2} \lambda_{imr}^{(h)} \lambda_{jmr}^{(h')} \widehat{p}_{mr} \left(1 - \widehat{p}_{mr} \right)} . \tag{3.13}$$

[1] A haplotype and local ancestry data set can be inferred from genotype data, as we will do in Section 3.3.1.

This particular design for the estimator $\widehat{\Gamma}_{ij}^{(r)}$ ensures that it is merely a special case of the well-known GRM $\mathbf{G}_{\text{VR}}$. One can show that in the case $R = 1$ equation (3.13) simplifies to the definition of $\mathbf{G}_{\text{VR}}$ given in (2.3), except that $v_{im} = w_{im}^{(1)} + w_{im}^{(2)}$. The equivalence with $\mathbf{G}_{\text{VR}}$ is also necessitates the factor $\frac{1}{2}$ in the denominator of (3.13), and this factor also ensures the correct standardization of the relatedness estimates. Namely, entries are standardized so that the diagonal entries $\widehat{\Gamma}_{ii}^{(h)}$ are close to 1 for individuals with $\pi_{ir} = 1$ in the absence of inbreeding. Note that using $\mathbf{G}_{\text{VR}}$ as a starting point for a group-specific GRM allows us to incorporate the local ancestry tracts in both the numerator and denominator sums in the definition of $\mathbf{G}_{\text{VR}}$. Thus, knowing the local ancestries adds value beyond just letting us utilize the summary statistics π_{ir} and $\theta_{ij}^{(r)}$.

Based on the covariance structure (3.8) we can now define the entries of the covariance matrices to be used in the animal model as

$$(\mathbf{G}_r)_{ij} = \widehat{\theta}_{ij}^{(r)} \cdot \widehat{\Gamma}_{ij}^{(r)} \tag{3.14}$$

for modeling group-specific additive genetic variance, and

$$(\mathbf{S}_{rr'})_{ij} = \frac{1}{2}\left(\widehat{\Delta}_{ij}^{(r)} + \widehat{\Delta}_{ij}^{(r')} - \sum_{r'',r^* \in \mathcal{R}\setminus\{r,r'\}} \widehat{\Delta}_{ij}^{(r''r^*)}\right) \tag{3.15}$$

for segregation variances. Thus, we have shown that the total genetic value can be written as

$$U_i = \sum_{r=1}^{R} \pi_{ir}\gamma_r + \sum_{r=1}^{R} g_i^{(r)} + \sum_{r=1}^{R-1}\sum_{r'=r+1}^{R} g_i^{(rr')}\,,$$

where $\boldsymbol{g}^{(r)} \sim \mathsf{N}\left(0, \sigma^2_{\mathrm{G}_r}\mathbf{G}_r\right)$ and $\boldsymbol{g}^{(rr')} \sim \mathsf{N}\left(0, \sigma^2_{\mathrm{S}_{rr'}}\mathbf{S}_{rr'}\right)$, and the distribution of the genomic total genetic value is

$$\boldsymbol{U} \sim \mathsf{N}\left(\boldsymbol{\Pi}\boldsymbol{\gamma}, \sum_{r=1}^{R}\sigma^2_{\mathrm{G}_r}\mathbf{G}_r + \sum_{r=1}^{R-1}\sum_{r'=r+1}^{R}\sigma^2_{\mathrm{S}_{rr'}}\mathbf{S}_{rr'}\right)\,,$$

where $\boldsymbol{\Pi}$ is the $N \times R$ matrix of group membership proportions. Now we can use U_i in place of g_i in the animal model equation (2.5) to fit a genomic genetic groups model, similar to what we did in the pedigree-case (2.10).

3.2 Data description

The Center for Biodiversity Dynamics (CBD) at NTNU has a long-running project where measurements of house sparrows on islands and the mainland of the Helgeland region in Northern Norway have been made annually since 1993 (see eg. Sæther et al. 1999; Jensen et al. 2008; Jensen et al. 2013). Said project provided the data used in this analysis, and the data we use includes phenotypic measurements from 1984 sparrows from a relatively isolated system of eight islands. These measurements were made in the years between 1993 and 2016. Because every bird is marked with a unique metal ring they are recognizable

throughout their lives. Several of the 1984 sparrows were measured repeatedly over the study period, so we have 4625 observations in total. Several phenotypic traits were measured for each sparrow, including wing length, body mass and tarsus length. The lengths were measured in millimeters, while body mass is given in grams. Other traits recorded for each sparrow include their sex and hatch year. In addition, we know the natal (birth) island for most birds. The date and island of each measurement were also recorded, and with the date and hatch year taken together we also know the age of each sparrow at the time of each measurement. Some data is missing for each of the phenotypic traits, with wing length missing 134 measurements, body mass missing 254 and tarsus length missing 130.

We also have access to genotype-data from 3116 individuals in the study system, including the 1984 phenotyped individuals. The genotyping was performed by taking blood samples from each individual, extracting DNA from the blood and genotyping the samples on an array with probes for 200 000 SNPs, as described in Lundregan et al. (2018). After quality control, 183 145 of the SNPs were retained in the analysis (with some missing genotypes). For all but 1782 SNPs we have a known relative position, which places SNPs on 30 different chromosomes (Lundregan et al. 2018). The genotypes have already been used in various applications, of which we will take advantage. A consistently scaled inbreeding coefficient F_{GRM} for every genotyped individual was computed by Niskanen et al. (2020), which we will use as a fixed effect accounting for inbreeding, as recommended by Reid and Keller (2010). The genetic assignment method of Kuismin et al. (2020) was applied to the data to infer the natal island of phenotyped sparrows that were missing this information (Saatoglu et al., in review). And notably, an extensive pedigree for the 3116 genotyped individuals was constructed from the data using the `R` package `SEQUOIA` (Huisman 2017). For further details on the construction of this pedigree, see Muff et al. (2019).

3.3 Statistical model

Within the study population, different subpopulations living on different islands are notably different, as subsets of islands differ in their habitats and environmental conditions (Muff et al. 2019). To account for possible genetic differences between these subpopulations originating from different island groups, we therefore partition the study population into genetic groups, where each group is associated with a set of islands. The populations on five islands closer to the mainland mostly live on dairy farms and enjoy more stability and larger population sizes than the sparrows on the islands further out to sea, which mostly live in local people's gardens. We label the former group of islands as the `inner` genetic group (encoded as 1), and the latter group as the `outer` genetic group (encoded as 2). The house sparrows on the remaining islands in the study system have not been SNP-genotyped yet, and we have notably fewer observations from them. We lump these islands together in a final genetic group `other` (encoded as 3).

We will use these genetic groups to test our new genomic genetic groups model and compare the results with an otherwise similar pedigree-based genetic groups model. The comparison will be performed in models using three different phenotypic responses: wing length, body mass and tarsus length.

3.3.1 Genetic group setup

Pedigree-based genetic groups

For the pedigree-based genetic groups, we must assign all phantom parents to the base population of one of the genetic groups, as described in Section 2.2.3. The assignment of phantom parents to groups is done based on the natal island of their offspring. In other words, for sparrows with unknown parents and a known natal island, the unknown (phantom) parent is assigned to the genetic group associated with the natal island. For the phantom parents whose offspring's natal island is unknown, we instead use the first island the offspring was observed on. The reason for this procedure is that having offspring that were present on an island in a given year is the best available evidence we have that the phantom parent also originated there. With the base population partitioned, both $\mathbf{Q}$ and $\mathbf{A}$ can be found found using functions from the `R` package `nadiv` (Wolak 2012). We can then proceed to find the group-specific relatedness matrices $\mathbf{A}_1$, $\mathbf{A}_2$ and $\mathbf{A}_3$ using the methods described in Section 2.2.3.

It is possible to obtain a measure of the relative sizes of the genetic groups within the population by summing over all individuals' expected group membership proportions q_{ir} in a given group and by dividing by the number of individuals. Considering only the phenotyped individuals, we estimate that a proportion 0.76 of the genetic material is expected to belong to `inner`, 0.18 to `outer` and 0.06 to `other`.

Genome-based genetic groups

In the case of our genome-based genetic groups, some extra steps are required. First note that about a third of the SNPs in the genotype data are heterozygous, and we have three genetic groups, justifying the need to use our extended genetic groups model. In the genomic case we have more freedom to define the purebred and admixed populations, as we are not forced to treat every founder individual as a purebred. The status as founder is irrelevant in the local ancestry approach, only genetic similarity matters. So we only need to partition the population into the purebreds from each genetic group, and the admixed individuals, which have partial group memberships. For the sake of easy comparison, we will simply consider any individual that is admixed in the pedigree-based setup to also be admixed in the genome-based setup. Conversely, individuals that are purebred in a single group based on the pedigree will be considered purebred in the same group in the genome-based model. Note that these purebred individuals include not only founders of the pedigree, but also offspring of purebred parents from the same genetic group. Among the 3116 genotyped individuals we have 1336 purebred `inner` individuals, 286 purebred `outer` individuals, 106 purebred `other` individuals and 1388 admixed individuals. The subset of 1984 phenotyped individuals contains 1002 purebred `inner` individuals, 144 purebred `outer` individuals, 50 `other` individuals and 788 admixed individuals. Reasonable amounts of phenotypic measurements are available for all four types of individuals, as missing phenotypic measurements are proportionally distributed between the admixed and purebred populations. For all phenotypes, the relevant subpopulation with the greatest proportion of missing phenotypes is purebred `inner`, which has 1.6% missing wing length measurements and 2.8% missing body mass measurements, and 1.6% missing tarsus length measurements.

As a first step in finding the genomic genetic groups, we must perform the gametic phasing of the genotype data. The phasing procedure determines haplotypes $w_{im}^{(h)}$, so we know not only the genotype at each locus, but also which chromosome copy each of the two alleles on a locus belongs to. After using `PLINK 1.9` (Chang et al. 2015) to convert the genomic data to the appropriate input format, we used `Beagle 5.1` with default settings to perform the gametic phasing (Browning, Zhou, and Browning 2018). The phasing was done separately on each of the purebred populations and the admixed individuals since they are assumed to be genetically distinct. 1782 of the SNPs were not assigned to a specific chromosome in the reference genome (Lundregan et al. 2018), precluding them from gametic phasing. These SNPs are therefore omitted from the remaining genomic analysis, leaving us with 181 363 SNPs. In addition to the gametic phasing, `Beagle` imputes any missing genotypes in the genomic data, that is, all missing values are inferred from the other data in the population.

Next, we need to perform the local ancestry inference to determine the group of origin of every considered allele in the admixed population. As output from the local ancestry inference we obtain the local ancestries $\lambda_{imr}^{(h)}$. To this end we have used the command-line version of the Python package `Loter` (Dias-Alves, Mairal, and Blum 2018), out of several possible alternatives (outlined in Geza et al. 2019). `Loter` requires as input phased and imputed genotype data, and is able to handle three genetic groups. However, using three rather than two groups disables `Loter`'s method of correcting for errors in the gametic phasing, making the correctness of the initial phasing (using eg. `Beagle`) more crucial. Thus, the authors of `Loter` recommend against using the length of ancestry tracts in analysis where 3 groups are present, which would be relevant in methods trying to establish group-specific *ancestries* (Dias-Alves, Mairal, and Blum 2019). The run time of the local ancestry inference was relatively slow, taking roughly one week when using eight Intel Xeon (2.6 GHz) CPUs on a shared computational server.

With the inferred values of the haplotypes $w_{im}^{(h)}$ and local ancestries $\lambda_{imr}^{(h)}$ we can compute group-specific GRMs $\mathbf{G}_1$, $\mathbf{G}_2$ and $\mathbf{G}_3$ and segregation covariance matrices $\mathbf{S}_{12}$, $\mathbf{S}_{13}$ and $\mathbf{S}_{23}$ using equations (3.10) - (3.15). However, we will disregard the segregation variance in this analysis, under the assumption of the infinitesimal model. A challenge of implementing these formulas is that we are dealing with very large data sets, namely $3116 \times 181\,363 \times 2$ haplotypes, and the same number of local ancestries for each group. The genomic data is too large to conveniently manipulate directly in-memory in `R`, and computing the matrix products in particular require a huge amount of memory. To overcome this memory limitation issue, we utilized the file-backed matrices implemented in the `R` package `BGData` (Grueneberg and de los Campos 2019). `BGData` has been developed specifically for the manipulation of large genomic data sets, and its linked file-backed matrices allows one to treat very large matrices as if they were loaded in-memory. The package also includes the function `getG()`, which is used to efficiently compute matrix products with parallel methods. Note that several of the estimators in Section 3.1.3 involve computations on this matrix-matrix product form.

3.3.2 Model description

Given the partial relatedness matrices $\mathbf{A}_r$ and group-specific GRMs $\mathbf{G}_r$ we can now formulate the full pedigree and genome-based models. As continuous fixed effects we included an intercept μ, age, the month of measurement (May through August treated numerically), the previously mentioned inbreeding coefficients denoted here as F_{GRM} and genetic group effects γ_r. Our only categorical fixed effect was sex, with 0 representing males and 1 representing females. We chose `inner` to be the reference group for identifiability reasons, so $\gamma_1 = 0$. The genetic group effects γ_2 and γ_3 then denote the deviation in the respective group's mean total additive genetic effect from `inner` (Wolak and Reid 2017). We use q_{ir} and $\widehat{\pi}_{ir}$ as the covariates used in estimating γ_r for the pedigree-based and genome-based models, respectively.

Random effects in the models include the group-specific genetic values for each of the genetic groups, hatch year, island of measurement, an individual identity effect and a residual random effect. The individual effect is included to account for permanent environmental effects since there are repeated measurements (Wilson et al. 2010). Group-specific genetic values $\boldsymbol{g}^{(r)}$ have covariance structure $\mathbf{G}_r$ in the genome-based model, while in the pedigree-based model $\boldsymbol{a}^{(r)}$ have structure $\mathbf{A}_r$. Despite only 1984 sparrows having phenotype data, all 3116 genotyped sparrows were used in setting up the group-specific kinship matrices $\mathbf{A}_r$ and $\mathbf{G}_r$ to obtain the most accurate possible relatedness estimates. After finding these matrices, but before fitting the model, we removed all rows and columns in $\mathbf{A}_r$ and $\mathbf{G}_r$ corresponding to sparrows that have not been phenotyped, leaving us with trimmed 1984×1984 matrices. All relatedness information between sparrows is retained, while we still only consider the relevant (i.e. phenotyped) individuals. Additionally, a very small value of 10^{-12} was added to the diagonals of the $\mathbf{G}_r$ matrices to make them positive definite, and thus proper covariance matrices. This addition is necessary because whenever there are individuals with $\pi_{ir} = 0$ for some group r (such as purebreeds) there will be zeros on the diagonal, and thus at least one eigenvalue will be zero.

Inclusion of the island of measurement as a common environment random effect is especially critical in this model. The effect not only deals with environmental covariance, but also ensures that the genetic groups estimate what they are intended to estimate. Our genetic groups are based on geographic origin, so the absence of an island effect might lead to the genetic group effects capturing environmental differences between the islands rather than capturing potential genetic differences. The presence of some dispersal between the islands (8.9% of recorded sparrows changed their island group during their lifetime; Saatoglu et al., in review) ensures that the data set contains sparrows measured on islands not corresponding to their genetic group, allowing the model to disentangle the two sources of variance.

A full mathematical statement of the pedigree-based model is

$$y_{ij} = \mu + \boldsymbol{x}_{ij}^{\top}\boldsymbol{\beta} + \sum_{r=1}^{2} q_{ir}\gamma_r + \sum_{r=1}^{3} a_i^{(r)} + id_i + island_{ij} + year_{ij} + \varepsilon_{ij}\ , \qquad (3.16)$$

where y_{ij} is the jth phenotypic measurement for individual i and $\boldsymbol{x}_{ij}$ is a vector storing the fixed covariates sex, age, month and F_{GRM}. Furthermore, $\boldsymbol{\beta}$ is a vector of the fixed effects, q_{ir} is i's expected group membership proportion in group r, and γ_r is the fixed genetic

group effects in groups. The group-specific genetic values $a_i^{(r)}$ are entries in the random vector $\boldsymbol{a}^{(r)} \sim \mathsf{N}\left(\mathbf{0}, \sigma^2_{\mathrm{A}_r}\mathbf{A}_r\right)$, while the random effects id_i, $island_{ij}$, $year_{ij}$ and ε_{ij} are distributed with $\mathsf{N}\left(0, \sigma^2_{\mathrm{ID}}\right)$, $\mathsf{N}\left(0, \sigma^2_{\mathrm{island}}\right)$, $\mathsf{N}\left(0, \sigma^2_{\mathrm{year}}\right)$ and $\mathsf{N}\left(0, \sigma^2_{\varepsilon}\right)$, respectively. Recall, the Bayesian modeling approach involves finding posterior distributions of the variance of each of these random variables. Using similar notation, the equivalent genome-based model can be stated as

$$y_{ij} = \mu + \boldsymbol{x}_{ij}^{\top}\boldsymbol{\beta} + \sum_{r=1}^{2}\widehat{\pi}_{ir}\gamma_r + \sum_{r=1}^{3} g_i^{(r)} + id_i + island_{ij} + year_{ij} + \varepsilon_{ij}\ , \tag{3.17}$$

where $\widehat{\pi}_{ir}$ is the group membership proportion estimated from local ancestries and $g_i^{(r)}$ is an entry in the random vector $\boldsymbol{g}^{(r)} \sim \mathsf{N}\left(\mathbf{0}, \sigma^2_{\mathrm{G}_r}\mathbf{G}_r\right)$.

