Two-phase subsampling for DNA sequencing with application to endangered species

Pei Luo

Supervisor: Prof. Thomas Lumley
Dr. Ben Stevenson

Department of Statistics
The University of Auckland

A thesis submitted in partial fulfilment of the requirements for the degree of Doctor of Philosophy in Statistics, the University of Auckland, 2024.

February 2024

Abstract

Whole-genome sequencing for New Zealand endangered parrot species kākāpō has been completed for the entire population. Despite the decreasing cost of DNA sequencing, this sort of effort is generally not feasible in conservation studies or large human cohorts. A cost-saving strategy is to obtain relatively inexpensive information for the whole sample, such as low-resolution genotype data, then resequence a small subsample from the original sample with higher resolution and use the combined data to infer the whole sample. Such sampling strategies are called two-phase sampling, where the initial sampling of the cohort is followed by a subsampling of the chosen individuals to be resequenced.

This thesis explores the two classes of approaches to handling incomplete data in two-phasing sampling designs under different situations. The first class of approaches is genotype imputation, which is a process of predicting the missing genotypes using low-resolution genotypes of the whole sample and high-resolution genotypes of the subsample. However, genotype imputation is much more complicated for endangered species than for well-studied species such as humans, livestock and other model organisms.

Alternatively, statistical inference of model parameters under two-phase sampling designs can be carried out by maximum likelihood approaches that account for the missing mechanisms of the data, which is another class of approaches that I explore. In genetic association studies, the polygenic model is often used to describe the architecture of complex traits as it allows the possibility that thousands of variants could contribute to the phenotypic variation in the population. Under such a proposition, mixed models can be used to measure the genetic effect of a particular variant while attributing the remaining variation to the population correlation structure.

In this thesis, I propose a weighted maximum likelihood approach for fitting mixed models that takes advantage of the fact that the kākāpō population relatedness structure is known, making it possible to incorporate the population covariance matrix rather than the sample covariance matrix into the model. The performance of the proposed method is evaluated using the kākāpō data and simulated data with a population structure similar to humans. Hence the method should provide a general solution for fitting mixed models under two-phase sampling designs in both endangered species and human populations.

Acknowledgements

First and foremost I would like to express my deepest appreciation to my supervisors, Prof. Thomas Lumley and Dr. Ben Stevenson for providing excellent supervision, their contribution, and patience during my PhD journey.

I would like to thank the kākāpō 125+ project for generating the modern genomics data and the New Zealand Department of Conservation (DOC) and Ngai Tahu for granting access to them. I also wish to acknowledge the use of New Zealand eScience Infrastructure (NeSI) high-performance computing facilities. A special thank you goes to the NeSI support team for their IT help and advice.

Lastly, I would like to express my gratitude to my parents and my friend for their continuous support, understanding and encouragement over the past few years, it would not be possible for me to complete my study without them.

Table of contents

List of figures

List of tables

Glossary

Genetic linkage	Genetic linkage is the phenomenon where genes are linked (i.e. physically close to each other along a chromosome), and hence they are likely to be inherited together.
Genetic marker	A genetic marker is a DNA sequence such that its location on the chromosome is known (e.g. SNPs).
Genome-wide association study	A genome-wide association study (GWAS or GWA study) is an approach to find the association between genetic variants and a particular trait by scanning SNPs across the genome in different individuals.
Genotype	A genotype is the genetic identity of an allele that is determined by the makeup of the allele.
Genotype imputation	Genotype imputation is the process of predicting the unobserved genotypes.
Haplotype	A haplotype is a combination of alleles that were inherited together from a single parent.
Heterogeneity	Heterogeneity refers to the substructure within a population that may be caused by subpopulations with different ethnic ancestries, different environments or different disease-related genetic factors.
Heterozygous	A diploid organism is heterozygous at a locus if the alleles are different from one another.
Identical by descent	An IBS segment is identical by descent (IBD) if the individuals inherited the DNA segment from the same ancestor.
Identical by state	A DNA segment is called identical by state (IBS) in two or more individuals if the nucleotide sequences in the DNA segment are identical between these individuals.

Inheritance vector An inheritance vector of a pedigree at a locus is a vector containing meiosis indicators of all non-founders that represents the inheritance pattern of the pedigree at the locus.

Linkage analysis Linkage analysis is a family-based approach that searches for genetic markers that cosegregate with a particular phenotype through families.

Linkage disequilibrium Linkage disequilibrium (LD) is the phenomenon where alleles at different loci are non-randomly associated in such a way that they are inherited together more frequently than expected if they were independent and randomly associated.

Maximum clique A clique is a subgraph that has all of its nodes connected to each other. A maximum clique is the largest clique that is not part of other cliques.

Meiosis indicator A meiosis indicator is a vector containing two binary number that indicate the pattern of allele transmission at this particular locus. The first (resp. second) binary number represents the allele transmission from the individual's mother (resp. father), and 0 (resp. 1) indicates the maternal (resp. paternal) copy of the allele is transmitted.
2. Law of independent assortment: Genes of different traits are segregated independently of one another during the formation of gametes.
3. Law of dominance: An organism with at least one dominant allele will display the effect of the dominant allele.

Pedigree A pedigree is a graph that illustrates the biological relationships between an organism and its ancestors.

Phase Phasing refers to the process of assigning alleles to the paternal and maternal chromosomes. A resulting pair of allele combinations on maternal and paternal chromosomes is called a phase.

Phenotype A phenotype is an observable characteristic or trait of an organism.

Recombination Recombination is a special form of genetic exchange such that two genetic sequences are combined into one sequence.

Single nucleotide polymorphism A single nucleotide polymorphism (SNP) is a variation at a single DNA building block, namely nucleotide.

Chapter 1

Introduction

The kākāpō, *Strigops habroptilus*, is a critically endangered species in New Zealand, and it is the world's largest, the only flightless, and the only lek-breeding parrot. Whole genome sequence data has been obtained for the entire kākāpō population. The DNA sequences data allows the kākāpō recovery team to perform numerous analyses of the kākāpō species providing insights into genetic management, disease, fertility and ageing [44, 53]. One of the major goals in the kākāpō conservation project is to find functional genetic variants that are associated with key traits using genome-wide association studies (GWA studies, or GWASs). A GWAS is a process of finding functional variants by testing hundreds of thousands of genetic variants across the genome in different individuals for association with the trait. In many GWA studies, the polygenic model is considered to be the founding principle as it allows the possibility that thousands of variants could contribute to the phenotypic variation in the population.

Under the polygenic model [51, 150], one can fit mixed models to measure the genetic effect of a particular variant while accounting the other variants as correlations between related individuals. Mixed-effects models have been widely used not only to carry out GWA studies but also to estimate heritability [8, 62, 163, 164, 166].

Since heritable genetic variation is a necessary condition for evolution by both natural and artificial selection, heritability is a critical concept in conservation genetics that measures the amount of phenotypic variation in a population caused by genetic factors relative to the total phenotypic variation due to both genetic and environmental factors. Thus the mixed-effects models play an important role in making informed decisions for genetic management of endangered species [6, 12, 49, 64, 79, 127].

However, sequencing the entire genome of every individual in a population is not cost-effective in general and GWA studies typically find associations with nearby genetic variants rather than the functional genetic variants. A cost-saving strategy is to do two-phase sampling,

that is, genotyping all individuals in the large sample in low resolution, fully sequencing a small subsample, and then using the combined data to impute the missing genotypes of the rest of the sample. Chapter 2 provides the background knowledge of the work in the subsequent chapters, covering topics such as the sequencing method, the use of the genomics data in GWA studies and the prediction of missing genotypes.