3.3.3 Implementation

The Bayesian genetic group animal models were implemented with the `R-INLA` package (Rue, Martino, and Chopin 2009). All models were rerun twice (using the `inla.rerun()` function) with the posterior modes from the previous run of the model as new starting points, in order to improve model stability and increase confidence in the results. Using the previously mentioned computational setup and trimmed kinship matrices, run-times (including the two reruns) for the genome-based models were roughly an hour for each of the phenotypes. For the pedigree-based models, the equivalent run-times were around three minutes.

In terms of priors, for the fixed effects we used wide normal distributions centered at zero, $\mathsf{N}\left(\mathbf{0}, 10^3\mathbf{I}\right)$. The random effect variances were given priors on their inverse, that is, their precision, as is common to work with in Bayesian statistics. The precisions were all given penalized complexity (PC) priors (Simpson et al. 2017). A PC prior $\mathsf{PC}(\nu, \alpha)$ for $\nu > 0$ and $\alpha \in (0, 1)$ on the precision $\frac{1}{\sigma^2}$ will assign the proportion α of the prior probability weight to the case $\sigma > \nu$. One can thus control how much to weigh values of σ over a certain threshold ν. All model variance components were given penalized complexity priors $\mathsf{PC}(1, 0.05)$. Thus we gave the most weight to simpler models with low variances, since the prior assumption is that $\mathsf{P}(\sigma > 1) = 0.05$ for all random effect variances σ^2 in the model. The data was forced to convince the prior that giving more probability weight to larger variances is acceptable.

Chapter 4

Results

4.1 Group membership proportions

We check the efficacy of the local ancestry inference by comparing the local ancestry-derived group membership proportions, $\widehat{\pi}_{ir}$, to their pedigree-derived counterparts, q_{ir}. Since the pedigree-based group membership proportions are true on expectation (given a correct and complete pedigree), they should mostly correspond to the realized group membership proportions. One way to investigate whether this correspondence occurs is to simply consider the correlation between q_{ir} and π_{ir} for each group r. Recall that π_{ir} was only estimated for admixed individuals and was assumed equal to q_{ir} for purebred sparrows. We therefore only check the correlation between $\widehat{\pi}_{ir}$ and q_{ir} within the phenotyped admixed subpopulation of 788 individuals. The correlations are 0.89, 0.90 and 0.79 for `inner`, `outer` and `other`, respectively.

We can also investigate the relationship between q_{ir} and π_{ir} by examining their scatter plots for different groups (Figure 4.1). Again, we limit the comparison to phenotyped admixed individuals, as $\pi_{ir} = q_{ir}$ for purebred individuals.[1] We see that for most individuals the two genetic group methods give corresponding results. The group membership proportions are especially concentrated along the diagonal of the scatter plots for `inner` and `outer`, indicating agreement. For the `inner` group more points are concentrated in the upper right corner, indicating that both methods assign the greatest genome proportions to this group. Conversely, the opposite pattern is found in `outer` and `other`, where more points are concentrated in the bottom left corner. Neither method assigns a large proportion of any admixed individual's genome to the `other` group, but the pedigree-based method is more likely to assign moderate `other` proportions. Furthermore, the genome-based method assigns larger group proportions to `inner` than the pedigree-based method, as more points lie above the diagonal than below. The opposite is true for `outer`. Notably, the two methods are never in strong disagreement, as no points are located in the top left or bottom right corners of the scatter plots.

[1] Hence, the equivalent figure containing the full phenotyped population would look identical except for clear dots in the top right and bottom left corners of each figure.

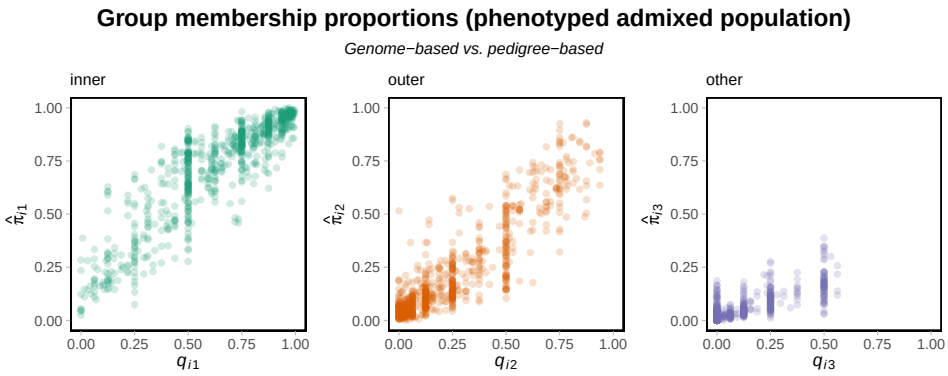


Figure 4.1: Scatter plots for group membership proportion derived from the pedigree (x-axes) and local ancestry inference (y-axes). Each point refers to one individual. The plots only contain points for phenotyped admixed individuals ($N = 788$). Points are partially transparent to show density patterns in areas with overlapping points.

4.2 Group-specific allele frequencies

Our genome-based genetic groups model allows for allele frequencies to differ within the genetic groups. Accordingly, we show how the estimated values of the group-specific allele frequencies are distributed across all combinations of loci m and DNA strand h (Figure 4.2). The alleles originating from the `outer` and `other` groups are likelier to have markers with very small allele frequencies. In `inner`, more markers have allele frequencies close to 0.2, and `inner` also has more alleles with frequency just below 0.5 and fewer alleles with frequency just above. The reason we see more allele frequencies in `outer` and `other` that are larger than 0.5 is that `inner` was used as the basis for which allele is considered the reference allele. Thus, nearly all alternate alleles have frequencies lower than 0.5 in `inner`, but some alternate allele frequencies have fluctuated to over 0.5 in the `outer` and `other` groups (possibly due to genetic drift and small sample effects). Overall, the distributions for the different groups seem to closely follow each other, suggesting there is not much difference in allele frequencies between groups. Moreover, the correlations between the estimated group-specific allele frequency vectors $\boldsymbol{p}_r$ are between 0.8 and 0.9 for all combinations of groups.

4.3 Posterior statistics

We report statistics for the posterior distributions of all fixed and random effects of the six models (Tables 4.1 and 4.2). Since we assume that the fixed effects have normal distributions, their posterior means equal their posterior modes. Thus, only the posterior means are reported for fixed effects. We also report a 95% HPD CI for every parameter. Additionally, the full posterior distributions of the most interesting parameters γ_r and $\sigma^2_{\mathrm{G}_r}$ are displayed graphically (Figures 4.3 and 4.4, respectively). As `inner` serves as the baseline for mean genetic value, γ_1 has no posterior distribution and is instead fixed at zero.

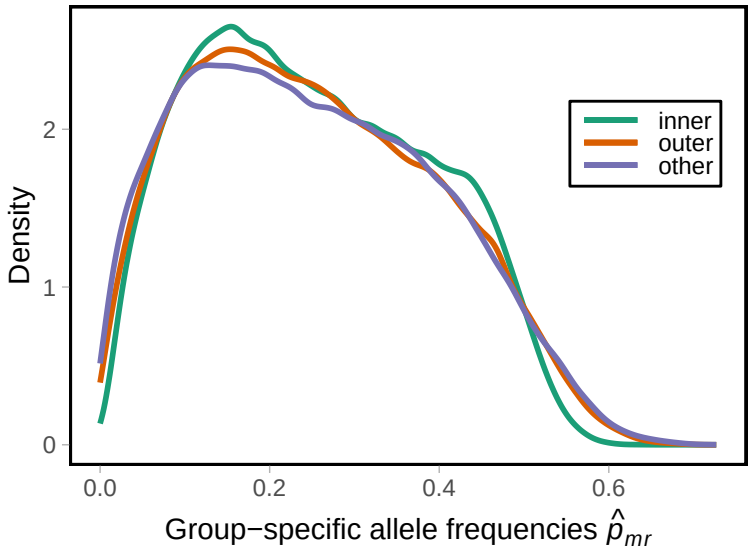


Figure 4.2: Distribution of allele frequencies of the alternate allele within the different genetic groups.

4.3.1 Wing length

For the wing length of house sparrows both the pedigree-based and genome-based models strongly indicate that females have shorter wings (Table 4.1). Conversely, both models find evidence that older birds have longer wings, while measurements made in later summer months generally find shorter wing lengths. Both models also indicate that the values of γ_2 and γ_3 are equally smaller than the reference $\gamma_1 = 0$, though this effect is somewhat more pronounced in the genome-based model (Figure 4.3, top panel). In other words, originating from a non-inner island makes a sparrow somewhat likely to have shorter wings, and the genome-based model implies this situation more strongly. Similarly, both models agree that being inbred likely has a negative impact on wing length, but the pedigree-based model finds this effect to be more distinct. The situation where inbreeding has a negative impact on a phenotype is known as "inbreeding depression" and is not uncommon in wild populations (Crnokrak and Roff 1999).

As for the decomposition of variance, the models agree that year and island of measurement explain little of the overall phenotypic variance in wing length, and that the residual environmental variance is close to 1 (Table 4.2). The permanent environmental variance σ^2_{ID} is found to be larger in the genome-based model than in the pedigree-based model. When it comes to additive genetic variances, both models find notable differences between the genetic groups (Figure 4.4, top panel). The models agree that the variance associated with `outer` is largest, followed by `inner` and then `other`. However, the posteriors for additive genetic variances are shifted towards higher values in the pedigree-based model compared to the genome-based model.

4.3.2 Body mass

In both models for body mass, the posterior fixed effects imply that female or older birds tend to have higher body mass, while inbreeding and later months are negatively associated with with body mass (Table 4.1). Both models also indicate that sparrows originating from `outer` or `other` islands have lower body mass (Figure 4.3, middle panel), but disagree on the magnitude of this difference. The genome-based model shows `outer` having a more negative impact and `other` having a less negative impact than the pedigree-based model does.

Similar to the results for wing length, both models for body mass show little contribution to the overall variance from year and island of measurement, a very similar residual variance and a larger variance in the individual effect in the genome based-model. There is good agreement between the models when it comes to the posteriors of the group-specific additive genetic variances (Table 4.2; Figure 4.4, middle panel), as the posteriors follow each other closely. However, the modes of the `inner` and `outer` variances are smaller in the genome-based model. Both models indicate a higher additive genetic variance in `inner` than in `other`, and an even higher additive genetic variance in `outer`.

Table 4.1: Posterior statistics for the fixed effects of the genetic group animal models. Each column corresponds to one model with a given response (i.e., phenotypic trait; wing length, body mass or tarsus length) and genetic group basis, and each row to a model parameter. For a given effect, the posterior mean is reported in the first row, and a 95% HPD CI is reported in the second row.

	Wing length		Body mass		Tarsus length	
Basis	Genome	Pedigree	Genome	Pedigree	Genome	Pedigree
Sex (f)	-2.77	-2.76	0.48	0.47	-0.08	-0.09
	(-2.90, -2.65)	(-2.89, -2.63)	(0.30, 0.65)	(0.29, 0.64)	(-0.15, -0.01)	(-0.15, -0.02)
F_{GRM}	-1.15	-1.38	-1.16	-1.15	-0.73	-0.77
	(-2.52, 0.23)	(-2.76, -0.00)	(-3.02, 0.70)	(-3.01, 0.71)	(-1.45, -0.01)	(-1.50, -0.04)
Month	-0.19	-0.19	-0.30	-0.30	0.03	0.03
	(-0.22, -0.15)	(-0.22, -0.15)	(-0.35, -0.24)	(-0.36, -0.24)	(0.02, 0.04)	(0.02, 0.04)
Age	0.47	0.47	0.08	0.08	0.00	0.00
	(0.43, 0.50)	(0.43, 0.50)	(0.02, 0.14)	(0.02, 0.14)	(-0.01, 0.01)	(-0.01, 0.01)
γ_2	-0.28	-0.17	-0.58	-0.47	-0.02	-0.01
	(-0.60, 0.05)	(-0.46, 0.14)	(-0.98, -0.17)	(-0.83, -0.11)	(-0.15, 0.11)	(-0.14, 0.11)
γ_3	-0.25	-0.17	-0.22	-0.38	0.07	-0.01
	(-0.70, 0.19)	(-0.49, 0.16)	(-0.77, 0.31)	(-0.83, 0.07)	(-0.17, 0.31)	(-0.19, 0.17)

4.3.3 Tarsus length

Posteriors for the tarsus length models are generally more narrow than for the other phenotypes, so there is less uncertainty in these models. Table 4.1 shows that female sparrows generally have shorter tarsi, while sparrows measured in later months have slightly longer tarsi. On the other hand, there was no evidence that age has an effect on tarsus length. As with all previous responses, inbreeding depression is prevalent for tarsus length, and the pedigree-based model finds the strongest evidence of this. The models mostly find little evidence of different genetic group means (Figure 4.3, bottom panel), except the genomic-based model shows evidence of longer mean tarsus within the `other` group compared to `inner` and `outer`.

In terms of the variances of the random effects in the tarsus models, the posteriors are almost identical between the genome-based and pedigree-based models (Table 4.2). We again find that island and year of measurement explain little phenotypic variance. As with the other phenotypes, the pedigree-based model and genome-based model agree on the residual environmental variance, but tarsus length is the only trait for which the models

Table 4.2: Posterior statistics for the random effect variances of the genetic group animal models. Each column corresponds to one model with a given response (i.e., phenotypic trait; wing length, body mass or tarsus length) and genetic group basis, and each row to a model parameter. For a given variance, the posterior mode and posterior mean (mode;mean) are reported in the first row, and a 95% HPD CI is reported in the second row.

	Wing length		Body mass		Tarsus length	
Basis	Genome	Pedigree	Genome	Pedigree	Genome	Pedigree
$\hat{\sigma}^2_{\text{year}}$	0.05;0.06	0.04;0.04	0.04;0.05	0.04;0.05	0.01;0.02	0.01;0.02
	(0.01, 0.16)	(0.01, 0.11)	(0.01, 0.15)	(0.01, 0.15)	(0.00, 0.04)	(0.00, 0.04)
$\hat{\sigma}^2_{\text{island}}$	0.06;0.08	0.10;0.12	0.10;0.12	0.11;0.13	0.00;0.01	0.00;0.01
	(0.02, 0.22)	(0.03, 0.33)	(0.02, 0.38)	(0.03, 0.39)	(0.00, 0.03)	(0.00, 0.03)
$\hat{\sigma}^2_{\text{ID}}$	0.45;0.46	0.32;0.33	1.13;1.13	1.03;1.05	0.36;0.36	0.36;0.36
	(0.32, 0.64)	(0.20, 0.53)	(0.85, 1.46)	(0.76, 1.42)	(0.32, 0.40)	(0.32, 0.41)
$\hat{\sigma}^2_{\text{G}_1}$	1.59;1.60	1.85;1.86	1.24;1.26	1.46;1.48	0.27;0.27	0.29;0.29
	(1.32, 1.90)	(1.56, 2.22)	(0.91, 1.73)	(1.08, 1.96)	(0.21, 0.35)	(0.22, 0.37)
$\hat{\sigma}^2_{\text{G}_2}$	1.92;1.94	2.27;2.30	1.90;1.96	2.08;2.12	0.14;0.15	0.14;0.14
	(1.39, 2.62)	(1.68, 3.06)	(1.15, 3.08)	(1.27, 3.24)	(0.07, 0.26)	(0.07, 0.26)
$\hat{\sigma}^2_{\text{G}_3}$	1.14;1.20	1.55;1.60	0.69;0.89	0.69;0.83	0.31;0.33	0.31;0.33
	(0.55, 2.16)	(0.84, 2.66)	(0.15, 2.76)	(0.15, 2.36)	(0.14, 0.61)	(0.15, 0.60)
$\hat{\sigma}^2_{\varepsilon}$	0.98;0.98	0.98;0.98	2.86;2.87	2.87;2.88	0.02;0.02	0.02;0.02
	(0.93, 1.04)	(0.93, 1.04)	(2.72, 3.05)	(2.72, 3.04)	(0.02, 0.02)	(0.02, 0.02)

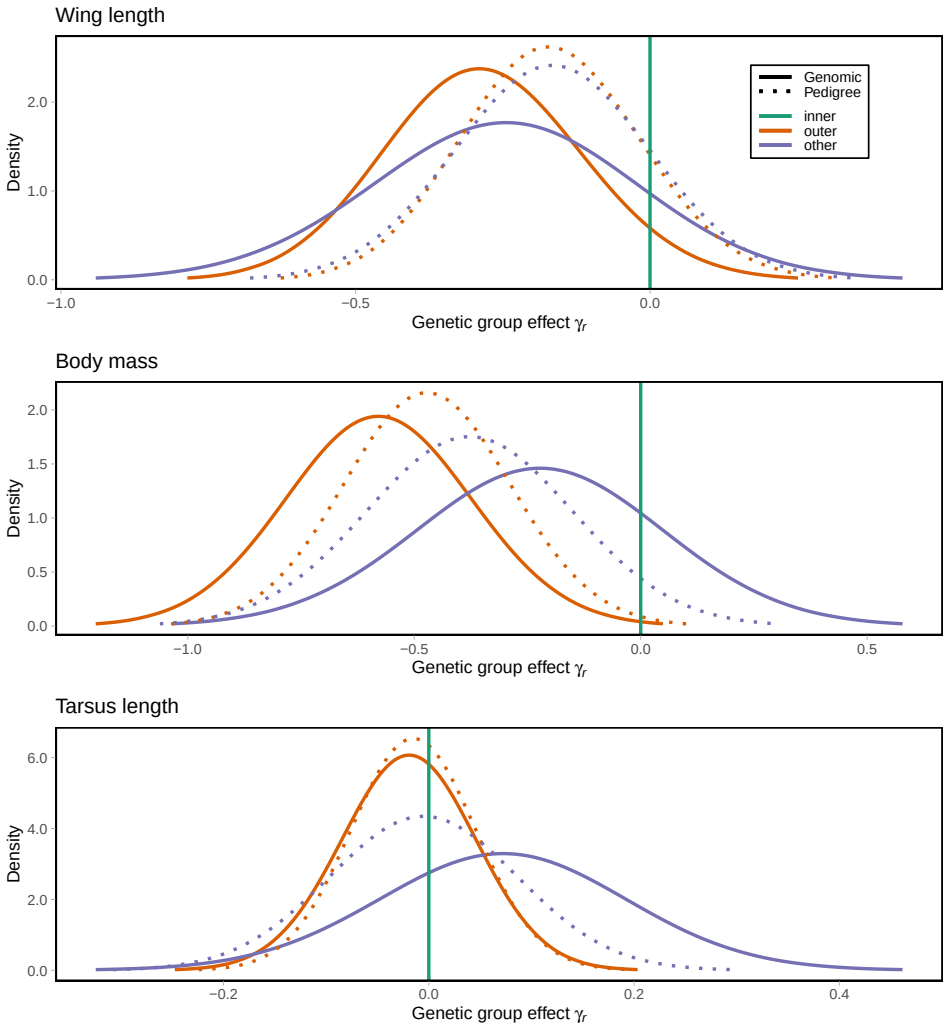


Figure 4.3: The posterior distribution of genetic group effects (i.e., mean genetic value) in the models for wing length (top), body mass (middle) and tarsus length (bottom). Posterior effects for the different genetic groups are shown in different colors. Genome-derived posteriors have solid lines, whereas the pedigree-based posteriors are shown with dotted lines. As `inner` is assumed to be the baseline mean, $\gamma_1 = 0$ is shown as a straight vertical line.

agree on the posterior for σ^2_{ID}. Meanwhile, the two models seem to agree that `inner` and `other` have similar additive genetic variances (with more uncertainty in `other`), which are larger than the additive genetic variances `outer` (Figure 4.4, bottom panel).

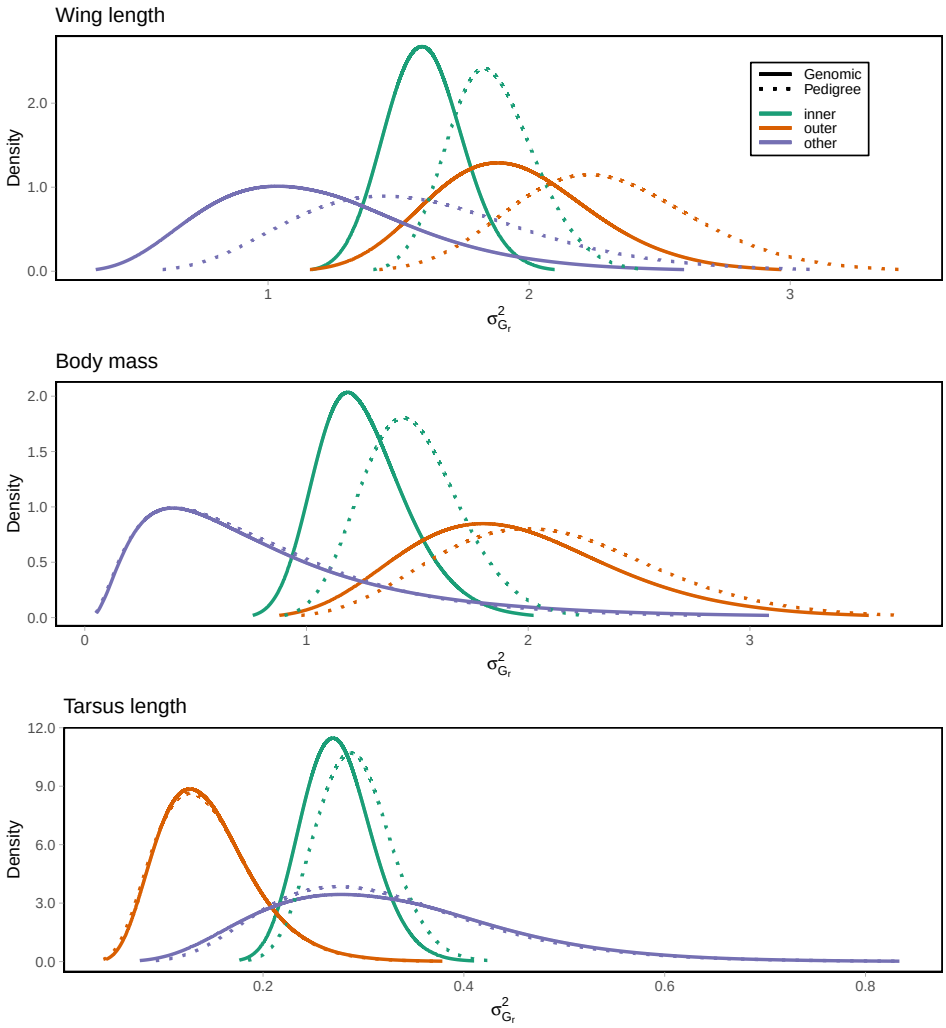


Figure 4.4: The posterior distribution of group-specific additive genetic variances in the models for wing length (top), body mass (middle) and tarsus length (bottom). Posterior variances for the different genetic groups are shown in different colors. Genome-derived posteriors have solid lines, whereas the pedigree-based posteriors are shown with dotted lines.

4.3.4 General findings

Some patterns are clear across the models for each phenotype. We see that the pedigree-based model finds slightly more extreme inbreeding depression than the genome-based model for wing length and tarsus length, although the differences are small with respect to the uncertainties in the estimates. The posteriors for group means indicate sparrows are lighter and have shorter wings relative to their tarsus length on the `outer` and `other`

islands. For wing length and body mass we find that the genome-based models attribute more variation to permanent environmental (i.e. ID) effects, and pedigree-based models find greater additive genetic effects, especially for wing length. There is also a tendency for additive genetic variances to differ between the groups, but the direction of difference varies between traits. Posteriors relating to `other` are usually quite flat, which is expected because we have the least data for this group. The pedigree-based and genome-based models are generally in agreement regarding the remaining model parameters.

Chapter 5

Discussion and conclusion

5.1 Comparison of pedigree-based and genome-based model results

Despite some overall differences in the genetic parameters, the genome-based model finds results that are in relatively good agreement with the pedigree-based results. Since the group-specific additive genetic variances tend to differ between the groups in all the models (regardless of genetic group basis and phenotypic response), the use of a genetic groups model is justified. In this section we will give some hypotheses that might explain the differences we see between the genome-based and pedigree-based model.

Recall that the base population of an animal model using SNP-based kinship is the population used to estimate the allele frequency p_m. However, in the genome-based genetic groups model we use *group-specific* allele frequencies p_{mr} estimated in equation (3.9) from the set of all alleles belonging to group r. This set of alleles includes the full genomes of purebred r-individuals and the r-specific portions of the genomes of admixed individuals. Therefore, the base population of each of the genetic groups would be its respective purebred individuals, in addition to the genome-portions of admixed individuals that are descended from the group. Thus, a genome-based genetic group's base population is not necessarily easily interpretable, as the admixed individuals usually are partial members of different base populations. In contrast, recall that pedigree-based genetic groups have base populations comprised of any phantom parents assigned to that group. Thus, despite the fact that the two types of genetic group models consider the same individuals to be purebred, they have different base populations. In the pedigree-based model only a subset of purebred individuals belong to the group base populations, while in the genome-based model the base populations consist of the purebred individuals plus some portions of admixed genomes. Comparison of the results from the two types of model is therefore not straight-forward, because the posteriors of the genetic parameters pertain to different base populations.

As discussed in Section 2.2.1, Legarra (2016) proposed a scaling which allows us to

explicitly compare genetic additive variance estimates that are based on different relatedness measures. However, the method in Legarra (2016) is not yet well-defined for the genetic groups situation, where admixed individuals have a mix of relatednesses from the different groups. Moreover, the method has not been derived in a Bayesian framework. In a preliminary attempt at a Bayesian group-wise version of the respective scaling, we can transform the posteriors of $\sigma^2_{\mathrm{G}_r}$ and $\sigma^2_{\mathrm{A}_r}$ to both refer to the same base population $\mathcal{B}_r$. We choose this population to consist of all r-purebred individuals and then multiply the posteriors by the appropriate scaling factors from equation (2.4). We then re-normalize to obtain proper posterior distributions. This version of Legarra's scaling is not rigorously derived, but note that the patterns between the genome-based and pedigree-based posteriors (found in Figure 4.4) persist after the group-wise scaling, because the scaling factors we obtain are all very close to 1 (the scaled results are found in Appendix C, Table C.1). Thus, the disagreements we see for group-specific additive genetic variances in wing length are not caused by the difference in base population between the two models.

Since the pedigree-based genetic groups model sometimes (especially for wing length) finds higher additive genetic variances than its genome-based counterpart, it is worth asking what the cause behind this additional variation is. Note that the patterns we see when comparing the group-specific additive genetic variances obtained from the genome-based and pedigree-based models (Figure 4.4) mirror the patterns we see when making a similar comparison between a non-genetic groups animal model simply based on $\mathbf{G}_{\mathrm{VR}}$ and an otherwise similar model based on $\mathbf{A}$ (results not shown). Just like in our genetic group results, the non-genetic groups model based on SNPs finds a smaller additive genetic variance for wing length, a slightly smaller additive genetic variance for body mass and an identical additive genetic variance for tarsus length. The similar pattern indicates that the difference between the models caused by is not caused by some incongruence stemming from our genetic group definitions, but rather inherent differences between the use of pedigrees versus the use of genomic data. Finding greater additive genetic variances in pedigree-based models compared to genome-based models is a well-known phenomenon, and variations of this issue are sometimes known as "missing heritability" problems (Manolio et al. 2009; for a wild animal example, see Bérénos et al. 2014). We will not dive deeply into missing heritabilities in this work, but simply note that our results indicate that the degree to which genetic effects are captured by the genomic data might be smaller in wing length compared to the other phenotypes. For example, the genetic architecture of wing length could be different in that there is less LD between QTL (quantitative trait loci, loci that have an actual impact on phenotype) and SNPs for this trait.

Another question is whether the differences between the pedigree-based and genome-based models relate to non-additive genetic effects. Our models assume an absence of dominance genetic effects, but Wilson et al. (2010, Table 1) points out that animal models can capture such effects, should they be present, in their estimates for additive genetic variances or σ^2_{ID}. Meanwhile, inbreeding depression and dominance effects are intrinsically connected, as explored by Wolak and Keller (2014). In fact, as they point out, a population that displays inbreeding depression must necessarily have a non-zero dominance variance component as part of the genetic variance. Dominance variances are expected to be negligible when allele frequencies are close to 0 or 1 and are only expected to reach magnitudes comparable to additive genetic effects when allele frequencies are close to 0.5 (Hill, God-

dard, and Visscher 2008). Indeed, group-specific allele frequencies in the proximity of 0.5 are not uncommon in the SNP data (Figure 4.2). So, the exact parameters that differ between the pedigree-based models and the genome-based models are the ones that interact with dominance variance effects, namely F_{GRM}, σ^2_{ID} and additive genetic variances. However, there is no clear explanation for why the two models would interact with dominance effects differently, that is, why dominance variance would inflate σ^2_{ID} only in the genome-based model. Exploring this angle further would require estimating dominance effects in each of the models. However, explicitly estimating dominance genetic effects requires large amounts of data in the pedigree-case and an expansion of the mathematical framework in the genome-case. Thus, we will not test these hypotheses here. We will note that neglecting dominance variance has been known to inflate additive genetic variance estimates (eg. Ovaskainen, Cano, and Merilä 2008; Lundregan et al. 2020). Thus, while it is imprudent to give a definitive explanation for the variance patterns, and we should be wary of over-interpreting differences between quite overlapping posteriors, it is worth keeping in mind that the effects the two model approaches disagree on, namely inbreeding depression, σ^2_{ID} (and thus possibly dominance effects) and additive genetic effects, are intricately connected.

5.2 Considerations regarding the genome-based genetic groups model

As previously mentioned, segregation variances were left out of the genome-based model. This omission was justified under the assumption of the infinitesimal model, where all variances $\sigma^2_{\text{S}_{rr'}}$ become zero. However, testing this assumption was not feasible for the data at hand. The inclusion of the segregation variances would involve another three random effect terms, each with a complicated covariance structure. Thus, even heavier stress would be placed on the statistical power of the model, possibly precluding model convergence. Furthermore, the computed $\mathbf{S}_{rr'}$ matrices are not positive semi-definite for this data set, and are thus not proper covariance matrices. Unlike the $\mathbf{G}_r$ matrices, the eigenvalues of the $\mathbf{S}_{rr'}$ are too negative to amend the definiteness issue by adding a small value to the diagonal, so another trick would need to be utilized. Note that the assumption that the segregation variance is zero was checked and found not to be crucial in Muff et al. (2019) for the pedigree-based model.

It might be more realistic to test the assumption $\sigma^2_{\text{S}_{rr'}} = 0$ using a model with only two genetic groups, considering the fact that such a model would have only a single segregation random effect. Additionally, the definition (3.15) of the segregation covariance matrix does not involve any subtractions for $R = 2$, making a positive semi-definite matrix more likely. The obvious candidate for a group to be left out of such a model is `other`. Only 50 phenotyped individuals are considered purebred `other`, and very few alleles in admixed individuals are assigned to have local ancestry originating from this group (Figure 4.1). Furthermore, the posteriors for the `other` genetic parameters are much wider than for the other groups, reducing our confidence in the results for all genetic parameters relating to `other`. An advantage of the genome-based model is that we could simply consider purebred `other` individuals to be admixed, which is not possible in the pedigree-based

approach. Alternatively, the `other` purebreds could be folded into `inner` or `outer` depending on biological considerations. However, we did not estimate any two-group models in this analysis because testing the validity of the extension to $R = 3$ was a higher priority.