In particular, this thesis focuses on predicting the missing genotypes in low-resolution data, because low-resolution genotype data is much cheaper and therefore can be obtained for every individual. Following some background on statistical methods for inferring missing genotypes in Chapter 2, Chapter 3 provides a simulation study that uses masked kākāpō genotype data to demonstrate the process of predicting the missing genotypes. The kākāpō sequence data makes it possible to assess the accuracy of genotype prediction, compare the performance of the statistical methods, and investigate the factors that affect prediction accuracy. The simulation study in Chapter 3 also demonstrates that genotype imputation can be computationally challenging in situations where the low-resolution genotype has a high error rate.

An alternative approach for missing data problem is to use the subsample data to estimate the same parameters in mixed-effects models as would be estimated with the complete sample data. When the subsample is selected at random or the selection depends on an observed covariate in the model, valid inference can be made from the subsample data using standard methods. However, random sampling is less efficient than outcome-dependent sampling for rare outcomes [16]. In order to obtain valid inference with incomplete genotype data, it is crucial to take into account the missingness structure when it comes to model fitting. Chapter 4 provides a review on two-phase sampling and the maximum likelihood estimation methods.

For mixed model inference under two-phase sampling, most estimation methods are developed for independent individuals and very few of them allow correlated individuals. In particular, methods that allow correlated individuals are usually only valid when sampling clusters are the same as model clusters, and such a sampling design is impossible for individuals with a complex correlation structure such as kākāpō. To the best of my knowledge, no methods have been proposed for mixed model inference under two-phase sampling where individuals have a complex correlation structure.

Therefore, this thesis focuses on a special case of two-phase sampling, where the individuals are related. In Chapter 5, I propose a weighted maximum likelihood estimation (MLE) approach for fitting linear mixed models that takes advantage of the fact that the kākāpō population kinship structure is known, making it possible to model the population covariance matrix rather than the sample covariance matrix. Since the population kinship structure is often known either exactly or approximately for endangered species, the proposed approach

provides a general solution for fitting linear mixed models under two-phase sampling designs in conservation genetics. The proposed method is written as an R package *WLMM*, which is available on GitHub (https://github.com/zoeluo15/WLMM).

For binary outcomes, McCulloch proposed a Monte-Carlo expectation-maximization (EM) algorithm with a Gibbs sampler to maximize the likelihood of a probit-normal model with independent random effects [98]. In Chapter 6, I consider an extension of McCulloch's model and propose a weighted maximum likelihood method for generalized linear mixed models with correlated random effects under two-phase designs.

Chapter 2

Genetic background

This chapter provides the background knowledge of the subsequent chapters. I start with the introduction to some fundamental genetic concepts in Section 2.1. Next, Section 2.2 gives an overview of the process of obtaining the genetic data used for methods evaluation in later chapters. Then, Section 2.3 gives an introduction to GWA studies that use such genetic data to identify variations in DNA sequence associated with a trait. Finally, since the interest of this thesis is reducing the cost of obtaining genetic data and making use of the low-resolution genotype, Section 2.4 describes the idea of predicting missing genotypes and relevant statistical methods.

2.1 Fundamental genetics concepts

Deoxyribonucleic acid (DNA) is a double-stranded helix that is composed of two sequences of nucleotides and describes the genetic information of an individual. A nucleotide is composed of a nitrogenous base, a sugar, and a phosphate group, where the sugar is deoxyribose attached to the phosphate group. There are four types of nucleotides in DNA, classified by the nitrogenous base, which may be either adenine (A), cytosine (C), guanine (G), or thymine (T). The nucleotides in a strand are joined together through the sugars and phosphates, and two strands are joined together by nitrogenous bases on different strands. Two nitrogenous bases from different strands that bond together are referred to as a base pair, where A and T always pair together and C and G always pair together. The two strands of DNA are parallel but oriented in opposite directions. Each strand begins with the 5' phosphate group of the first nucleotide in the strand (referred to as 5') and terminates with the 3' hydroxyl of the last nucleotide (referred to as 3').

The complete set of DNA, known as the genome, is organised into a number of different chromosomes, and the number varies between species. The chromosomes with similar

lengths and sequences come into sets. For humans and animals, there are two chromosomes in a set, with one inherited from the maternal parent and one inherited from the paternal parent, and such species are called diploid. For example, humans have 23 pairs of chromosomes including one pair of sex chromosomes (XX for females or XY for males), and kākāpō have 26 pairs of chromosomes including one pair of sex chromosomes (ZW for females or ZZ for males).

A source of genetic variation in the DNA inheritance process that makes an individual different to their parents is meiosis. Meiosis is a process that reduces the number of chromosomes in the parent cell by half and produces cells for sexual reproduction. The process of meiosis involves four steps: replication, crossover, the first cell division and second cell division. In the first step, a double-stranded DNA is replicated to create an identical copy, where each strand serves as a template to produce a new strand. The two copies of a chromosome are linked together to form an X-shape, where the pair of linked copies is referred to as sister chromatids and the pair of unlinked copies is referred to as non-sister chromatids. Errors that occur in the DNA replication process (e.g. a single nucleotide base is changed, inserted or deleted from a DNA sequence) are a source of point mutations. For humans, the copy error rate is approximately 1 per 10,000 nucleotides.

Next, copies of the same chromosome are paired up to exchange DNA materials between non-sister chromatids. This process is known as crossover or recombination, which results in two unique non-identical sister chromatids. In other words, recombination creates a new combination of DNA materials in an individual compared to the DNA materials found in their parents. Then, the two pairs of chromatids are randomly divided into two cells in the first cell division, and the sister chromatids within both pairs are randomly divided into four gametes. A gamete is a reproductive cell of an individual, also known as a sex cell, which is haploid because it only contains one set of chromosomes. For example, in humans, gametes contain one chromosome from each of the 23 chromosome pairs. The process of cell division results in a new combination of maternal and paternal chromosomes, which is also referred to as independent assortment. As a result of recombination and independent assortment, genetic variation is introduced to an individual, making the individual different to their parents.

2.2 Genotyping by sequencing

Genotyping by sequencing (GBS) is a high-throughput sequencing method for discovering genetic variation in order to perform genotyping studies. GBS is suitable for genetic studies of endangered species, livestock and plants because it is capable of efficiently producing high-density genotype data of a large number of DNA samples at a low cost, and can be

generalized to non-human organisms. GBS provides a rapid and low-cost This section provides an introduction to the GBS process as the kākāpō genetic data used in the thesis is generated using this method [53]. GBS is comprised of the following steps:

Step 1. Hair, tissue or blood samples of target individuals are collected;

Step 2. Multiple copies of DNA are extracted from each sample;

Step 3. Copies of DNA are cut into many small pieces using restriction enzyme(s), a protein produced by bacteria, at specific sequences that are recognised by the restriction enzyme;

Step 4. Adapters and barcodes are attached to the ends of the DNA fragments, where adapters are short sequences needed in PCR amplification in step 6, and barcodes are unique sequences of nucleotides used to identify samples from different individuals;

Step 5. DNA from different samples are pooled together;

Step 6. DNA fragments are amplified using a polymerase chain reaction (PCR) step, which is a technique that clones the targeted parts of a DNA sequence and produces thousands to millions of copies of target DNA;

Step 7. DNA fragments that are either too short or too long are removed, where the size selection depends on the laboratory setting;

Step 8. One end of the DNA fragments is sequenced with a fixed number of nucleotides using a sequencing machine, and the end of the fragment is called a sequence read;

More details about the GBS laboratory protocol can be found in Elshire et al. [47] and the size selection step in Dodds et al. [43].