The genome-based model allows the allele effects β_{mr} and allele frequencies p_{mr} to depend on the group r for a given m. The former assumption implies that otherwise identical alleles can have different effects depending on which group they are descended from, regardless of their haplotype. The existence of such group-specific allele effects was shown in maize (Rio et al. 2020b). Group-specific allele effects can also result from different levels of LD between SNPs and QTL in the different groups. For instance, the degree of LD between genotyped SNPs and QTL differed among breeds (genetic groups) in cattle (De Roos et al. 2008) and in maize (Technow et al. 2012). As for house sparrows, Hagen et al. (2020) found that levels of LD were generally higher and remained higher over longer distances along the chromosomes on islands in our study system with smaller effective population sizes. Island genetic groups were not explicitly considered by Hagen et al. (2020), but note that `inner` islands generally have larger effective population sizes. The decision to allow group-specific allele effects thus has some justification for this study system. However, allowing allele frequencies to be group-specific seems less necessary in this system. Based on the correlations given in Section 4.2, allele frequencies at specific loci are very similar across the genetic groups. Larger differences in allele frequencies are more likely to be found in populations that are more isolated from each other, with less dispersal between the island groups.

In a wild system such as the one we consider, accurately assigning individuals as purebred or admixed is difficult. In fact, due to the dispersal between the islands, all individuals are likely to be at least somewhat admixed if we look far enough back in time. However, performing local ancestry inference relies on using purebred individuals as references, so we are forced to make a choice of how to partition the population into purebred and admixed populations. Here we simply considered sparrows that are admixed in the pedigree-based model to also be admixed in the genome-based model. This decision has obvious drawbacks because it makes the genome-based model partially rely on pedigree-information. The drawbacks inherent to pedigrees (especially in the wild) might therefore have impacted our choice of purebred individuals in the genome-based model, which is a sacrifice we made for the sake of easy comparison between the models. Luckily, the partition is such that a sizable proportion of phenotyped individuals are admixed (788 out of 1984). Thus, potential differences between the pedigree-based models and the genome-based models should be apparent in the results. An alternative that uses somewhat less pedigree information could be that only founders (rather than purebreds) of the pedigree are eligible to be purebred in the genome-based model. This choice would reduce the impact of potential pedigree-errors on the purebred/admixed partition, and increase the number of admixed individuals we would have to perform local ancestry inference on, based on even fewer reference-genomes. The amount of reference genetic material for `other` would be especially small, which is why we decided against using only founders as purebred in the genome-based model.

In the future it might be useful to develop rigorous guidelines on how to partition the population into purebred and admixed individuals, particularly in the absence of a pedi-

gree. There are several factors to consider, such as what number of purebreds is necessary to obtain accurate local ancestry inferences. Tools such as the one suggested by Kuismin, Ahlinder, and Sillanpää (2017) allows us to detect population structure. Thus, a rigorous approach to choosing purebred and admixed subpopulations might involve designating individuals as purebred based on genetic clusters in the population. Individuals that are not clear members of any cluster can be considered admixed. This "blind" approach is feasible, as it is similar to the analysis performed on the house sparrows by Ranke et al. (2020). However, the blind approach does not take into account informative previous biological knowledge about the system, such as knowledge that the different island-group populations are genetically distinct (Jensen et al. 2013; Niskanen et al. 2020).

A weakness of the genome-based genetic groups model is the long run times of the local ancestry inference. One reason for these run times is the three-way genetic group structure. Indeed, finding a method which reliably works with three genetic groups was challenging when performing the local ancestry inference. For instance, the `R` package `EILA` (Yang et al. 2013) was found to be incapable of handling three groups well and only converged when applied separately on chromosomes with few ($M < 1000$) SNPs. Furthermore, as mentioned in Section 3.3.1, `Loter` disables some quality-control features when $R = 3$. Run times were also severely affected by the particular partition of the genotyped population into purebreds and admixed individuals. A `Loter` run with an alternative group partition where almost all individuals were admixed, for example, had a run time of less than two days: a massive improvement. An improvement in local ancestry inference run times would probably also be seen if we only used a subset of SNP markers in our analysis. We utilized as many SNPs in our analysis as possible, namely all $181\,363$. However, Bérénos et al. (2014, Figure 5) and Rio et al. (2020a, Figure 4) found that their additive genetic variance estimates and predictive ability, respectively, stabilized at numbers of SNPs around $20\,000$. Thus, we might not need all of our SNP markers to obtain reasonable results and could leave some SNPs out to speed up the calculations in the local ancestry inference. However, we would need an investigation into how many SNPs are necessary to obtain accurate results for gametic phasing and local ancestry inference, which would depend on LD-patterns in the populations. Although local ancestry is slow, one upside is that it only needs to be performed once for a given genotype data set and group partition and can then be utilized in many different models.

Despite having developed the theoretical framework for genetic group models with an arbitrary number of groups, we are limited by the capabilities of local ancestry inference software. For models with a large number groups we are also limited by the size of the available data sets, since the number of random effects to be estimated increases with R, and even more so when segregation variances are included. Indeed, we might already be imposing too high a of a demand on the data at hand. A prior sensitivity analysis reveals that our estimates of $\sigma^2_{\mathrm{A}_3}$ and $\sigma^2_{\mathrm{G}_3}$ in particular are not very stable. A much more rigid prior $\mathrm{PC}(0.2, 0.05)$ gave $\sigma^2_{\mathrm{G}_3}$ and $\sigma^2_{\mathrm{A}_3}$ posteriors very close to 0 for all responses, while a less restrictive prior $\mathrm{PC}(3, 0.25)$ gave posteriors shifted to somewhat higher values than in Table 4.2. For the rest of the model parameters the choice of prior was less crucial. Posteriors pertaining to `inner` and `outer` shifted somewhat when the more rigid prior was used, and were barely affected by the less restrictive prior. However, note that using PC priors weighed towards small values for the model variances might actually be undesirable, since

what we want to do is partition the phenotypic variance into various components. Fuglstad et al. (2018) propose a framework for selecting priors to decompose total variance hierarchically, which could be a preferable future alternative to PC priors.

Another step in testing the validity of the genome-based model could involve fitting the model on data from simulated scenarios, as was done with the pedigree-based heterogeneous-variance genetic groups model in Muff et al. (2019). We could then evaluate model performance comprehensively on different genetic architectures, while also having the ability to compare the results to true parameter values.

5.3 Conclusion

Based on the MAGBLUP-RI model introduced by Rio et al. (2020a), we have developed a genome-based genetic groups animal model for wild animal systems. We obtained posterior distributions of genetic parameters in a house sparrow metapopulation using the genome-based model and an equivalent pedigree-based model for different phenotypic responses. Results from the two types of models are generally in agreement, though the genome-based model sometimes finds less inbreeding depression, a larger permanent environmental variance and smaller group-specific additive genetic variances.

Utilizing the genome-based genetic groups model requires some extra steps, including gametic phasing and local ancestry inference. The need for local ancestry inference was computationally very demanding in our case, and the accuracy of this step might be limited in the three-group model. These trade-offs have to be weighed against the potential weaknesses of pedigree-based models. The causes behind the different results from the pedigree-based and genome-based model approaches require further investigation, but we have conceptually introduced genome-based genetic group models in wild populations.

Future work on the genome-based genetic groups model would include a simulation study allowing us to explore the accuracy of model results. We would also like to check our segregation variance assumptions using a 2-group model or a much larger data set. Finally, we would like to explore how the model performs on more and diverse data sets.

Bibliography

1000 Genomes Project Consortium. 2015. “A global reference for human genetic variation.” *Nature* 526 (7571): 68–74.

Al Abri, Mohammed A., et al. 2017. “Application of genomic estimation methods of inbreeding and population structure in an Arabian Horse Herd.” *Journal of Heredity* 108 (4): 361–368.

Beck, Jon A., et al. 2000. “Genealogies of mouse inbred strains.” *Nature Genetics* 24 (1): 23–25.

Bérénos, Camillo, et al. 2014. “Estimating quantitative genetic parameters in wild populations: a comparison of pedigree and genomic approaches.” *Molecular Ecology* 23 (14): 3434–3451.

Browning, Brian L., Ying Zhou, and Sharon R. Browning. 2018. “A one-penny imputed genome from next-generation reference panels.” *The American Journal of Human Genetics* 103 (3): 338–348.

Cantet, R. J. C., and Rohan L. Fernando. 1995. “Prediction of breeding values with additive animal models for crosses from 2 populations.” *Genetics Selection Evolution* 27 (4): 1–12.

Chang, Christopher C., et al. 2015. “Second-generation PLINK: rising to the challenge of larger and richer datasets.” *Gigascience* 4 (1): s13742–015.

Charmantier, Anne, Dany Garant, and Loeske E. B. Kruuk. 2014. *Quantitative genetics in the wild*. Oxford: Oxford University Press.

Charmantier, Anne, et al. 2016. “Mediterranean blue tits as a case study of local adaptation.” *Evolutionary Applications* 9 (1): 135–152.

Chase, Sherret S. 1952. “Production of Homozygous Diploids of Maize from Monoploids 1.” *Agronomy Journal* 44 (5): 263–267.

Cohen, Jacob, et al. 2013. *Applied Multiple Regression/Correlation Analysis for the Behavioral Sciences*. 3rd ed. Abingdon-on-Thames: Routledge.

Conner, Jeffrey K., and Daniel L. Hartl. 2004. *A Primer of Ecological Genetics.* Sunderland: Sinauer Associates Incorporated.

Crnokrak, Peter, and Derek A. Roff. 1999. “Inbreeding depression in the wild.” *Heredity* 83 (3): 260–270.

Crossa, José, et al. 2017. “Genomic selection in plant breeding: methods, models, and perspectives.” *Trends in Plan Science* 22 (11): 961–975.

De Roos, A. P. W., et al. 2008. “Linkage disequilibrium and persistence of phase in Holstein–Friesian, Jersey and Angus cattle.” *Genetics* 179 (3): 1503–1512.

Dias-Alves, Thomas, Julien Mairal, and Michael G. B. Blum. 2019. “Local Ancestry Example.” Visited on 12/15/2020. `https://github.com/bcm-uga/Loter/blob/master/python-package/Local_Ancestry_Example.ipynb`.

— . 2018. “Loter: A software package to infer local ancestry for a wide range of species.” *Molecular Biology and Evolution* 35 (9): 2318–2326.

Edwards, David. 2015. “Two molecular measures of relatedness based on haplotype sharing.” *BMC Bioinformatics* 16 (1): 383.

Excoffier, Laurent, Guillaume Laval, and David J. Balding. 2003. “Gametic phase estimation over large genomic regions using an adaptive window approach.” *Human Genomics* 1 (1): 1–13.

Falconer, D. S., and Trudy F. C. Mackay. 1996. *Introduction to Quantitative Genetics*. 4th ed. Essex: Longman.

Faraway, Julian J. 2016. *Extending the Linear Model with R: Generalized Linear, Mixed Effects and Nonparametric Regression Models*. 2nd ed. Boca Raton: CRC press.

Flanagan, Sarah P., and Adam G. Jones. 2019. “The future of parentage analysis: From microsatellites to SNPs and beyond.” *Molecular Ecology* 28 (3): 544–567.

Fuglstad, Geir-Arne, et al. 2018. “Intuitive joint priors for variance parameters.” *Bayesian Analysis*.

Galwey, Nicholas W. 2014. *Introduction to Mixed Modelling: Beyond Regression and Analysis of Variance*. 3rd ed. Hoboken: John Wiley & Sons.

García-Cortés, Luis Alberto, and Miguel Ángel Toro. 2006. “Multibreed analysis by splitting the breeding values.” *Genetics Selection Evolution* 38 (6): 1–16.

Gelman, Andrew. 2005. “Analysis of variance—why it is more important than ever.” *The Annals of Statistics* 33 (1): 1–53.

Geza, Ephifania, et al. 2019. “A comprehensive survey of models for dissecting local ancestry deconvolution in human genome.” *Briefings in Bioinformatics* 20 (5): 1709–1724.

Gienapp, Phillip, et al. 2017. “Genomic quantitative genetics to study evolution in the wild.” *Trends in Ecology & Evolution* 32 (12): 897–908.

Givens, Goef H., and Jennifer A. Hoeting. 2012. *Computational Statistics*. 2nd ed. Hoboken: John Wiley & Sons.

Grueneberg, Alexander, and Gustavo de los Campos. 2019. “BGData-A Suite of R Packages for Genomic Analysis with Big Data.” *G3: Genes, Genomes, Genetics* 9 (5): 1377–1383.

Hagen, Ingerid Julie, et al. 2020. “A genome-wide linkage map for the house sparrow (Passer domesticus) provides insights into the evolutionary history of the avian genome.” *Molecular Ecology Resources* 20 (2): 544–559.

Hayes, Ben John, Peter M. Visscher, and Michael E. Goddard. 2009. “Increased accuracy of artificial selection by using the realized relationship matrix.” *Genetics Research* 91 (1): 47–60.

He, Yi, and James S Hodges. 2008. “Point estimates for variance-structure parameters in Bayesian analysis of hierarchical models.” *Computational Statistics & Data Analysis* 52 (5): 2560–2577.

Henderson, C. R. 1984. *Applications of Linear Models in Animal Breeding*. Guelph: University of Guelph Press.

Hill, William G., Michael E. Goddard, and Peter M. Visscher. 2008. “Data and theory point to mainly additive genetic variance for complex traits.” *PLoS Genetics* 4 (2): e1000008.

Hill, William G., and Bruce S. Weir. 2011. “Variation in actual relationship as a consequence of Mendelian sampling and linkage.” *Genetics Research* 93 (1): 47–64.

Huisman, Jisca. 2017. “Pedigree reconstruction from SNP data: parentage assignment, sibship clustering and beyond.” *Molecular Ecology Resources* 17 (5): 1009–1024.

Jensen, Henrik, et al. 2008. “Evolutionary dynamics of a sexual ornament in the house sparrow (Passer domesticus): the role of indirect selection within and between sexes.” *Evolution: International Journal of Organic Evolution* 62 (6): 1275–1293.

Jensen, Henrik, et al. 2013. “Genetic variation and structure of house sparrow populations: is there an island effect?” *Molecular Ecology* 22 (7): 1792–1805.

Jones, Adam G., and William R. Ardren. 2003. “Methods of parentage analysis in natural populations.” *Molecular Ecology* 12 (10): 2511–2523.

Keller, Lukas F., et al. 2001. “Heritability of morphological traits in Darwin’s finches: misidentified paternity and maternal effects.” *Heredity* 87 (3): 325–336.

Kruuk, Loeske E. B. 2004. “Estimating genetic parameters in natural populations using the ‘animal model’.” *Philosophical Transactions of the Royal Society of London. Series B: Biological Sciences* 359 (1446): 873–890.

Kruuk, Loeske E. B., and Jarrod D. Hadfield. 2007. “How to separate genetic and environmental causes of similarity between relatives.” *Journal of Evolutionary Biology* 20 (5): 1890–1903.

Kuismin, Markku O., Jon Ahlinder, and Mikko J. Sillanpää. 2017. “CONE: community oriented network estimation is a versatile framework for inferring population structure in large-scale sequencing data.” *G3: Genes, Genomes, Genetics* 7 (10): 3359–3377.

Kuismin, Markku O., et al. 2020. “Genetic assignment of individuals to source populations using network estimation tools.” *Methods in Ecology and Evolution* 11 (2): 333–344.

Legarra, Andres. 2016. “Comparing estimates of genetic variance across different relationship models.” *Theoretical Population Biology* 107:26–30.

Lo, L. L., Rohan L. Fernando, and M. Grossman. 1993. "Covariance between relatives in multibreed populations: additive model." *Theoretical and Applied Genetics* 87 (4): 423–430.

Lundregan, Sarah L., et al. 2018. "Inferences of genetic architecture of bill morphology in house sparrow using a high-density SNP array point to a polygenic basis." *Molecular Ecology* 27 (17): 3498–3514.

Lundregan, Sarah L., et al. 2020. "Resistance to gapeworm parasite has both additive and dominant genetic components in house sparrows, with evolutionary consequences for ability to respond to parasite challenge." *Molecular Ecology*.

Lynch, Michael, and Bruce Walsh. 1998. *Genetics and Analysis of Quantitative Traits*. Vol. 1. Sunderland: Sinauer Associates Incorporated.

Makgahlela, Mahlako Linah, et al. 2013. "Across breed multi-trait random regression genomic predictions in the Nordic Red dairy cattle." *Journal of Animal Breeding and Genetics* 130 (1): 10–19.

Manolio, Teri A., et al. 2009. "Finding the missing heritability of complex diseases." *Nature* 461 (7265): 747–753.

Meuwissen, Theo H. E., Ben John Hayes, and Michael E. Goddard. 2016. "Genomic selection: A paradigm shift in animal breeding." *Animal Frontiers* 6 (1): 6–14.

Mrode, Raphael A. 2014. *Linear Models for the Prediction of Animal Breeding Values*. 3rd ed. Wallingford: Cabi.