In order to be used in genetic analysis (e.g. genome-wide association study introduced in the following section), the sequence reads need to be processed into a Variant Call Format (VCF) file [38], which contains information on genetic variation. First, DNA fragments belonging to the same individual are combined into a file by identifying reads with the same unique barcode. Then, the location of sequence reads on the genome is determined by mapping to a reference genome, which is a representation of the genome sequence for a particular species. Finally, the genetic variation is identified by scanning for variation at a single base pair across the reads from all individuals that are mapped to the same location on

the reference genome. Such a genetic variation is called a single nucleotide polymorphism (SNP) and caused by point mutation.

For each SNP called from the GBS data, the VCF file records the number of reads for the reference allele (when the observed DNA state at a position is the same as the reference genome) and alternative allele (when the observed DNA state at a position is different to the reference genome) for each sample. The read count data are then used to infer the genotype of the called SNPs.

A genotype refers to an unordered set of alleles carried by an individual at a particular genetic marker. A genetic marker refers to a variant in DNA sequences with a known physical location on a chromosome, which includes SNPs but can also be a sequence of DNA. For a diploid individual that has two copies of each marker, the genotype of a biallelic SNP (only one alternative allele) is coded as either "0/0", "0/1" or "1/1", where 0 indicates reference allele, and 1 indicates alternative allele. A genotype is called homozygous if the two alleles are identical and heterozygous otherwise. The forward slash (/) notation means it is unknown which is the maternal/paternal allele. An ordered sequence of such alleles along a single chromosome is called a haplotype, where the alleles in a haplotype tends to be inherited together.

However, haplotype reconstruction and genotype inference cannot be carried out directly due to sequencing errors and missing parental alleles. Sequencing error refers to the mistake in called nucleotide base, which can occur during the sequencing process in the laboratory or result from incorrect alignment to the reference genome. The problem of missing parental alleles arises when there is no reads cover one of the two parentally inherited chromosomes, which is a consequence of low read depth. While the first problem can result in false homozygous calls at heterozygous sites, the second problem can result in false heterozygous calls at homozygous sites [54], and both problems are missleading in locating the recombinations. Therefore, quality control of the called variants is necessary for haplotype construction and genotype imputation.

2.3 Genome-wide association studies

The goal of a GWAS is to find the association between genetic variants and a particular trait (e.g. human disease, measurable characteristics etc.), by scanning SNPs across the genome in different individuals, and searching for SNPs that are found more frequently in individuals with the trait than individuals without the trait. If any SNPs are identified to be associated with the trait, then regions near these SNPs may contain a gene or genes that are responsible for the trait. This is based on the assumption that only low levels of recombination

occurred between the associated SNPs and the risk gene in many past generations for the majority of the population [104].

The first successful GWAS, about myocardial infarction, was published in 2002 [105]. In 2005, the first result of significant association from a GWAS was reported: two SNPs were found that change the allele frequency in patients with age-related macular degeneration [77]. By September 2018, over 70,000 associations between genetic variants and traits have been found in over 5,000 human GWA studies [26]. Although a lot of GWA studies are focused the association between SNPs and human diseases, they can be applied to other organisms in the same manner.

Many GWA studies rely on SNP chips (or SNP arrays). SNP chips are designed to identify the specific nucleotides present at hundreds of thousands of locations across the genome where SNPs are known to exist. The selection of SNP locations depends on the population minor allele frequency (MAF), in other words, most SNP chips mainly cover common SNPs but cover very limited rare variants or even no rare variants. Therefore, despite the fact that GWA studies have successfully discovered associations between many common variants and human diseases, GWA studies are underpowered to detect associations with rare variants through linkage disequilibrium with common SNPs [84, 104]. Linkage disequilibrium refers to the phenomenon where alleles at nearby loci are correlated in such a way that they are inherited together more frequently than expected if they were independent and randomly associated. Unlike common variants which can be found in many individuals in the population regardless of ethnicity and relatedness, and a certain proportion of moderately rare variants can be found in other subpopulations [56, 143], very rare variants are usually shared only within ethnic groups or families. Therefore, population-based GWA studies need to either take a larger sample size or include more inividuals with a particular trait in order to find association between rare variants and the trait. For example. in a case-control GWAS which aims to determine whether a SNP is more frequently associated with cases (diseased individuals) or controls (healthy individuals), increasing the number of cases can enrich the genetic loci containing variants that are rare in the population.

Another problem with population-based GWA studies is the presence of a substructure within a population, which can be caused by subpopulations with different ethnic ancestries, different environments or different disease-related genetic factors. Population stratification can lead to a difference in allele frequencies within case and control groups, and hence increase the rate of type 1 error in population-based GWA studies. For example, previous studies have showed differences in genetic predictions of height among the European populations, but the differences were overestimated due to unrecognized population stratification in genome-wide association studies [135]. Family-based GWA studies overcome the challenge

of unidentified population heterogeneity because they are conditional on the genomes of the founders of each pedigree instead of population allele frequencies.

Linkage analysis is a family-based approach that searches for genetic markers that cosegregate with a particular phenotype through families. It has successfully identified genetic variants that cause rare diseases such as Huntington disease [92]. In other words, a linkage study aims to identify loci that are physically close to each other along a chromosome, and such loci are defined to be linked. If two or more loci are linked, their alleles tend to be inherited together within families. Linked loci can be broken up by recombinations, and the closer the loci of two markers, the lower the probability of recombination, and vice versa. Thus, recombination frequency (frequency of a recombination occur between two genetic markers) is used to identify whether two loci are linked and how tightly [14]. Therefore, linkage analysis is a powerful tool to localize potential causal genes (i.e. genes that are responsible for a particular disease or trait) for related individuals by the linkage between those genes with genetic markers.

In contrast to linkage analysis, which relies on the fact that disease/trait-causing genes are inherited together with genetic markers within a family, GWAS studies rely on linkage disequilibrium between disease/trait-associated variants and causal variants within population. Genetic linkage and linkage disequilibrium are different in terms of scale which is demonstrated in Figure 2.1. Morgan (M) is a unit for recombination frequency that measures the relative distance between genes on a chromosome, and one morgan means that the expected number of recombination between two genes on a chromosome is one. Linkage analysis looks at genetic markers that are separated by multiple centimorgans (cM, and 1 cM=100 M). In humans, 1 cM approximately equals to 1 Megabase pair (Mbp), and 1 Mbp equals to 1,000 kilobase pairs (kbp). While linkage can be detected within a long interval ($\leq$ 5 Mbp), association requires a much finer scale ($\leq$ 100 kbp) in order to be detected [104]. Since linkage can be detected in long intervals, the number of genetic markers needed for linkage analysis is much less than association analysis. However, this also leads to a wide range of candidate regions containing the potential causal genes (e.g. 10–20 Mbp) [87]. Therefore, it is common in family-based designs to use linkage analysis to first identify candidate regions, and then conduct an association analysis which requires denser genetic markers to narrow the region of interest (e.g. [35, 59–61, 82, 94]).