Muff, Stefanie, et al. 2019. "Animal models with group-specific additive genetic variances: extending genetic group models." *Genetics Selection Evolution* 51 (1): 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." *Proceedings of the Royal Society B: Biological Sciences* 284 (1850): 20162763.

Niskanen, Alina K., et al. 2020. "Consistent scaling of inbreeding depression in space and time in a house sparrow metapopulation." *Proceedings of the National Academy of Sciences*.

Ovaskainen, Otso, José Manuel Cano, and Juha Merilä. 2008. "A Bayesian framework for comparative quantitative genetics." *Proceedings of the Royal Society B: Biological Sciences* 275 (1635): 669–678.

Padhukasahasram, Badri. 2014. "Inferring ancestry from population genomic data and its applications." *Frontiers in Genetics* 5:204.

Pinheiro, José, and Douglas Bates. 2006. *Mixed-Effects Models in S and S-PLUS*. Berlin: Springer Science & Business Media.

Ponzi, Erica, Lukas F. Keller, and Stefanie Muff. 2019. "The simulation extrapolation technique meets ecology and evolution: A general and intuitive method to account for measurement error." *Methods in Ecology and Evolution* 10 (10): 1734–1748.

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." *Evolution* 72 (10): 1992–2004.

Quaas, R. L. 1988. “Additive genetic model with groups and relationships.” *Journal of Dairy Science* 71 (5): 1338–1345.

Quaas, R. L., and EJ Pollak. 1981. “Modified equations for sire models with groups.” *Journal of Dairy Science* 64 (9): 1868–1872.

Ranke, Peter Sjolte, et al. 2020. “Multi-generational genetic consequences of reinforcement in a bird metapopulation.” *Conservation Genetics*: 1–10.

Reid, Jane M., and Lukas F. Keller. 2010. “Correlated inbreeding among relatives: occurrence, magnitude, and implications.” *Evolution: International Journal of Organic Evolution* 64 (4): 973–985.

Rio, Simon, et al. 2020a. “Accounting for group-specific allele effects and admixture in genomic predictions: theory and experimental evaluation in maize.” *Genetics* 216 (1): 27–41.

Rio, Simon, et al. 2020b. “Disentangling group specific QTL allele effects from genetic background epistasis using admixed individuals in GWAS: an application to maize flowering.” *PLoS Genetics* 16 (3): e1008241.

Rue, Håvard, Sara Martino, and Nicolas Chopin. 2009. “Approximate Bayesian inference for latent Gaussian models by using integrated nested Laplace approximations.” *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 71 (2): 319–392.

Schaeffer, L. R. 1991. “CR Henderson: Contributions to predicting genetic merit.” *Journal of Dairy Science* 74 (11): 4052–4066.

Searle, Shayle R., George Casella, and Charles E McCulloch. 2006. *Variance Components*. 1st ed. Vol. 391. Hoboken: John Wiley & Sons.

Simpson, Daniel, et al. 2017. “Penalising model component complexity: A principled, practical approach to constructing priors.” *Statistical Science* 32 (1): 1–28.

Slatkin, Montgomery, and Russell Lande. 1994. “Segregation variance after hybridization of isolated populations.” *Genetics Research* 64 (1): 51–56.

Speed, Doug, and David J. Balding. 2015. “Relatedness in the post-genomic era: is it still useful?” *Nature Reviews Genetics* 16 (1): 33–44.

Steinsland, Ingelin, and Henrik Jensen. 2010. “Utilizing Gaussian Markov random field properties of Bayesian animal models.” *Biometrics* 66 (3): 763–771.

Strandén, Ismo, and Esa A. Mäntysaari. 2013. “Use of random regression model as an alternative for multibreed relationship matrix.” *Journal of Animal Breeding and Genetics* 130 (1): 4–9.

Sæther, Bernt-Erik, et al. 1999. “Spatial and temporal variation in demography of a house sparrow metapopulation.” *Journal of Animal Ecology* 68 (3): 628–637.

Saatoglu, Dilan, et al. *Dispersal in a house sparrow metapopulation - identifying “cryptic” dispersers using genetic assignment*. Tech. rep.

Technow, Frank, et al. 2012. “Genomic prediction of hybrid performance in maize with models incorporating dominance and population specific marker effects.” *Theoretical and Applied Genetics* 125 (6): 1181–1194.

Thompson, Elizabeth A. 2013. “Identity by descent: variation in meiosis, across genomes, and in populations.” *Genetics* 194 (2): 301–326.

VanRaden, Paul M. 2008. “Efficient methods to compute genomic predictions.” *Journal of Dairy Science* 91 (11): 4414–4423.

Wang, Bowen, Serge Sverdlov, and Elizabeth A. Thompson. 2017. “Efficient estimation of realized kinship from single nucleotide polymorphism genotypes.” *Genetics* 205 (3): 1063–1078.

Weir, Bruce S., Amy D. Anderson, and Amanda B. Hepler. 2006. “Genetic relatedness analysis: modern data and new challenges.” *Nature Reviews Genetics* 7 (10): 771–780.

Weir, Bruce S, and Jérôme Goudet. 2017. “A unified characterization of population structure and relatedness.” *Genetics* 206 (4): 2085–2103.

Wientjes, Yvonne C. J., et al. 2017. “Multi-population genomic relationships for estimating current genetic variances within and genetic correlations between populations.” *Genetics* 207 (2): 503–515.

Wilson, Alastair J., et al. 2010. “An ecologist’s guide to the animal model.” *Journal of Animal Ecology* 79 (1): 13–26.

Wolak, Matthew E. 2012. “nadiv: an R package to create relatedness matrices for estimating non-additive genetic variances in animal models.” *Methods in Ecology and Evolution* 3 (5): 792–796.

Wolak, Matthew E., and Lukas F. Keller. 2014. “Dominance genetic variance and inbreeding in natural populations.” In *Quantitative Genetics in the Wild*, ed. by Anne Charmantier, Dany Garant, and Loeske E. B. Kruuk, 104–127. Oxford: Oxford University Press.

Wolak, Matthew E., and Jane M. Reid. 2017. “Accounting for genetic differences among unknown parents in microevolutionary studies: how to include genetic groups in quantitative genetic animal models.” *Journal of Animal Ecology* 86 (1): 7–20.

— . 2016. “Is pairing with a relative heritable? Estimating female and male genetic contributions to the degree of biparental inbreeding in song sparrows (Melospiza melodia).” *The American Naturalist* 187 (6): 736–752.

Wright, Sewall. 1922. “Coefficients of inbreeding and relationship.” *The American Naturalist* 56 (645): 330–338.

Yang, James J., et al. 2013. “Efficient inference of local ancestry.” *Bioinformatics* 29 (21): 2750–2756.

Yang, Jian, et al. 2011. “GCTA: a tool for genome-wide complex trait analysis.” *The American Journal of Human Genetics* 88 (1): 76–82.

Zuur, Alain, et al. 2009. *Mixed Effects Models and Extensions in Ecology with R*. Berlin: Springer Science & Business Media.

Ødegård, Jørgen, et al. 2018. “Large-scale genomic prediction using singular value decomposition of the genotype matrix.” *Genetics Selection Evolution* 50 (1): 1–12.

Appendix A

Miscellaneous calculations

A.1 Mean genetic value

We calculate the mean genetic value of an individual i as follows.

$$\begin{aligned}\mathsf{E}\left(U_i\right) &= \sum_{r=1}^{R}\sum_{m=1}^{M}\mathsf{E}\left\{\Lambda_{imr}^{(h)}\left[\beta_{mr}^{\text{ref}}+W_{im}^{(h)}\left(\beta_{mr}^{\text{alt}}-\beta_{mr}^{\text{ref}}\right)\right]\right\}\\ &= \sum_{r=1}^{R}\sum_{m=1}^{M}\left[\pi_{ir}\beta_{mr}^{\text{ref}}+\mathsf{E}\left(W_{im}^{(h)}\Lambda_{imr}^{(h)}\right)\left(\beta_{mr}^{\text{alt}}-\beta_{mr}^{\text{ref}}\right)\right]\ .\end{aligned}$$

Recall that $\mathsf{E}\left(W_{im}^{(h)}\Lambda_{imr}^{(h)}\right)=p_{mr}\pi_{ir}$, and so

$$\begin{aligned}\mathsf{E}\left(U_i\right) &= \sum_{r=1}^{R}\sum_{m=1}^{M}\left[\pi_{ir}\beta_{mr}^{\text{ref}}+p_{mr}\pi_{ir}\left(\beta_{mr}^{\text{alt}}-\beta_{mr}^{\text{ref}}\right)\right]\\ &= \sum_{r=1}^{R}\sum_{m=1}^{M}\left[\pi_{ir}\left(\beta_{mr}^{\text{ref}}+p_{mr}\left(\beta_{mr}^{\text{alt}}-\beta_{mr}^{\text{ref}}\right)\right)\right]\\ &= \sum_{r=1}^{R}\sum_{m=1}^{M}\pi_{ir}\gamma_{mr}=\sum_{r=1}^{R}\pi_{ir}\gamma_{r}\ ,\end{aligned}$$

a sum of genetic group effects, weighted by group membership proportions.

A.2 Derivation of equivalent model for genetic value

In this section we show an alternative but equivalent expression of genetic value found using mean-centered versions of $\Lambda_{imr}^{(h)}$ and $W_{im}^{(h)}$. We start with the original definition

(3.1) of genetic value and first insert the mean centered haplotype variable $\widetilde{W}_{im}^{(h)}$;

$$\begin{aligned}
U_i &= \sum_{m=1}^{M}\sum_{r=1}^{R}\sum_{h=1}^{2}\frac{1}{2}\left[\Lambda_{imr}^{(h)}\left(\beta_{mr}^{\text{ref}}+W_{im}^{(h)}\left(\beta_{mr}^{\text{alt}}-\beta_{mr}^{\text{ref}}\right)\right)\right] \\
&= \sum_{m=1}^{M}\sum_{r=1}^{R}\sum_{h=1}^{2}\frac{1}{2}\left[\Lambda_{imr}^{(h)}\left(\beta_{mr}^{\text{ref}}+\left(\frac{\widetilde{W}_{imr}^{(h)}}{\Lambda_{imr}^{(h)}}+p_{mr}\right)\left(\beta_{mr}^{\text{alt}}-\beta_{mr}^{\text{ref}}\right)\right)\right] \\
&= \sum_{m=1}^{M}\sum_{r=1}^{R}\sum_{h=1}^{2}\frac{1}{2}\left[\Lambda_{imr}^{(h)}\beta_{mr}^{\text{ref}}+\left(\widetilde{W}_{imr}^{(h)}+\Lambda_{imr}^{(h)}p_{mr}\right)\left(\beta_{mr}^{\text{alt}}-\beta_{mr}^{\text{ref}}\right)\right] \\
&= \sum_{m=1}^{M}\sum_{r=1}^{R}\sum_{h=1}^{2}\frac{1}{2}\left[\Lambda_{imr}^{(h)}\left(\beta_{mr}^{\text{ref}}+p_{mr}\left(\beta_{mr}^{\text{alt}}-\beta_{mr}^{\text{ref}}\right)\right)+\widetilde{W}_{imr}^{(h)}\left(\beta_{mr}^{\text{alt}}-\beta_{mr}^{\text{ref}}\right)\right] \\
&= \sum_{m=1}^{M}\sum_{r=1}^{R}\sum_{h=1}^{2}\frac{1}{2}\left[\Lambda_{imr}^{(h)}\gamma_{mr}+\widetilde{W}_{imr}^{(h)}\left(\beta_{mr}^{\text{alt}}-\beta_{mr}^{\text{ref}}\right)\right] .
\end{aligned}$$

For the first term in the above brackets, we split out the R term in the sum over genetic groups, as follows

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

Now, since $\Lambda_{imR}^{(h)} = 1-\sum_{r=1}^{R-1}\Lambda_{imR}^{(h)}$ we can write

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

Note that $1-\sum_{r=1}^{R-1}\pi_{ir} = \pi_{iR}$, so

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

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

The term $\sum_{m=1}^{M} \sum_{h=1}^{2} \sum_{r=1}^{R} \pi_{ir} \gamma_{mr}$ equals the mean value of U_i, so we can split it out and finally write

$$U_i = \mathsf{E}\left(U_i\right) + \sum_{m=1}^{M} \sum_{h=1}^{2} \frac{1}{2} \left[\sum_{r=1}^{R-1} \widetilde{\Lambda}_{imr}^{(h)} \left(\gamma_{mr} - \gamma_{mR} \right) + \sum_{r=1}^{R} \widetilde{W}_{imr}^{(h)} \left(\beta_{mr}^{\text{alt}} - \beta_{mr}^{\text{ref}} \right) \right] .$$

A.3 Haplotype covariances

A.3.1 Between-individual, between-locus, within-group

We calculate the covariance between haplotypes of alleles on different loci ($m \neq m'$) in different individuals, so that $i \neq j$. The values of h and h' are irrelevant throughout. First off, because the centered random variables have mean 0, we can say

$$\mathsf{Cov}\left(\widetilde{W}_{imr}^{(h)}, \widetilde{W}_{jm'r}^{(h')} \right) = \mathsf{E}\left(\widetilde{W}_{imr}^{(h)} \widetilde{W}_{jm'r}^{(h')} \right) - \mathsf{E}\left(\widetilde{W}_{imr}^{(h)} \right) \left(\widetilde{W}_{jm'r}^{(h')} \right) = \mathsf{E}\left(\widetilde{W}_{imr}^{(h)} \widetilde{W}_{jm'r}^{(h')} \right) .$$

Using the law of total expectation, and the fact that $\widetilde{W}_{imr}^{(h)} \widetilde{W}_{jm'r}^{(h')} = 0$ if at least one of $\Lambda_{imr}^{(h)}$ or $\Lambda_{jm'r}^{(h')}$ are zero, we can write

$$\begin{aligned}
&\mathsf{Cov}\left(\widetilde{W}_{imr}^{(h)}, \widetilde{W}_{jm'r}^{(h')} \right) \\
&= \mathsf{E}\left(\widetilde{W}_{imr}^{(h)} \widetilde{W}_{jm'r}^{(h')} \,\middle|\, \Lambda_{imr}^{(h)} = 1, \Lambda_{jm'r}^{(h')} = 1 \right) \mathsf{P}\left(\Lambda_{imr}^{(h)} = 1, \Lambda_{jm'r}^{(h')} = 1 \right) \\
&+ \mathsf{E}\left(\widetilde{W}_{imr}^{(h)} \widetilde{W}_{jm'r}^{(h')} \,\middle|\, \Lambda_{imr}^{(h)} = 1, \Lambda_{jm'r}^{(h')} = 0 \right) \mathsf{P}\left(\Lambda_{imr}^{(h)} = 1, \Lambda_{jm'r}^{(h')} = 0 \right) \\
&+ \mathsf{E}\left(\widetilde{W}_{imr}^{(h)} \widetilde{W}_{jm'r}^{(h')} \,\middle|\, \Lambda_{imr}^{(h)} = 0, \Lambda_{jm'r}^{(h')} = 1 \right) \mathsf{P}\left(\Lambda_{imr}^{(h)} = 0, \Lambda_{jm'r}^{(h')} = 1 \right) \\
&+ \mathsf{E}\left(\widetilde{W}_{imr}^{(h)} \widetilde{W}_{jm'r}^{(h')} \,\middle|\, \Lambda_{imr}^{(h)} = 0, \Lambda_{jm'r}^{(h')} = 0 \right) \mathsf{P}\left(\Lambda_{imr}^{(h)} = 0, \Lambda_{jm'r}^{(h')} = 0 \right) \\
&= \mathsf{E}\left(\widetilde{W}_{imr}^{(h)} \widetilde{W}_{jm'r}^{(h')} \,\middle|\, \Lambda_{imr}^{(h)} = 1, \Lambda_{jm'r}^{(h')} = 1 \right) \mathsf{P}\left(\Lambda_{imr}^{(h)} = 1, \Lambda_{jm'r}^{(h')} = 1 \right) .
\end{aligned}$$