For the human genome, there are possibly at least 10 million common genetic variants (MAF>0.05) [72], and GWA studies typically genotype only a fraction (e.g. 100,000–1,000,000) of all genetic variants. In some scenarios, we may have genotype information with even lower coverage or low-depth sequencing data with high sequencing error. A GWAS requires high-density SNPs across the genome in order to narrow down intervals that contain probable

causal variants. Although the genotyped genetic variants might not be enough to narrow the interval of interest, they provide useful information for inferring the non-genotyped variants in the same group of individuals.

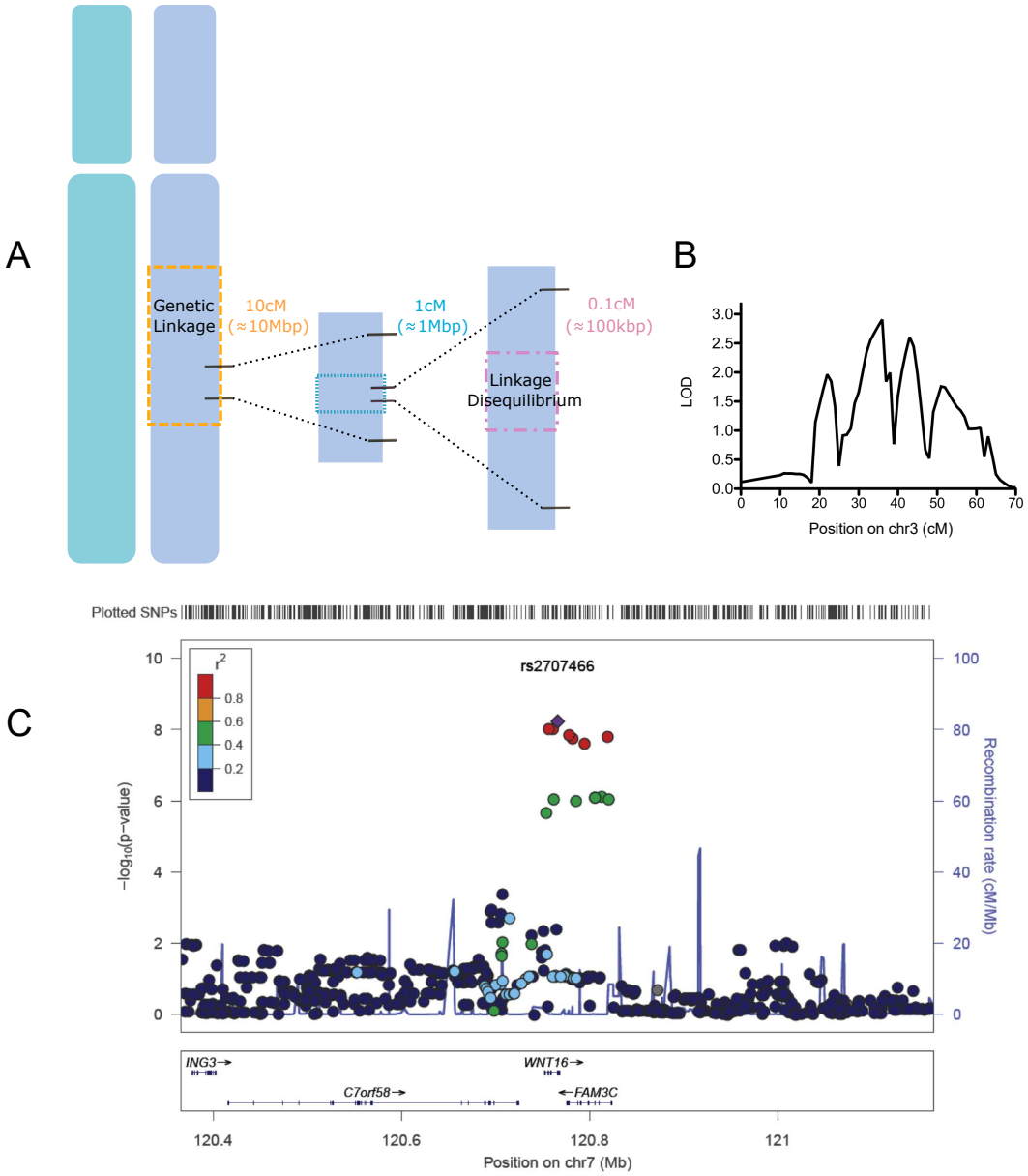

Fig. 2.1 A: different scales of genetic linkage and linkage equilibrium on human genome; B: Linkage of two genes on Chromosome 3 and the LOD score is an estimates of relative probability that two loci on a chromosome are physically close enough to each other and hence they are likely to be inherited together [69]; C: SNP rs2707466 regional association plot of the discovery genome-wide meta-analysis. Circles show GWA meta-analysis p-value of SNPs on Chromosome 7, with different colors indicating varying linkage disequilibrium with rs2707466 (diamond) [170]. Note that 1Mb $\approx$ 1cM.

2.4 Genotype imputation

The process of predicting the unobserved genotypes is referred as genotype imputation, which is a strategy that can boost the power of a GWAS without extra sequencing. There are two types of genotype imputation, one is family-based imputation with related samples, the other one is population-based imputation with unrelated samples.

2.4.1 Family-based genotype imputation

For a family-based GWAS in which candidate regions have been identified by linkage analysis, the goal is to collect sequencing data of the candidate regions in order to test for associations between genetic variants and the trait. However, the budget may allow sequencing candidate regions for only a subset of the sampled individuals. On the other hand, sequencing every individual in the family is equivalent to effectively sequencing many stretches of chromosome more than once because relatives share long stretches of chromosome. Therefore, it is cost-saving and efficient to sequence a subset of individuals in a pedigree, and use their sequencing data to infer genotype of their relatives.

To illustrate the process of imputation for related individuals, consider the example two-generation pedigree in Figure 2.2 which contains two parents and four offspring. For individuals with identical nucleotide sequences in the stretch of shared chromosome, the common DNA segment is called identical by state (IBS) in these individuals. An IBS segment is identical by descent (IBD) if the individuals inherited the segment from the same ancestor. In Figure 2.2A, all individuals in the family have pre-existing genotype information for a set of markers (the loci colored in red), and the genotype for remaining markers are obtained for parents only but left to be imputed for offspring (indicated in black). Then offspring with partial genotype information are compared with parents with complete genotype information in this DNA segment to identify stretches of shared chromosome within the pedigre. This process is demonstrated in Figure 2.2B, where each haplotype transmitted between two generations is given a unique color. Finally, as shown in Figure 2.2C, the missing genotypes of offspring are imputed if the haplotypes they inherited are IBD to copies of the same haplotypes.