We have assumed the LD to be zero, so

$$\begin{aligned}
&\mathsf{Corr}\left(\widetilde{W}_{imr}^{(h)}, \widetilde{W}_{jm'r}^{(h')} \mid \Lambda_{imr}^{(h)} = 1,\ \Lambda_{jm'r}^{(h')} = 1 \right) \\
&= \mathsf{Corr}\left(W_{im}^{(h)}, W_{jm'}^{(h')} \mid \Lambda_{imr}^{(h)} = 1,\ \Lambda_{jm'r}^{(h')} = 1 \right) = 0 ,
\end{aligned}$$

leading to

$$\mathsf{E}\left(\widetilde{W}_{imr}^{(h)} \widetilde{W}_{jm'r}^{(h')} \,\middle|\, \Lambda_{imr}^{(h)} = 1, \Lambda_{jm'r}^{(h')} = 1 \right) = \mathsf{Cov}\left(\widetilde{W}_{imr}^{(h)} \widetilde{W}_{jm'r}^{(h')} \,\middle|\, \Lambda_{imr}^{(h)} = 1, \Lambda_{jm'r}^{(h')} = 1 \right)$$

$$= 0 \times \sqrt{p_{mr}\left(1-p_{mr}\right)p_{m'r}\left(1-p_{m'r}\right)} = 0$$

which implies

$$\mathsf{Cov}\left(\widetilde{W}_{imr}^{(h)}, \widetilde{W}_{jm'r}^{(h')}\right) = 0\,.$$

A.3.2 Within-individual, within-locus, between-group

We calculate $\mathsf{Cov}\left(\widetilde{W}_{imr}^{(h)}, \widetilde{W}_{imr'}^{(h)}\right)$, using the definition of $\widetilde{W}$, as follows

$$\begin{aligned}
\mathsf{Cov}\left(\widetilde{W}_{imr}^{(h)}, \widetilde{W}_{imr'}^{(h)}\right) &= \mathsf{Cov}\left(\Lambda_{imr}^{(h)}\left(W_{im}^{(h)} - p_{mr}\right), \Lambda_{imr'}^{(h')}\left(W_{im}^{(h)} - p_{mr'}\right)\right) \\
&= \mathsf{Cov}\left(\Lambda_{imr}^{(h)}W_{im}^{(h)}, \Lambda_{imr'}^{(h)}W_{im}^{(h)}\right) \\
&- \mathsf{Cov}\left(\Lambda_{imr}^{(h)}W_{im}^{(h)}, \Lambda_{imr'}^{(h)}\right)p_{mr'} \\
&- \mathsf{Cov}\left(\Lambda_{imr}^{(h)}, \Lambda_{imr'}^{(h)}W_{im}^{(h)}\right)p_{mr} \\
&+ \mathsf{Cov}\left(\Lambda_{imr}^{(h)}, \Lambda_{imr'}^{(h)}\right)p_{mr}p_{mr'} \\
&= \mathsf{E}\left(\Lambda_{imr}^{(h)}W_{im}^{(h)}\Lambda_{imr'}^{(h)}W_{im}^{(h)}\right) \\
&- \mathsf{E}\left(\Lambda_{imr}^{(h)}W_{im}^{(h)}\right)\mathsf{E}\left(\Lambda_{imr'}^{(h)}W_{im}^{(h)}\right) \\
&- \left[\mathsf{E}\left(\Lambda_{imr}^{(h)}W_{im}^{(h)}\Lambda_{imr'}^{(h)}\right) - \mathsf{E}\left(\Lambda_{imr}^{(h)}W_{im}^{(h)}\right)\mathsf{E}\left(\Lambda_{imr'}^{(h)}\right)\right]p_{mr'} \\
&- \left[\mathsf{E}\left(\Lambda_{imr}^{(h)}\Lambda_{imr'}^{(h)}W_{im}^{(h)}\right) - \mathsf{E}\left(\Lambda_{imr}^{(h)}\right)\mathsf{E}\left(\Lambda_{imr'}^{(h)}W_{im}^{(h)}\right)\right]p_{mr} \\
&+ \left[\mathsf{E}\left(\Lambda_{imr}^{(h)}\Lambda_{imr'}^{(h)}\right) - \mathsf{E}\left(\Lambda_{imr}^{(h)}\right)\mathsf{E}\left(\Lambda_{imr'}^{(h)}\right)\right]p_{mr}p_{mr'}\,.
\end{aligned}$$

Since at least one of $\Lambda_{imr}^{(h)}$ or $\Lambda_{imr'}^{(h)}$ equal zero, all the expectations above containing a product between these two will also equal zero, so we have

$$\begin{aligned}
\mathsf{Cov}\left(\widetilde{W}_{imr}^{(h)}, \widetilde{W}_{imr'}^{(h)}\right) &= -\mathsf{E}\left(\Lambda_{imr}^{(h)}W_{im}^{(h)}\right)\mathsf{E}\left(\Lambda_{imr'}^{(h)}W_{im}^{(h)}\right) \\
&+ \mathsf{E}\left(\Lambda_{imr}^{(h)}W_{im}^{(h)}\right)\mathsf{E}\left(\Lambda_{imr'}^{(h)}\right)p_{mr'} \\
&+ \mathsf{E}\left(\Lambda_{imr}^{(h)}\right)\mathsf{E}\left(\Lambda_{imr'}^{(h)}W_{im}^{(h)}\right)p_{mr} \\
&- \mathsf{E}\left(\Lambda_{imr}^{(h)}\right)\mathsf{E}\left(\Lambda_{imr'}^{(h)}\right)p_{mr}p_{mr'} \\
&= (-1+1+1-1)\pi_{ir}p_{mr}\pi_{ir'}p_{mr'} \\
&= 0\,.
\end{aligned}$$

A.3.3 Between individual, within-locus, within-group

We calculate the covariance between haplotypes of allele on the same locus m in different individuals, $i \neq j$. By the same logic as in Appendix A.3.1, we obtain

$$\begin{aligned}
&\mathsf{Cov}\left(\widetilde{W}_{imr}^{(h)}, \widetilde{W}_{jmr}^{(h')}\right) \\
&= \mathsf{E}\left(\widetilde{W}_{imr}^{(h)} \widetilde{W}_{jmr}^{(h')} \,\middle|\, \Lambda_{imr}^{(h)} = 1, \Lambda_{jmr}^{(h')} = 1\right) \mathsf{P}\left(\Lambda_{imr}^{(h)} = 1, \Lambda_{jmr}^{(h')} = 1\right) \ .
\end{aligned}$$

Since they are indicator variables that only take the values of 1 and 0, we can observe that $\mathsf{P}\left(\Lambda_{imr}^{(h)} = 1, \Lambda_{jmr}^{(h')} = 1\right) = \mathsf{P}\left(\Lambda_{imr}^{(h)} = 1, \Lambda_{jmr}^{(h)} = 1\right) = \mathsf{E}\left(\Lambda_{imr}^{(h)} \Lambda_{jmr}^{(h)}\right)$, which is also the definition of $\theta_{ij}^{(r)}$. Further note that

$$\begin{aligned}
&\mathsf{Corr}\left(\widetilde{W}_{imr}^{(h)}, \widetilde{W}_{jmr}^{(h')} \,\middle|\, \Lambda_{imr}^{(h)} = 1,\ \Lambda_{jmr}^{(h')} = 1\right) \\
&= \mathsf{Corr}\left(W_{im}^{(h)}, W_{jm}^{(h')} \,\middle|\, \Lambda_{imr}^{(h)} = 1,\ \Lambda_{jmr}^{(h')} = 1\right) = \Gamma_{ij}^{(r)}
\end{aligned}$$

and that

$$\begin{aligned}
&\mathsf{Var}\left(\widetilde{W}_{imr}^{(h)} \,\middle|\, \Lambda_{imr}^{(h)} = 1,\ \Lambda_{jmr}^{(h')} = 1\right) \\
&= \mathsf{Var}\left(W_{im}^{(h)} \,\middle|\, \Lambda_{imr}^{(h)} = 1,\ \Lambda_{jmr}^{(h')} = 1\right) \\
&= p_{mr}\left(1 - p_{mr}\right) \ ,
\end{aligned}$$

with the same result for the conditional variance of $\widetilde{W}_{jmr}^{(h')}$. With this in hand we can write

$$\begin{aligned}
\mathsf{E}\left(\widetilde{W}_{imr}^{(h)} \widetilde{W}_{jmr}^{(h')} \,\middle|\, \Lambda_{imr}^{(h)} = 1, \Lambda_{jmr}^{(h')} = 1\right) &= \mathsf{Cov}\left(\widetilde{W}_{imr}^{(h)} \widetilde{W}_{jmr}^{(h')} \,\middle|\, \Lambda_{imr}^{(h)} = 1, \Lambda_{jmr}^{(h')} = 1\right) \\
&= \Gamma_{ij}^{(r)} p_{mr}\left(1 - p_{mr}\right)
\end{aligned}$$

by the definition of correlation. Then finally, we arrive at

$$\begin{aligned}
&\mathsf{Cov}\left(\widetilde{W}_{imr}^{(h)}, \widetilde{W}_{jmr}^{(h')}\right) \\
&= \mathsf{E}\left(\widetilde{W}_{imr}^{(h)} \widetilde{W}_{jmr}^{(h')} \,\middle|\, \Lambda_{imr}^{(h)} = 1, \Lambda_{jmr}^{(h')} = 1\right) \mathsf{P}\left(\Lambda_{imr}^{(h)} = 1, \Lambda_{jmr}^{(h')} = 1\right) \\
&= \theta_{ij}^{(r)} \Gamma_{ij}^{(r)} p_{mr}\left(1 - p_{mr}\right) \ .
\end{aligned}$$

A.4 Between-individual, between-locus local ancestry covariance

We find the covariance between local ancestries on different loci in different individuals, first within groups and then between groups.

A.4.1 Within-group

The within-group case was already shown in the supplementary material 1 of Rio et al. (2020a) (they do not need the between-group case), but we include it here for completeness, using our own notation. From the definition of $\widetilde{\Lambda}^{(h)}_{imr}$, we can say

$$\begin{aligned}
\mathsf{Cov}\left(\widetilde{\Lambda}^{(h)}_{imr}, \widetilde{\Lambda}^{(h')}_{jm'r}\right) &= \mathsf{Cov}\left(\Lambda^{(h)}_{imr}, \Lambda^{(h')}_{jm'r}\right) \\
&= \mathsf{E}\left(\Lambda^{(h)}_{imr}\Lambda^{(h')}_{jm'r}\right) - \mathsf{E}\left(\Lambda^{(h)}_{imr}\right)\mathsf{E}\left(\Lambda^{(h')}_{jm'r}\right) \\
&= \mathsf{P}\left(\Lambda^{(h)}_{imr} = 1, \Lambda^{(h')}_{jm'r} = 1\right) - \pi_{ir}\pi_{jr} \\
&= \mathsf{P}\left(\Lambda^{(h)}_{imr} = 1 \,\middle|\, \Lambda^{(h')}_{jm'r} = 1\right)\mathsf{P}\left(\Lambda^{(h')}_{jm'r} = 1\right) - \pi_{ir}\pi_{jr} \\
&= \mathsf{P}\left(\Lambda^{(h)}_{imr} = 1 \,\middle|\, \Lambda^{(h')}_{jm'r} = 1\right)\pi_{jr} - \pi_{ir}\pi_{jr} \,. \qquad \text{(A.1)}
\end{aligned}$$

We now consider $\mathsf{P}\left(\Lambda^{(h)}_{imr} = 1 \,\middle|\, \Lambda^{(h')}_{jm'r}\right)$. We use the law of total expectation by conditioning on $\Lambda^{(h')}_{im'r} = 1$, so that

$$\begin{aligned}
&\mathsf{P}\left(\Lambda^{(h)}_{imr} = 1 \,\middle|\, \Lambda^{(h')}_{jm'r} = 1\right) \\
=&\mathsf{P}\left(\Lambda^{(h)}_{imr} = 1 \,\middle|\, \Lambda^{(h')}_{im'r} = 1, \Lambda^{(h')}_{jm'r} = 1\right)\mathsf{P}\left(\Lambda^{(h')}_{im'r} = 1 \,\middle|\, \Lambda^{(h')}_{jm'r} = 1\right) \\
+&\mathsf{P}\left(\Lambda^{(h)}_{imr} = 1 \,\middle|\, \Lambda^{(h')}_{im'r} = 0, \Lambda^{(h')}_{jm'r} = 1\right)\mathsf{P}\left(\Lambda^{(h')}_{im'r} = 0 \,\middle|\, \Lambda^{(h')}_{jm'r} = 1\right) \\
=&\mathsf{P}\left(\Lambda^{(h)}_{imr} = 1 \,\middle|\, \Lambda^{(h')}_{im'r} = 1, \Lambda^{(h')}_{jm'r} = 1\right)\frac{\mathsf{P}\left(\Lambda^{(h')}_{im'r} = 1, \Lambda^{(h')}_{jm'r} = 1\right)}{\mathsf{P}\left(\Lambda^{(h')}_{jm'r} = 1\right)} \\
+&\mathsf{P}\left(\Lambda^{(h)}_{imr} = 1 \,\middle|\, \Lambda^{(h')}_{im'r} = 0, \Lambda^{(h')}_{jm'r} = 1\right)\frac{\mathsf{P}\left(\Lambda^{(h')}_{im'r} = 0, \Lambda^{(h')}_{jm'r} = 1\right)}{\mathsf{P}\left(\Lambda^{(h')}_{jm'r} = 1\right)} \\
=&\mathsf{P}\left(\Lambda^{(h)}_{imr} = 1 \,\middle|\, \Lambda^{(h')}_{im'r} = 1, \Lambda^{(h')}_{jm'r} = 1\right)\frac{\theta^{(r)}_{ij}}{\pi_{jr}} \\
+&\mathsf{P}\left(\Lambda^{(h)}_{imr} = 1 \,\middle|\, \Lambda^{(h')}_{im'r} = 0, \Lambda^{(h')}_{jm'r} = 1\right)\frac{\pi_{jr} - \theta^{(r)}_{ij}}{\pi_{jr}} \,. \qquad \text{(A.2)}
\end{aligned}$$

Consider that $\mathsf{P}\left(\Lambda^{(h)}_{imr} = 1 \,\middle|\, \Lambda^{(h')}_{im'r} = 1, \Lambda^{(h')}_{jm'r} = 1\right)$ will be the group membership proportion of i in group r, on a shrunken set of alleles, since 1 out of the $2M$ alleles are already known to be in group r. Thus, we can say

$$\mathsf{P}\left(\Lambda^{(h)}_{imr} = 1 \,\middle|\, \Lambda^{(h')}_{im'r} = 1, \Lambda^{(h')}_{jm'r} = 1\right) = \frac{\pi_{ir} - \frac{1}{2M}}{1 - \frac{1}{2M}} = \frac{2M\pi_{ir} - 1}{2M - 1} \,. \qquad \text{(A.3)}$$

Similarly, $\mathsf{P}\left(\Lambda_{imr}^{(h)} = 1 \,\middle|\, \Lambda_{im'r}^{(h')} = 0, \Lambda_{jm'r}^{(h')} = 1\right)$ is the group membership on the same shrunken set of alleles, but group r is known not to have been shrunk, so

$$\mathsf{P}\left(\Lambda_{imr}^{(h)} = 1 \,\middle|\, \Lambda_{im'r}^{(h')} = 0, \Lambda_{jm'r}^{(h')} = 1\right) = \frac{\pi_{ir}}{1 - \frac{1}{2M}} = \frac{2M\pi_{ir}}{2M - 1} \,. \tag{A.4}$$

Insert (A.3) and (A.4) into (A.2), which we insert into (A.1) to end up at

$$\begin{aligned}
\mathsf{Cov}\left(\widetilde{\Lambda}_{imr}^{(h)}, \widetilde{\Lambda}_{jm'r}^{(h')}\right) &= \left(\frac{2M\pi_{ir} - 1}{2M - 1} \times \frac{\theta_{ij}^{(r)}}{\pi_{jr}} + \frac{2M\pi_{ir}}{2M - 1} \times \frac{\pi_{jr} - \theta_{ij}^{(r)}}{\pi_{jr}}\right)\pi_{jr} - \pi_{ir}\pi_{jr} \\
&= \frac{\left(2M\pi_{ir} - 1\right)\theta_{ij}^{(r)} + 2M\pi_{ir}\left(\pi_{jr} - \theta_{ij}^{(r)}\right)}{2M - 1} - \pi_{ir}\pi_{jr} \\
&= \frac{2M\pi_{ir}\pi_{jr} - \theta_{ij}^{(r)}}{2M - 1} - \pi_{ir}\pi_{jr} = \frac{\pi_{ir}\pi_{jr} - \theta_{ij}^{(r)}}{2M - 1} \\
&= -\frac{\Delta_{ij}^{(r)}}{2M - 1}, \quad m \neq m' \,.
\end{aligned}$$