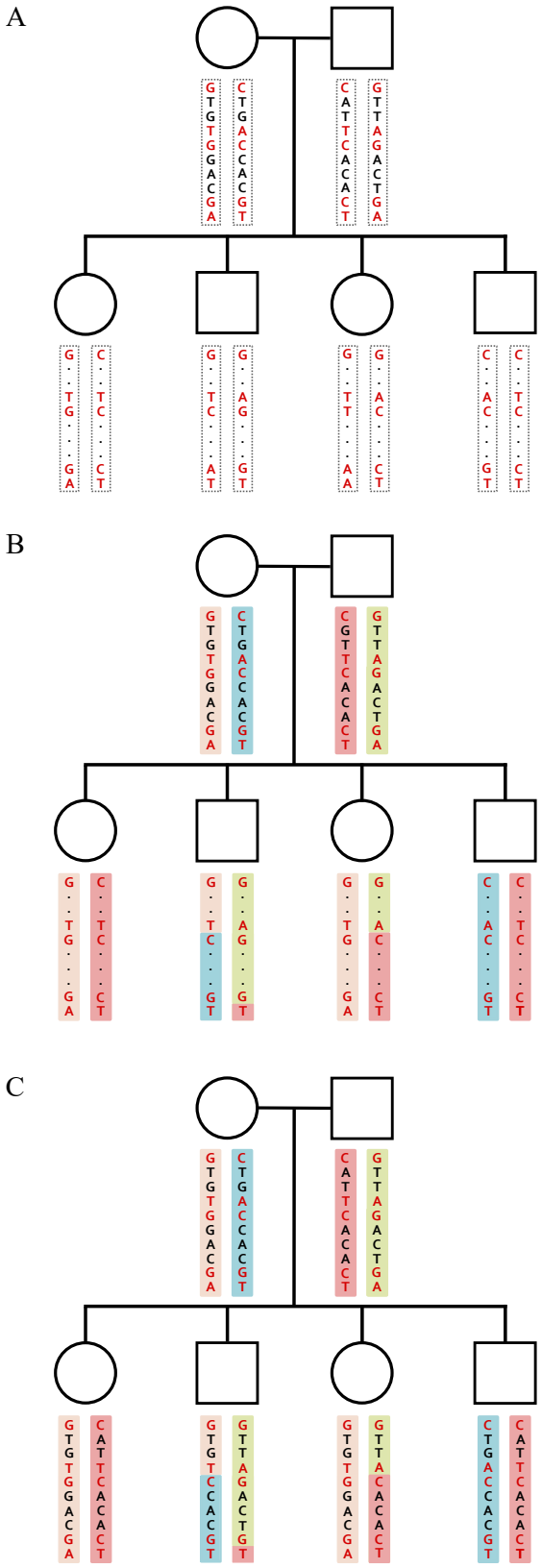

Fig. 2.2 The process of family-based genotype imputation. The pedigree shows the relationships between members in a two-generation family. Parents are the first generation at the top and offspring are the second generation at the bottom. Females are represented by circles and males are represented by squares.

The IBD pattern of a pedigree at a single locus can be described by a quantity called the inheritance vector (IV) [1]. For example, consider a pedigree structure at a particular locus given in Figure 2.3. The two top generation individuals and the first inividual on the left at the second generation are called founders because they don't have parents in the pedigree. Each non-founder has two binary numbers in the bracket, namely meiosis indicators, which indicate the pattern of allele transmission at this particular locus. The first (resp. second) binary number represents the allele transmission from the individual's mother (resp. father), and 0 (resp. 1) indicates the maternal (resp. paternal) copy of the allele is transmitted. Note that founders do not have a pair of meiosis indicators since there is no information on their parents. The vector that contains meiosis indicators of all non-founders at a locus is then called the IV. If there are N non-founders in a pedigree, then the IV at any locus on a chromosome is a vector with $2N$ elements. Since an allele can be either maternal (denoted by 0) or paternal (denoted by 1), in total there are 2^{2N} possible inheritance patterns of the pedigree at the locus.

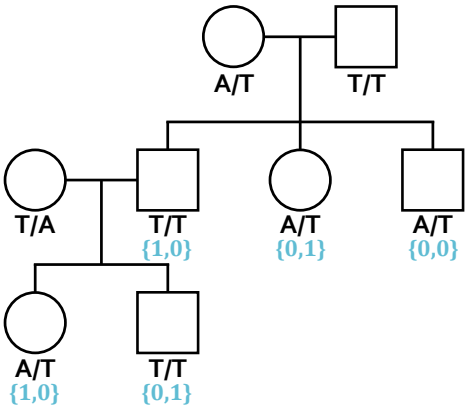

Fig. 2.3 An IV (indicated in blue) of non-founders shows the inheritance pattern in a pedigree at a particular locus. The pedigree is a three-generation pedigree with grandparents at the top and grandchildren at the bottom. Females are represented by circles and males are represented by squares.

`Merlin` [1] and `GIGI` [34] coupled with `gl_auto` [142] (part of the `MORGAN` program) are two family-based imputation approaches for general pedigrees that rely on IBD computation, but `GIGI` is compatible with large pedigrees whereas `Merlin` requires splitting the pedigree. While `Merlin` computes IBD internally based on the Lander-Green algorithm [81], `GIGI` uses `gl_auto` for IBD computation. The idea of the Lander-Green algorithm is to model the flow of alleles of a pedigree at multiple loci as a Markov process with hidden states being IV at each locus [5]. For small pedigrees, `gl_auto` infers the shared segments of a chromosome by sampling IVs with probabilities obtained from exact computations based

on the Lander-Green algorithm. For large pedigrees, `gl_auto` uses a Markov Chain Monte Carlo (MCMC) sampler based on both the Lander-Green algorithm and the Elston-Stewart algorithm [48] to approximate the likelihood of observed genotype on a pedigree [157].

IVs are sampled for a set of framework markers which are sparsely distributed on the chromosomes based on the observed genotypes of the framework markers using `gl_auto`. In general, most information of IVs in a pedigree can be extracted by a moderate number of framework markers [158, 159]. Hence, IVs at the position of dense markers between two framework markers can be sampled based on the IVs sampled at the two framework markers as IVs at nearby positions are highly correlated.

Let S_v denotes the IV at the position of a dense marker v, and s denotes a configuration of the IV. `GIGI` estimates the probability distribution of the missing genotype G_{iv} of individual i of dense marker v being a particular genotype configuration g conditional on the observed genotypes of all framework markers G_F^{ob}, the observed genotypes G_v^{ob} of dense marker v.

$$
\begin{aligned}
P(G_{iv} = g | G_F^{ob}, G_v^{ob}) &= \sum_s P(G_{iv} = g \,|\, S_v = s, G_F^{ob}, G_v^{ob}) P(S_v = s \,|\, G_F^{ob}, G_v^{ob}) \\
&\cong \sum_s P(G_{iv} = g \,|\, S_v = s, G_F^{ob}, G_v^{ob}) P(S_v = s \,|\, G_F^{ob}) \qquad (2.1) \\
&\cong \sum_s P(G_{iv} = g \,|\, S_v = s, G_v^{ob}) P(S_v = s \,|\, G_F^{ob}) \qquad (2.2) \\
&= \sum_s \frac{P(G_{iv} = g, G_v^{ob} \,|\, S_v = s)}{\sum_k P(G_{iv} = k, G_v^{ob} \,|\, S_v = s)} P(S_v = s \,|\, G_F^{ob}) \qquad (2.3)
\end{aligned}
$$

The exact equality in Eq 2.1 holds when dense marker v is one of the framework markers. Eq 2.2. And Eq 2.1 is approximately equal to the LHS by assuming the influence of including additional genotype of dense marker v on the IV inference at the position of v is small. Eq 2.2 is an exact equality when the framework markers are in linkage equilibrium with dense marker v, and a good approximation otherwise as genotype of distant marker is unlikely to be informative. In Eq 2.3, each term in the fraction can be be computed efficiently [80, 134] where the bottom is summed over all possible genotype configurations, and the remaining term can be estimated by averaging over the sampled IVs at position v. Then, a Monte Carlo estimator for genotype G_{iv} is

$$
\widehat{P}(G_{iv} = g | G_F^{ob}, G_v^{ob}) = \frac{1}{n^*} \sum_{j=1}^{n} P(G_{iv} = g \,|\, S_v^j, G_v^{ob}),
$$

where $S_v^j n$ is the IV sampled at iteration j for $j = 1, \ldots, n$, and n^* is the number of iterations that S_v^j is consistent with the observed genotypes G_v^{ob} at dense marker v. Finally, `GIGI` calls

the most likely genotype based on the estimated probabilities. Alternatively, one can set a threshold, and the complete genotype can be called if

$$\widehat{P}(G_{iv} = g|G_F^{ob}, G_v^{ob}) > t_1,$$

where t_1 is a user-defined threshold. When there is ambiguity in genotype calling (i.e. the estimated probability of the missing genotype being the heterozygous configuration is equal to t_1), the algorithm will call the more likely allele from the two alleles, that is:

$$\widehat{P}(G_{iv} = a/\cdot|G_F^{ob}, G_v^{ob}) > t_2,$$

where $a/\cdot$ denotes the genotype contains an a allele, and $t_2 = t_1 + (1 - t_1/2)$.

2.4.2 Population-based imputation

The intuition behind population-based imputation is that apparently unrelated individuals still share short genome sequences inherited from distant ancestors. Sequencing a small panel allows common haplotypes to be measured, and the haplotypes are put together into imputed genotypes according to the SNP information.

Genotype imputation for unrelated individuals follows a similar process to family-based imputation, and the idea is illustrated in Figure 2.4. The unrelated individuals in Figure 2.4B are first genotyped at a modest number of genetic markers, but genotypes at most markers remain unknown. The next step is to assign observed alleles to the paternal and maternal chromosomes, and this process is called phasing or haplotype estimation. Then the estimated haplotypes are compared to reference haplotypes in Figure 2.4A to identify shared stretches (Figure 2.4C). Unlike family-based imputations, for which the observed genotypes of a relatively small subset of individuals can make useful predictions of the genotypes of the rest pedigree, population-based imputation require a much larger set of genotyped individuals from the same ethnic origin to serve as a reference panel. The accuracy of population-based imputation increases as the size of reference panel increases, particularly for rare variants [22]. An example of a reference panel of the human genome is HapMap, which consists of genotypes at several million genetic markers for 269 individuals from different ethnic groups [71]. Finally, the missing genotypes are imputed using the matching reference haplotypes (Figure 2.4D).

In contrast to IBD-based phasing in related individuals, population-based phasing uses the linkage disequilibrium information pooled from hundreds to thousands of individuals to model the haplotype frequencies. A detailed review on haplotype phasing is given in [24],

with a focus on the developement of population-based phasing methods. While the phasing accuracy in common variants can be improved by increasing the sample size, marker density and so forth, accurate phasing for rare variants remains a problem for population-based approaches [24].

In particular, both family-based and population-based phasing methods are unable to phase *de novo* mutations (i.e. mutations that are arise for the first time in an individual that are not found in the genome of its parents). Read-based phasing (that uses sequencing reads covering at least two heterozygous variants to construct haplotypes) provides a more comprehensive understanding of the genome because of its ability to phase rare variants and *de novo* mutations. Read-based approaches such as `WhatsHap` [97] and `HapCut2` [45] phase individual genomes by aligning the sequencing reads to the reference genome, and reads that cover at least two heterozygous variants are partitioned into two groups that correspond to the pair of haplotypes. When pedigree information is available, `WhatsHap` is able to utilize the fact that the haplotypes of an offspring are a recombination of the haplotypes of its parents to increase the phasing accuracy [97]. Due to the length of sequencing reads, the number and type of variants can be phased by read-based approaches has been limited, until the rapid development of long read sequencing technologies in recent years.

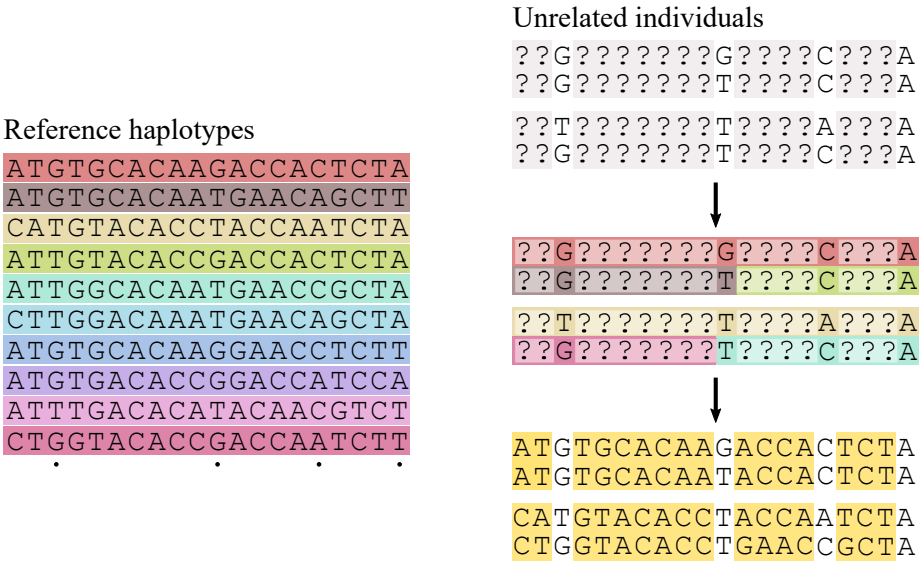

Fig. 2.4 The process of population-based genotype imputation. A: the reference set of haplotypes; B: unrelated individuals with partial genotype information; C: the observed haplotypes are colored according to their matching with haplotypes in the reference set; D: missing genotypes are imputed using the matching reference haplotypes.

Several population-based genotype imputation methods are computationally feasible for a GWAS purpose (e.g. [22, 66, 86, 124]). These methods typically use hidden Markov

models (HMMs) to sample haplotype pairs for each individual conditional on their unphased low-density genotype data, and then the missing genotypes are imputed by the inferred haplotypes. The basic idea of HMMs is that there are hidden states underlying the observed data, and these hidden states have a Markov structure (i.e. the next state only depends on the current state and does not depend on any previous states). In the context of haplotyping and genotype imputation, the observed data is the unphased low-density genotype, and the hidden state is the haplotype phase. And the observed genotype at a marker only depends on the haplotype phase at that marker.

`FastPHASE` [124], `MaCH` [86] and `IMPUTE2` [66] are based on the Li and Stephens framework [85]. Under this framework, each reference haplotype is a hidden state of the HMM at each marker. The underlying true haplotypes are assumed to be combinations of reference haplotypes where the switches from one reference haplotype to another represent the recombinations. In order to account for mutation and genotype error, the framework also allows the observed alleles to be different to the alleles on the underlying true haplotypes.

The model parameters are estimated using the expectation-maximization (EM) algorithm [41], which is an iterative method to obtain maximum likelihood estimation in the presence of latent variables. In this situation, the EM algorithm starts with an arbitrary guess of the haplotype phase and missing genotypes and uses them for the maximum likelihood estimation of model parameters such as recombination rate, mutation rate and genotype error rate. Then, the parameter estimates are used together with the observed genotype data to re-estimate the haplotype phase and mixing genotypes. The EM algorithm stops once the convergence of the parameter estimates is reached.

In `MaCH` and `IMPUTE2`, the hidden states in HMM are the reference haplotypes. During each EM algorithm, the haplotype pair of each individual is re-estimated using the reference haplotypes previously estimated for the other individuals. In contrast to `MaCH` which estimates recombination rates and mutation rates between markers in the model fitting process, `IMPUTE2` requires a recombination map to derive these model parameters. In order to better capture the complex patterns of LD, `fastPHASE` uses a cluster of similar haplotypes as each reference haplotype and allows the cluster membership changes along a chromosome as the parameters are updated in the model fitting process. While phasing and imputation accuracy can be improved by enriching the reference haplotypes, the computation time of these methods grows quadratically as the number of haplotype clusters/reference haplotypes increases.