A.4.2 Between-group

As for $\mathsf{Cov}\left(\widetilde{\Lambda}_{imr}^{(h)}, \widetilde{\Lambda}_{jm'r'}^{(h')}\right)$ for $m \neq m'$ and $r \neq r'$, we find

$$\mathsf{Cov}\left(\widetilde{\Lambda}_{imr}^{(h)}, \widetilde{\Lambda}_{jm'r'}^{(h')}\right) = \mathsf{P}\left(\Lambda_{imr}^{(h)} = 1 \,\middle|\, \Lambda_{jm'r'}^{(h')} = 1\right)\pi_{jr'} - \pi_{ir}\pi_{jr'} \,, \tag{A.5}$$

similarly to the approach in Appendix A.4.1. When considering $\mathsf{P}\left(\Lambda_{imr}^{(h)} = 1 \,\middle|\, \Lambda_{jm'r'}^{(h')}\right)$, we this time instead condition on $\Lambda_{im'r'}^{(h')}$ in the law of total expectation, so that

$$\begin{aligned}
&\mathsf{P}\left(\Lambda_{imr}^{(h)} = 1 \,\middle|\, \Lambda_{jm'r'}^{(h')} = 1\right) \\
=&\mathsf{P}\left(\Lambda_{imr}^{(h)} = 1 \,\middle|\, \Lambda_{im'r'}^{(h')} = 1, \Lambda_{jm'r'}^{(h')} = 1\right)\frac{\theta_{ij}^{(r')}}{\pi_{jr'}} \\
+&\mathsf{P}\left(\Lambda_{imr}^{(h)} = 1 \,\middle|\, \Lambda_{im'r'}^{(h')} = 0, \Lambda_{jm'r'}^{(h')} = 1\right)\frac{\pi_{jr'} - \theta_{ij}^{(r')}}{\pi_{jr'}} \,.
\end{aligned} \tag{A.6}$$

Again, $\mathsf{P}\left(\Lambda_{imr}^{(h)} = 1 \,\middle|\, \Lambda_{im'r'}^{(h')} = 1, \Lambda_{jm'r'}^{(h')} = 1\right)$ will be the group membership proportion of i in group r, on a shrunken set of alleles, where 1 out of the $2M$ alleles are already known to be in another group r'. Thus,

$$\mathsf{P}\left(\Lambda_{imr}^{(h)} = 1 \,\middle|\, \Lambda_{im'r'}^{(h')} = 1, \Lambda_{jm'r'}^{(h')} = 1\right) = \frac{\pi_{ir}}{1 - \frac{1}{2M}} \,. \tag{A.7}$$

When it comes to $\mathsf{P}\left(\Lambda_{imr}^{(h)} = 1 \,\middle|\, \Lambda_{im'r'}^{(h')} = 0, \Lambda_{jm'r'}^{(h')} = 1\right)$, all we know is that i's loci allele h' at m' is *not* in group r'. Thus, it can either be in group r, or in neither group r or

r'. We weight the results from each of these cases by their respective probability, so

$$\begin{aligned}&\mathsf{P}\left(\Lambda_{imr}^{(h)}=1\,\middle|\,\Lambda_{im'r'}^{(h')}=0,\Lambda_{jm'r'}^{(h')}=1\right)\\&=\frac{\theta_{ij}^{(rr')}}{\pi_{jr'}-\theta_{ij}^{(r')}}\times\frac{\pi_{ir}-\frac{1}{2M}}{1-\frac{1}{2M}}+\frac{\pi_{jr'}-\theta_{ij}^{(r')}-\theta_{ij}^{(rr')}}{\pi_{jr'}-\theta_{ij}^{(r')}}\times\frac{\pi_{ir}}{1-\frac{1}{2M}}\,.\end{aligned}\tag{A.8}$$

We insert (A.7) and (A.8) into (A.6), which we insert into (A.5) to end up at

$$\begin{aligned}&\mathsf{Cov}\left(\widetilde{\Lambda}_{imr}^{(h)},\widetilde{\Lambda}_{jm'r'}^{(h')}\right)\\=&\left[\left(\frac{\theta_{ij}^{(rr')}}{\pi_{jr'}-\theta_{ij}^{(r')}}\times\frac{\pi_{ir}-\frac{1}{2M}}{1-\frac{1}{2M}}+\frac{\pi_{jr'}-\theta_{ij}^{(r')}-\theta_{ij}^{(rr')}}{\pi_{jr'}-\theta_{ij}^{(r')}}\times\frac{\pi_{ir}}{1-\frac{1}{2M}}\right)\times\frac{\pi_{jr'}-\theta_{ij}^{(r')}}{\pi_{jr'}}\right.\\&\left.+\frac{\pi_{ir}}{1-\frac{1}{2M}}\times\frac{\theta_{ij}^{(r')}}{\pi_{jr'}}\right]\pi_{jr'}-\pi_{ir}\pi_{jr'}\\=&\left[\frac{\theta_{ij}^{(rr')}\left(\pi_{ir}-\frac{1}{2M}\right)+\left(\pi_{jr'}-\theta_{ij}^{(r')}-\theta_{ij}^{(rr')}\right)\pi_{ir}+\pi_{ir}\theta_{ij}^{(r')}}{1-\frac{1}{2M}}\right]-\pi_{ir}\pi_{jr'}\\=&\left[\frac{\pi_{ir}\pi_{jr'}-\theta_{ij}^{(rr')}\frac{1}{2M}}{1-\frac{1}{2M}}\right]-\pi_{ir}\pi_{jr'}=\frac{2M\pi_{ir}\pi_{jr'}-\theta_{ij}^{(rr')}-(2M-1)\,\pi_{ir}\pi_{jr'}}{2M-1}\\=&\frac{\pi_{ir}\pi_{jr'}-\theta_{ij}^{(rr')}}{2M-1}=-\frac{\Delta_{ij}^{(rr')}}{2M-1}\,.\end{aligned}$$

A.5 Covariances between haplotypes and local ancestry

We calculate the covariance $\mathsf{Cov}\left(\widetilde{\Lambda}_{imr}^{(h)},\widetilde{W}_{jm'r'}^{(h')}\right)$, using the definition of the centered variables as follows.

$$\begin{aligned}&\mathsf{Cov}\left(\widetilde{\Lambda}_{imr}^{(h)},\widetilde{W}_{jm'r'}^{(h')}\right)\\=&\mathsf{Cov}\left(\Lambda_{imr}^{(h)},\Lambda_{jm'r'}^{(h')}W_{jm'}^{(h')}\right)-\mathsf{Cov}\left(\Lambda_{imr}^{(h)},\Lambda_{jm'r'}^{(h')}\right)p_{m'r'}\\=&\mathsf{E}\left(\Lambda_{imr}^{(h)}\Lambda_{jm'r'}^{(h')}W_{jm'}^{(h')}\right)-\mathsf{E}\left(\Lambda_{imr}^{(h)}\right)\mathsf{E}\left(\Lambda_{jm'r'}^{(h')}W_{jm'}^{(h')}\right)+\frac{\Delta_{ij}^{(rr')}}{2M-1}\times p_{m'r'}\\=&\mathsf{P}\left(\Lambda_{imr}^{(h)}=1,\Lambda_{jm'r'}^{(h')}=1,W_{jm'}^{(h')}=1\right)-\pi_{ir}\pi_{jr}p_{m'r'}+\frac{\Delta_{ij}^{(rr')}}{2M-1}\times p_{m'r'}\,.\end{aligned}$$

Note that

$$\begin{aligned}&\mathsf{P}\left(\Lambda_{imr}^{(h)}=1,\Lambda_{jm'r'}^{(h')}=1,W_{jm'}^{(h')}=1\right)\\&=\mathsf{P}\left(W_{jm'}^{(h')}=1\,\middle|\,\Lambda_{imr}^{(h)}=1,\Lambda_{jm'r'}^{(h')}=1\right)\times\mathsf{P}\left(\Lambda_{imr}^{(h)}=1,\Lambda_{jm'r'}^{(h')}=1\right)\end{aligned}$$

$$
\begin{aligned}
&= p_{m'r'} \times \mathsf{E}\left(\Lambda^{(h)}_{imr}\Lambda^{(h')}_{jm'r'}\right) \\
&= p_{m'r'} \times \left(\mathsf{Cov}\left(\Lambda^{(h)}_{imr}, \Lambda^{(h')}_{jm'r'}\right) + \mathsf{E}\left(\Lambda^{(h)}_{imr}\right)\mathsf{E}\left(\Lambda^{(h')}_{jm'r'}\right)\right) \\
&= p_{m'r'} \times \left(-\frac{\Delta^{(rr')}_{ij}}{2M-1} + \pi_{ir}\pi_{jr}\right) .
\end{aligned}
$$

Thus

$$
\mathsf{Cov}\left(\widetilde{\Lambda}^{(h)}_{imr}, \widetilde{W}^{(h')}_{jm'r'}\right) = 0 .
$$

A.6 Covariance between total genetic values

To recap, the only nonzero covariances are (regardless of the values of h and h')

- $\mathsf{Cov}\left(\widetilde{W}^{(h)}_{imr}, \widetilde{W}^{(h')}_{jmr}\right) = \theta^{(r)}_{ij}\Gamma^{(r)}_{ij}p_{mr}\left(1-p_{mr}\right)$,
- $\mathsf{Cov}\left(\widetilde{\Lambda}^{(h)}_{imr}, \widetilde{\Lambda}^{(h')}_{jmr'}\right) = \Delta^{(rr')}_{ij}$,
- $\mathsf{Cov}\left(\widetilde{\Lambda}^{(h)}_{imr}, \widetilde{\Lambda}^{(h')}_{jm'r}\right) = -\frac{\Delta^{(rr')}_{ij}}{2M-1}, \quad (m \neq m')$,

which we will use to find the covariance between two individuals

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

We note that $\sum_{m=1}^{M}\sum_{m'\neq m}^{M} a_{mm'} = \sum_{m=1}^{M}\sum_{m'=m}^{M} a_{mm'} - \sum_{m=1}^{M} a_{mm}$, so

$$
\begin{aligned}
&\mathsf{Cov}(U_i, U_j \mid \boldsymbol{\pi}_i, \boldsymbol{\pi}_j, \boldsymbol{\theta}_{ij}, \boldsymbol{\Gamma}_{ij}) \\
&= \sum_{r=1}^{R-1}\sum_{r'=1}^{R-1} \Delta_{ij}^{(rr')} \quad \Bigg[\sum_{m=1}^{M} (\gamma_{mr} - \gamma_{mR})(\gamma_{mr'} - \gamma_{mR}) \\
&\qquad\qquad + \frac{1}{2M-1}\sum_{m=1}^{M} (\gamma_{mr} - \gamma_{mR})(\gamma_{mr'} - \gamma_{mR}) \\
&\qquad\qquad - \frac{1}{2M-1}\sum_{m=1}^{M}\sum_{m'=m}^{M} (\gamma_{mr} - \gamma_{mR})(\gamma_{m'r'} - \gamma_{m'R}) \Bigg] \\
&+ \sum_{r=1}^{R} \sigma^2_{\mathrm{G}_r} \theta_{ij}^{(r)} \Gamma_{ij}^{(r)} \\
&= \sum_{r=1}^{R-1}\sum_{r'=1}^{R-1} \Delta_{ij}^{(rr')} \quad \Bigg[\frac{2M}{2M-1}\sum_{m=1}^{M} (\gamma_{mr} - \gamma_{mR})(\gamma_{mr'} - \gamma_{mR}) \\
&\qquad\qquad - \frac{1}{2M-1}\sum_{m=1}^{M}\sum_{m'=m}^{M} (\gamma_{mr} - \gamma_{mR})(\gamma_{m'r'} - \gamma_{m'R}) \Bigg] \\
&+ \sum_{r=1}^{R} \sigma^2_{\mathrm{G}_r} \theta_{ij}^{(r)} \Gamma_{ij}^{(r)} \ .
\end{aligned}
$$

Now note that $\sum_{m=1}^{M}\sum_{m'=1}^{M} a_m b_{m'} = \left(\sum_{m=1}^{M} a_m\right)\left(\sum_{m'=1}^{M} b_{m'}\right)$, so that

$$
\begin{aligned}
&\mathsf{Cov}(U_i, U_j \mid \boldsymbol{\pi}_i, \boldsymbol{\pi}_j, \boldsymbol{\theta}_{ij}, \boldsymbol{\Gamma}_{ij}) \\
&= \sum_{r=1}^{R-1}\sum_{r'=1}^{R-1} \Delta_{ij}^{(rr')} \quad \Bigg[\frac{2M}{2M-1}\sum_{m=1}^{M} (\gamma_{mr} - \gamma_{mR})(\gamma_{mr'} - \gamma_{mR}) \\
&\qquad\qquad - \frac{1}{2M-1}(\gamma_r - \gamma_R)(\gamma_{r'} - \gamma_R) \Bigg] \\
&+ \sum_{r=1}^{R} \sigma^2_{\mathrm{G}_r} \theta_{ij}^{(r)} \Gamma_{ij}^{(r)} \ .
\end{aligned}
$$

Note that the contents of the above brackets can be rewritten as segregation variance terms because

$$
\begin{aligned}
&\sum_{m=1}^{M} (\gamma_{mr} - \gamma_{mR})(\gamma_{mr'} - \gamma_{mR}) \\
&= \sum_{m=1}^{M} \left[\gamma_{mr}\gamma_{mr'} - \gamma_{mr}\gamma_{mR} - \gamma_{mr'}\gamma_{mR} + \gamma_{mR}^2\right]
\end{aligned}
$$

$$
\begin{aligned}
&= \sum_{m=1}^{M} \left[-\frac{1}{2} \left(\gamma_{mr} - \gamma_{mr'} \right)^2 + \frac{1}{2} \gamma_{mr}^2 + \frac{1}{2} \gamma_{mr'}^2 - \gamma_{mr} \gamma_{mR} - \gamma_{mr'} \gamma_{mR} + \gamma_{mR}^2 \right] \\
&= \frac{1}{2} \sum_{m=1}^{M} \left[\left(\gamma_{mr} - \gamma_{mR} \right)^2 + \left(\gamma_{mr'} - \gamma_{mR} \right)^2 - \left(\gamma_{mr} - \gamma_{mr'} \right)^2 \right] ,
\end{aligned}
$$

and similarly

$$
\begin{aligned}
&\left(\gamma_r - \gamma_R \right) \left(\gamma_{r'} - \gamma_R \right) \\
&= \frac{1}{2} \left[\left(\gamma_r - \gamma_R \right)^2 + \left(\gamma_{r'} - \gamma_R \right)^2 - \left(\gamma_r - \gamma_{r'} \right)^2 \right] ,
\end{aligned}
$$

and thus

$$
\begin{aligned}
&\mathsf{Cov}(U_i, U_j \mid \boldsymbol{\pi}_i, \boldsymbol{\pi}_j, \boldsymbol{\theta}_{ij}, \boldsymbol{\Gamma}_{ij}) \\
&= \frac{1}{2} \sum_{r=1}^{R-1} \sum_{r'=1}^{R-1} \Delta_{ij}^{(rr')} \left[\frac{2M}{2M-1} \sum_{m=1}^{M} \left(\gamma_{mr} - \gamma_{mR} \right)^2 - \frac{1}{2M-1} \left(\gamma_r - \gamma_R \right)^2 \right. \\
&\qquad\qquad + \frac{2M}{2M-1} \sum_{m=1}^{M} \left(\gamma_{mr'} - \gamma_{mR} \right)^2 - \frac{1}{2M-1} \left(\gamma_{r'} - \gamma_R \right)^2 \\
&\qquad\qquad \left. - \left(\frac{2M}{2M-1} \sum_{m=1}^{M} \left(\gamma_{mr} - \gamma_{mr'} \right)^2 - \frac{1}{2M-1} \left(\gamma_r - \gamma_{r'} \right)^2 \right) \right] \\
&+ \sum_{r=1}^{R} \sigma_{\mathsf{G}_r}^2 \theta_{ij}^{(r)} \Gamma_{ij}^{(r)} \\
&= \frac{1}{2} \sum_{r=1}^{R-1} \sum_{r'=1}^{R-1} \Delta_{ij}^{(rr')} \left[\sigma_{\mathsf{S}_{rR}}^2 + \sigma_{\mathsf{S}_{r'R}}^2 - \sigma_{\mathsf{S}_{rr'}}^2 \right] + \sum_{r=1}^{R} \sigma_{\mathsf{G}_r}^2 \theta_{ij}^{(r)} \Gamma_{ij}^{(r)} .
\end{aligned}
$$