In contrast to the above methods based on the Li and Stephens framework, the Browning model in `BEAGLE` [22, 23] is more parsimonious, and there are some important differences. First, the number of states at each marker varies along the genome to handle the difference in

complexity at different markers. Note that both BEAGLE and fastPHASE consider clusters of similar haplotypes as hidden states, but the number of clusters is fixed in fastPHASE. Second, BEAGLE does not explicitly model recombinations and mutations but accounts for them in the transitions between states. Third, there are at most k transitions from one state at a marker to states at the next marker, where k is the number of observed alleles for the next marker (e.g., $k = 2$ for SNPs). Fourth, there is only one possible outcome (i.e. the observed allele) from each state in BEAGLE whereas methods under Li and Stephens framework include both the observed allele and mutation in the model. These differences together restrict the possible transitions in HMM given observed genotypes, thus reducing the computation time in parameter estimation.

2.4.3 Subject selection strategies

In population-based imputation, unobserved genotypes can be imputed using external sequencing data as a reference. For example, the HapMap [70–72] haplotypes can be used to impute the missing genotypes if the ancestry of the sample is close to one of the ancestry groups in the HapMap project. However, when there are no such reference haplotypes available, a subsample selected from the original sample can serve as the reference. While population-based imputation allows both types of reference data, family-based imputation requires selecting individuals for sequencing from the same pedigree as the original samples [34, 122]. In many cases, particularly for wildlife populations, the budget constraint still remains a problem for sequencing a large number of individuals. Thus, it is important to prioritize the individuals chosen for sequencing.

The choice of individuals for sequencing has a direct impact on the imputation quality of the non-sequenced individuals, hence it is important to carefully choose which subset of individuals to sequence. Subject selection depends on the type of the study. In population-based studies, individuals are considered unrelated or distantly related, and hence share little common genetic information. Therefore, it is natural to select a subsample for sequencing in a way that the reference panel covers as many distinct haplotypes as possible to achieve a high imputation accuracy. On the other hand, individuals in family-based studies are closely related, and genotype imputations profit from the fact that closely related individuals share longer stretches of IBD DNA segments. Thus the ideal reference individuals would contain the most genetic information of their relatives. There has been a number of subject selection strategies developed in order to attain a higher accuracy of predicting the genotype of non-sequenced individuals [137, 36, 34]. This section provides a review on three prevalent subject selection strategies including PRIMUS, ExomePicks and GIGI-Pick, which are favourable for different situations in genotype imputation. Based on their designs, these selection

strategies can be classified into two categories: (1) a method for population-based imputation: PRIMUS; and (2) methods for family-based imputation: ExomePicks and GIGI-Pick.

PRIMUS

PRIMUS is a subject selection strategy that identifies a set of maximumly unrelated individuals [137]. The program takes the estimates of pairwise kinship (i.e. a measure of relatedness) as input and transforms them into an undirected graph consisting of multiple family networks, with nodes being individuals and edges being relationships above a user-defined threshold of relatedness. Individuals from different family networks are considered as unrelated.

Within each family network, PRIMUS identifies the maximumly unrelated set of individuals by searching the maximum clique in the complement graph. In the complement graph, edges are relationships below the user-defined threshold (i.e. original edges in the graph are removed because they correspond to relationships above the threshold). A clique is a subgraph that has all of its nodes connected to each other. A maximum clique is the largest clique that is not part of other cliques. Then, the maximumly unrelated set in all individuals is the combination of maximumly unrelated set within each family network. When there are more than one maximum clique, a unique strength of PRIMUS is to weight the maximum clique given additional criteria (e.g. disease status, data completeness .etc).

In terms of GWA studies, when the variants are assumed to be rare in the population, PRIMUS selects subjects by obtaining more copies of the variants to find the associations between variants and the trait [99]. For population based-imputation, PRIMUS should also lead to a higher accuracy than random selection as maximumly unrelated individuals carry more unique haplotypes than randomly selected individuals. On the contrary, PRIMUS can be worse than random selection in family-based imputations as maximumly unrelated individuals provide little information on the other individuals [147].

ExomePicks

For each pedigree in the dataset, ExomePicks starts from the oldest generation and moves towards the youngest generation of the pedigree, selecting individuals from each generation for sequencing [36]. The program requires only a pedigree file, which describes the relationship between individuals, and a data file, which describes the content of the pedigree file. In particular, the data file should at least indicate the individuals in the pedigree that have been genotyped at some markers, and hence are eligible to be selected for sequencing.

If pre-existing genotype information is available for all individuals in the pedigree, `ExomePicks` selects all the founders and at least one of their offspring. The founders are selected for identification of all chromosomes segregating in the pedigree, and at least one offspring for each founder is selected because they provide information needed for phasing. If a founder has not been genotyped, an additional offspring from this founder or an offspring from the same generation will be selected.

However, the IBD sharing inferred by pedigree structure does not include the variance that is introduced by the probabilistic process of Mendelian segregation [144]. For example, an offspring has exactly 50% IBD sharing with a parent but IBD sharing with a sibling has an expectation of 50% and non-zero variance. It is important to take variation in IBD sharing into account because the IBD sharing between sequenced individuals and genotyped but non-sequenced individuals provide information needed for phasing, and hence imputing the missing genotypes [29]. Therefore, `ExomePicks` might be a desirable choice for subject selection only when high-density SNP markers are available on all individuals, two examples are given in [4] and [145].

`GIGI-Pick`

A pedigree file does not specify the allele inheritance pattern in a pedigree, however, the inheritance vector does. `GIGI-Pick` is a subject selection strategy that profits from the inheritance vector. Similar to `ExomePicks`, `GIGI-Pick` also makes use of the IBD sharing, but `GIGI-Pick` requires relatively sparse markers in comparison to `ExomePicks`. In other words, the strength of `GIGI-Pick` is dealing with uncertainty in the data. Given the genotype of the markers, marker map positions, pedigree structure and population allele frequencies, the program `gl_auto` uses MCMC to sample IVs that are consistent with the pedigree structure and pre-existing genotypes [33]. For each sampled IV, ability to impute genotypes for any choice of selected subject can be measured by calculating coverage.

Given a selection of subjects, coverage is defined as the expected percentage of alleles that are either observed or can be imputed by the observed genotype conditional on a given IV, and it is calculated as follows. Let I be the number of disjoint IBD graph in an IV. In an IBD graph, the nodes are copies of distinct chromosomes from subjects with observed genotypes, and each pair of nodes are connected by the subjects who inherted those copies of distinct chromosomes. When the observed genotypes can be phased, a total of F_i alleles can be inferred for the i-th IBD graph,

$$F_i = w_i + x_i,$$

where w_i is the number of copies of alleles already genotyped, and x_i is the number of copies of alleles can be imputed because they are on a copy of the same chromosome as copies of alleles with observed genotypes. When the observed genotypes cannot be phased, a total of G_i alleles can be inferred for the i-th IBD graph,

$$G_i = w_i + y_i,$$

where y_i is double the number of subjects whose genotype is unobserved but both their alleles are IBD with alleles of subjects with observed genotypes. Then for a sampled IV,

$$\text{coverage} = \frac{1}{2N}\sum_i (F_i p_i + G_i q_i),$$

where p_i is the probability that the observed genotype can be phased, and $q_i = 1 - p_i$ is the probability the observed genotypes cannot be phased.