Now let $\mathcal{R}$ be the set $\{1, \ldots, R\}$. We note that per definition $\sigma_{\mathsf{S}_{rr}}^2 = 0$ and $\sigma_{\mathsf{S}_{rr'}}^2 = \sigma_{\mathsf{S}_{r'r}}^2$, so we can simplify to

$$
\begin{aligned}
&\mathsf{Cov}(U_i, U_j \mid \boldsymbol{\pi}_i, \boldsymbol{\pi}_j, \boldsymbol{\theta}_{ij}, \boldsymbol{\Gamma}_{ij}) \\
&= \sum_{r=1}^{R-1} \Delta_{ij}^{(r)} \sigma_{\mathsf{S}_{rR}}^2 + \frac{1}{2} \sum_{r=1}^{R-2} \sum_{r'=r+1}^{R-1} \left(\Delta_{ij}^{(rr')} + \Delta_{ij}^{(r'r)} \right) \left[\sigma_{\mathsf{S}_{rR}}^2 + \sigma_{\mathsf{S}_{r'R}}^2 - \sigma_{\mathsf{S}_{rr'}}^2 \right] \\
&+ \sum_{r=1}^{R} \sigma_{\mathsf{G}_r}^2 \theta_{ij}^{(r)} \Gamma_{ij}^{(r)} \\
&= \sum_{r=1}^{R-1} \left(\Delta_{ij}^{(r)} + \frac{1}{2} \sum_{r' \neq r}^{R-1} \left(\Delta_{ij}^{(rr')} + \Delta_{ij}^{(r'r)} \right) \right) \sigma_{\mathsf{S}_{rR}}^2 - \frac{1}{2} \sum_{r=1}^{R-2} \sum_{r'=r+1}^{R-1} \left(\Delta_{ij}^{(rr')} + \Delta_{ij}^{(r'r)} \right) \sigma_{\mathsf{S}_{rr'}}^2 \\
&+ \sum_{r=1}^{R} \sigma_{\mathsf{G}_r}^2 \theta_{ij}^{(r)} \Gamma_{ij}^{(r)} .
\end{aligned}
$$

Since $\sum_{r=1} \Lambda_{imr}^{(h)} = 1$, we can find the identity

$$\Delta_{ij}^{(r)} = \sum_{\mathcal{R}\setminus\{r\}} \Delta_{ij}^{(r')} + \sum_{r',r''\in\mathcal{R}\setminus\{r\}} \Delta_{ij}^{(rr')} , \tag{A.9}$$

which we can use to rewrite the coefficient

$$\Delta_{ij}^{(r)} + \frac{1}{2}\sum_{r'\neq r}^{R-1} \left(\Delta_{ij}^{(rr')} + \Delta_{ij}^{(r'r)}\right) = \frac{1}{2}\left(\Delta_{ij}^{(r)} + \Delta_{ij}^{(R)} - \sum_{r',r''\in\mathcal{R}\setminus\{r,R\}} \Delta_{ij}^{(r'r'')}\right) .$$

Again using eq. (A.9) and the similar identities

$$\Delta_{ij}^{(r)} = -\sum_{r'\neq r}^{R} \Delta_{ij}^{(rr')}, \quad \Delta_{ij}^{(r)} = -\sum_{r'\neq r}^{R} \Delta_{ij}^{(r'r)} ,$$

we can also rewrite the other coefficient

$$-\frac{1}{2}\left(\Delta_{ij}^{(rr')} + \Delta_{ij}^{(r'r)}\right) = \frac{1}{2}\left(\Delta_{ij}^{(r)} + \Delta_{ij}^{(r')} - \sum_{r'',r^*\in\mathcal{R}\setminus\{r,r'\}} \Delta_{ij}^{r''r^*}\right) .$$

So, finally, we can write

$$\begin{aligned}
&\mathsf{Cov}(U_i, U_j \mid \boldsymbol{\pi}_i, \boldsymbol{\pi}_j, \boldsymbol{\theta}_{ij}, \boldsymbol{\Gamma}_{ij}) \\
&= \frac{1}{2}\sum_{r=1}^{R-1}\left(\Delta_{ij}^{(r)} + \Delta_{ij}^{(R)} - \sum_{r',r''\in\mathcal{R}\setminus\{r,R\}} \Delta_{ij}^{(r'r'')}\right)\sigma^2_{\mathsf{S}_{rR}} \\
&+ \frac{1}{2}\sum_{r=1}^{R-2}\sum_{r'=r+1}^{R-1}\left(\Delta_{ij}^{(r)} + \Delta_{ij}^{(r')} - \sum_{r'',r^*\in\mathcal{R}\setminus\{r,r'\}} \Delta_{ij}^{r''r^*}\right)\sigma^2_{\mathsf{S}_{rr'}} \\
&+ \sum_{r=1}^{R}\sigma^2_{\mathsf{G}_r}\theta_{ij}^{(r)}\Gamma_{ij}^{(r)} \\
&= \frac{1}{2}\sum_{r=1}^{R-1}\sum_{r'=r+1}^{R}\left(\Delta_{ij}^{(r)} + \Delta_{ij}^{(r')} - \sum_{r'',r^*\in\mathcal{R}\setminus\{r,r'\}} \Delta_{ij}^{r''r^*}\right)\sigma^2_{\mathsf{S}_{rr'}} \\
&+ \sum_{r=1}^{R}\sigma^2_{\mathsf{G}_r}\theta_{ij}^{(r)}\Gamma_{ij}^{(r)} .
\end{aligned}$$

Appendix B

R code and calls to other software

This appendix contains all R code and all calls to other software that were used to generate the model results. All `R` code can be found in the `Git` repositoyory located at `https://github.com/kennaas/GGG`.

B.1 Pedigree-based kinship matrices

The `R` file `data_and_GG_setup.R` is based on code used in Muff et al. (2019), and prepares the phenotypic and pedigree data for use. We compute the genetic relatedness matrix $\mathbf{A}$ and compute the matrix $\mathbf{Q}$ containing as entries expected group membership proportions q_{ir} in each group for all individuals in the pedigree. Using $\mathbf{A}$ and $\mathbf{Q}$ we compute group-specific relatedness matrices $\mathbf{A}_r$ in the file `GG_A.R` as described in section 2.2.3.

B.2 Genome-based kinship matrices

The partition of the genotyped individuals into three reference populations and an admixed population is preformed in the `R` file `ref_adm_partition.R`. We generate `.txt` files containing all genotyped individuals *not* contained in each of these populations, which we use in the gametic phasing.

B.2.1 Gametic phasing

The genotype data is available on the `PLINK 1.9` genomic data file format `.ped` with an accompanying `.map` file. The `.ped` file contains counts of the alternate allele at each SNP for every individual, while the `.map` contains additional information such as the chromosome each SNP is located on. `Beagle 5.1`, the software we use to phase the data (Browning, Zhou, and Browning 2018), takes `.vcf` files as input, so we use `PLINK`

to convert the data to this format.

```
plink.exe --ped genotypes.ped --map genotypes.map
          --chr-set 32 --recode vcf-iid --out vcf_genotypes
```

The following `Beagle` commands were used to phased/impute the data. Inner:

```
java -jar beagle.18May20.d20.jar gt=vcf_genotypes.vcf
     excludemarkers=no_chrom_SNPs.txt
     excludesamples=NOT_inner_inds.txt out=inner_phased
```

Outer:

```
java -jar beagle.18May20.d20.jar gt=vcf_genotypes.vcf
     excludemarkers=no_chrom_SNPs.txt
     excludesamples=NOT_outer_inds.txt out=outer_phased
```

Other:

```
java -jar beagle.18May20.d20.jar gt=vcf_genotypes.vcf
     excludemarkers=no_chrom_SNPs.txt
     excludesamples=NOT_other_inds.txt out=other_phased
```

Admixed:

```
java -jar beagle.18May20.d20.jar gt=vcf_genotypes.vcf
     excludemarkers=no_chrom_SNPs.txt
     excludesamples=NOT_admixed_inds.txt out=admixed_phased
```

These calls to `Beagle` produces imputed haplotype data for each of the populations, on the `.vcf` format. Note that we specific why individuals to exclude from each imputation/phasing, as well as which SNPs, namely the SNPs not placed on specific chromosomes.

In the `R` file `W_setup.R` the imputed/phased haplotype data for each reference and admixed population are merged into a single file-backed haplotype matrix $\mathbf{W}$. Rows of $\mathbf{W}$ correspond to individuals, while columns correspond to alleles. Alleles are ordered so that odd-numbered columns have alleles with $h = 1$, and even-numbered columns have $h = 2$. In other words, the first few entries in the i^{th} row of $\mathbf{W}$ are $w_{i1}^{(1)}$, $w_{i1}^{(2)}$, $w_{i2}^{(1)}$, $w_{i2}^{(2)}$, $w_{i3}^{(1)}$, $w_{i3}^{(2)}$, etc. We use the functions contained in the utility-file `file_backed_mat.R` for the initializing of file-backed matrices.

B.2.2 Local ancestry inference

After the gametic phasing we have haplotype available on the `.vcf` format, separately for each of the three reference populations and the admixed population. We run local ancestry

inference on the admixed population using the command-line version of `Loter` with the following command:

```
loter_cli -f 'vcf' -r outer_phased.vcf.gz inner_phased.vcf.gz
          other_phased.vcf.gz -a admixed_phased.vcf -n 8
          -o loter_out.txt -v
```

The use of the `-f` option makes `Loter` accept the haplotype data on the `.vcf` format, while the `-r` option lists the reference populations, the `-a` option lists the admixed population, the `-n` option lists the number of cores to be used, `-o` names the output file and `-v` tells `Loter` to use the verbose option, that is, it outputs more information while running.

The local ancestry data is outputted on a `.txt` format where entries 0, 1 and 2 indicate group membership in `outer`, `inner` and `other`, respectively. In the `R` file `loter_result_conversion.R` we convert this data to three separate local ancestry matrices $\mathbf{\Lambda}_r$ whose entries refer to the same alleles as the entries of $\mathbf{W}$. The entries in $\mathbf{\Lambda}_r$ are 0 or 1, indicating whether or not the allele has membership in group r.

B.2.3 Construction of genome-based relatedness matrices

To implement equation (3.13), we take several steps. Group-specific allele frequencies are computed in the file `Group-specific_allele_freq.R`. The numerator of (3.13) is computed as a sum of four matrix product in the file `gamma_numerator.R`, which relies on the matrices computed in `V_matrix.R`. Similarly, the denominator of (3.13) is computed as a sum of four matrix product in the file `gamma_denominator.R`, which relies on the matrices computed in `L_matrix.R`. The estimators for $\theta_{ij}^{(r)}$ are computed in `theta.R`. Using all these results, in the file `GG_GRM_setup.R` we find the final group specific GRMs $\mathbf{G}_r$ for each group r, as well as the estimated group membership proportion vectors $\boldsymbol{\pi}_r$ and the segregation covariance matrix components $\mathbf{\Delta}^{(r)}$.

B.3 INLA model

The model is fit using `R-INLA` in the file `GG_Animal_Model.R`, which when ran from command line has options for whether to use pedigree-based or genome-based genetic groups, which phenotypic response to use, and more. The model results are saved and can be used to generate figures and tables of posterior distributions of the parameters. Legarra-scaling of the additive genetic variance posteriors is performed in `Legarra.R`.

Appendix C

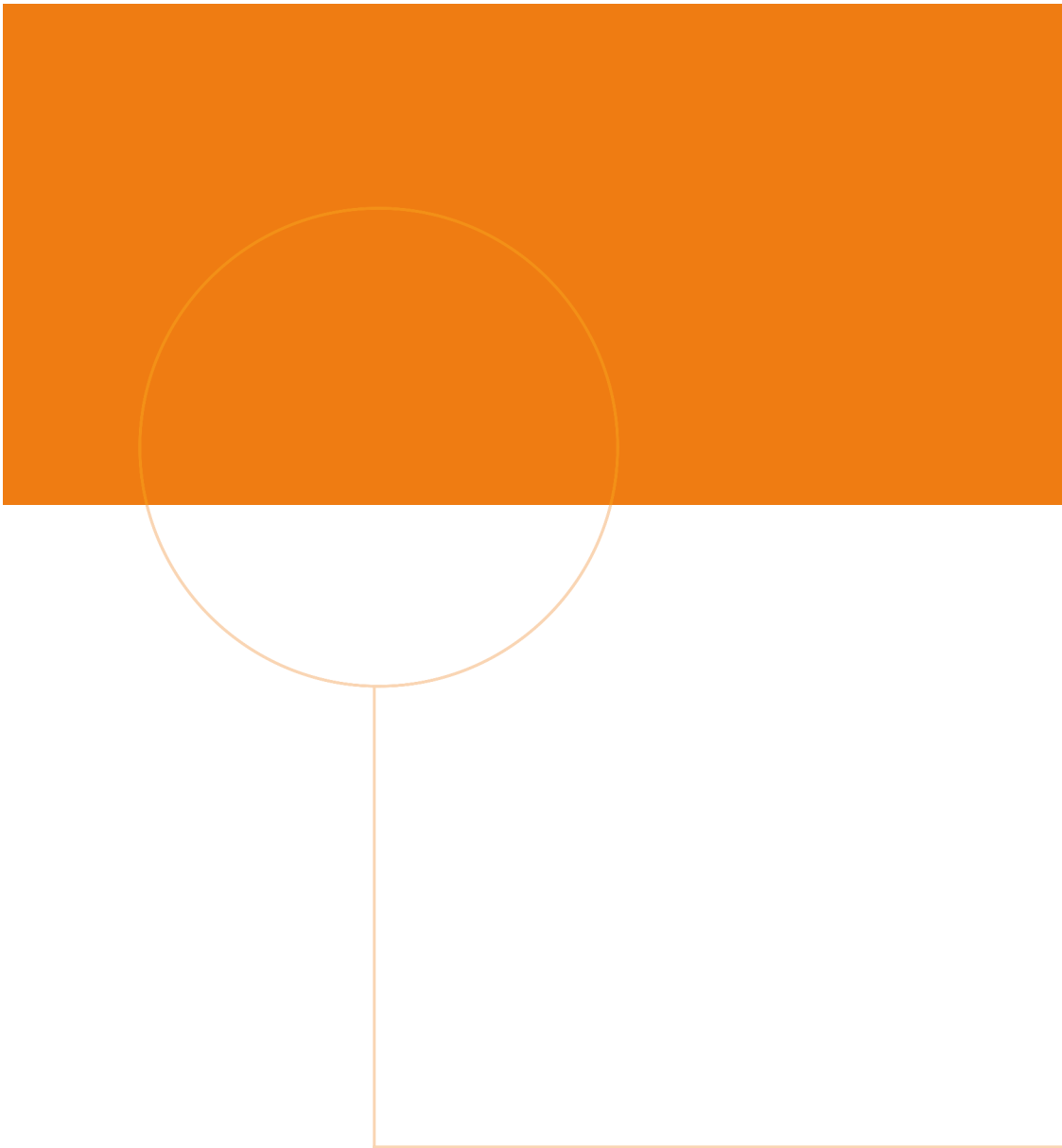
Legarra-scaled additive genetic variances

Table C.1 shows the group-specific additive genetic variances, scaled to refer to the same base populations: $\mathcal{B}_1$, $\mathcal{B}_2$ and $\mathcal{B}_3$ for `inner`, `outer` and `other`, respectively. The base population $\mathcal{B}_r$ contains the individuals that are purebred in group r.

Table C.1: Posterior statistics for the Legarra-scaled group-specific additive genetic variances. Each column corresponds to one model with a given response and genetic group basis, and each row to a model parameter. For a given base population, we report the posterior mode and posterior mean (mode;mean) in the first row, and a 95% HPD CI in the second row.

Legarra-scaled group-specific additive genetic variances

	Wing length		Body mass		Tarsus length	
Basis	Genome	Pedigree	Genome	Pedigree	Genome	Pedigree
$\hat{\sigma}^2_{\mathcal{B}_1}$	1.65;1.65	1.87;1.88	1.28;1.30	1.48;1.49	0.28;0.28	0.29;0.29
	(1.36, 1.96)	(1.57, 2.24)	(0.95, 1.79)	(1.09, 1.98)	(0.22, 0.36)	(0.22, 0.37)
$\hat{\sigma}^2_{\mathcal{B}_2}$	1.98;2.00	2.25;2.28	1.96;2.01	2.06;2.10	0.14;0.15	0.14;0.14
	(1.43, 2.70)	(1.66, 3.03)	(1.18, 3.17)	(1.25, 3.21)	(0.07, 0.27)	(0.07, 0.25)
$\hat{\sigma}^2_{\mathcal{B}_3}$	1.17;1.22	1.51;1.56	0.71;0.91	0.67;0.81	0.32;0.33	0.30;0.32
	(0.56, 2.21)	(0.83, 2.61)	(0.15, 2.82)	(0.14, 2.31)	(0.14, 0.62)	(0.15, 0.59)



NTNU

Kunnskap for en bedre verden