The genotype imputation ability for a particular choice of subject selection at a random locus is measured by calculating the average coverage over all sampled IVs. The selection algorithm begins with calculating the coverage for all subjects available for sequencing and subjects with top coverages are kept as current top choices. Then the same calculation is applied on all possible combinations of the kept subjects from the last iteration and the remaining subjects, and the top choices are updated with an additional subject. This process is terminated when the number of chosen subjects reaches a specified threshold.

By sampling IVs, `GIGI-Pick` allows for stochastic variation in IBD sharing and also for the probability of recombination between markers, which is important when the markers are relatively sparse. In the simulation study presented by Cheung et al. [33], it was shown that `GIGI-Pick` outperforms `ExomePicks` for subject selection in a single pedigree dataset. When the pedigree is too complex or the number of pedigrees is large (e.g. in a population-based study), `GIGI-Pick` may be computationally infeasible due to constraints of the IV and the number of IVs need to be drawn for all pedigrees.

Other sampling strategies

There are other sampling strategies that are designed for different purposes, such as for association studies. Wang et al. [151] proposed G-STRATEGY, which aims to select a subset of individuals that reinforce the power of detecting an association by maximizing the objective function involving the enrichment value and the pairwise kinship coefficient. G-STRATEGY is appropriate for situations when none of the individuals are genotyped but all individuals

are phenotyped, and Wang et al. [151] showed that it has an outstanding performance in subject selection in datasets with relatively low relatedness among individuals.

David et al. [39] recently developed a sampling method for association studies that prioritize the individuals to be phenotyped when genotype data are available. Their method optimizes the sampling by maximizing the D criterion, which is a criterion that quantifies the merit of a design. In David et al. [39], the D criterion quantifies estimation errors of the target parameter, and the optimal subsample can be selected by maximizing the D criterion using STPGA (Selection of Training Populations by Genetic Algorithm) [2, 3]. David et al. [39] showed that optimized designs improve the precision of the joint estimation of breeding values and genetic effect of the locus, particularly for small sample sizes. Moreover, optimized designs are more efficient than simple random sampling for estimating locus effects for traits with simple genetic architecture.

A summary of the sampling strategies discussed in this section is provided in Table 2.1.

Subject selection strategy	Study type	Information required	Feature	Limitation
PRIMUS [137]	Population-based genotype imputation	Pairwise IBD	Allows preferential selection of a maximumly unrelated set	Underpowered in subject selection for family-based imputation
ExomePicks [36]	Family-based genotype imputation	Dense markers	Quickly suggest subject selection choice in large pedigree	Does not take into account of the variation in IBD sharing
GIGI-Pick [33]	Family-based genotype imputation	Sparse markers	Allows uncertainty in the data	Computationally expensive for large pedigrees
G-STRATEGY [151]	Association studies	Phenotypes and pedigree	Computationally feasible for larger dataset	To be investigated
D optimality [39]	Association studies	Genotypes	Works the best for traits with simple genetic architecture	Estimations are difficult for complex traits

Table 2.1 Summary of the subject selection strategies.

Chapter 3

Integrating the kākāpō data and simulations of existing selection strategies

Since imputed genotypes can be used to carry out numerous analyses such as GWAS, there have been many studies on the factors that can affect the quality of genotype imputation for humans and livestock [63, 67, 96, 131, 147, 148]. Previous studies showed that factors include the size and quality of the reference haplotype panel, composition of the reference haplotypes, MAF of the target SNPs, genotype density, the relationships between imputed individuals and reference individuals and so forth. However, there are few such studies for wildlife populations or endangered species because it is too expensive to sequence the whole population. Therefore, the kākāpō WGS data provide a great opportunity to learn the differences in the genome and pedigree structure between kākāpō and humans in the context of genotype imputation. The aim of this chapter is to compare the performance of various combinations of selection strategies and imputation methods given the masked kākāpō genomics data. The masked genomics data is simulated by hiding the sequences in chosen genome regions.

This chapter begins with a description of different ways of simulating low-density kākāpō genotype data in Section 3.2, and followed by a brief comment on the variant calling and quality control of the kākāpō genomic data in Section 3.3. Then, Section 3.4 discusses the application of existing strategies for selecting reference individuals in genotype imputation. Finally, Section 3.5 shows the results for haplotype phasing and genotype imputation, and discusses the factors that affect imputation quality for the kākāpō data.

3.1 The kākāpō GBS data

The kākāpō GBS data is provided by the Kākāpō125+, which is a gene sequencing project established by Kākāpō Recovery and the Genetic Rescue Foundation following the initial genome sequencing of Jane, a reference kākāpō, in 2015 at Duke University and Pacific Biosciences. The Kākāpō125+ consortium led by Genomics Aotearoa has been producing global analyses of the dataset focusing on genetic management, disease, fertility and ageing [44, 53]. This work has been undertaken in partnership with Ngāi Tahu, who are the traditional guardians (kaitiaki) for this kākāpō data set. The kākāpō GBS data contains the genotype information of 169 kākāpō (76 females and 93 males).

Among the 169 individuals, 125 of them were alive at the time, and samples from live kākāpō were collected as part of regular monitoring activities by the Kākāpō Recovery Team. DNA was extracted using a phenol/chloroform extraction protocol [123]. For the 44 historical samples, DNA was extracted from the toepads of the 13 historical birds using a DNeasy Blood & Tissue Kit (Qiagen, Hilden, Germany) with appropriate precautions taken to minimize the risk of contamination [78].

The extracted DNA was cut into many small fragments using restriction enzymes PstI and MspI (New England Biolabs, Ipswich, USA), which resulted in recognition sequences each containing a PstI restriction site (CTGCAnG) and a MspI restriction site (CnCGG). In the reference kākāpō genome which is 1.14 Gbp long (Gb = giga base pairs = 1,000,000,000 bp), the total hit count of the two cut sites is 1.74 million.

After the DNA was digested, barcoded adapters were added to the ends of the DNA fragments to distinguish between samples, which allows DNA from different samples to be pooled together. All the pooled DNA was then amplified using a PCR step, followed by removing DNA fragments that are shorter than 193 bp and longer than 500 bp (SAGE Science, Beverly, USA).

Finally, the remaining DNA fragments were sequenced using Illumina HiSeq 2500 with a 2x150 bp setup in at New Zealand Genomics Limited in Palmerston North, New Zealand for modern samples, and Illumina HiSeqX with a 2x150 bp setup at the SciLifeLab sequencing facility in Stockholm, Sweden for historical samples. More details of the DNA extraction and sequencing process are described in Dussex et al. [44].

3.2 Subsampling of sequencing reads

The best way to do low-density genotyping depends on what genotyping technologies are available at the time, and it will change over time, thus three approaches for simulating

low-density genotype data are considered here. For kākāpō, the sequenced reads across all individuals in the population were obtained using GBS. The most straightforward way to do low-density genotyping is to perform low-depth WGS, and this can be simulated by random sampling of the sequenced reads. More details on how I simulate low-depth WGS data are described in the following paragraph.

For humans, approximately an average read depth of 30-fold is considered to be sufficient for genotyping. This is relatively close to the read depth of the kākāpō data. The average read depth per sample in the complete kākāpō data has a mean of 17.03-fold and a median of 19.15-fold (with a minimum of 10- and a maximum of 40.19-fold), and the average read depth for each kākāpō is shown in the top figure of Figure 3.1. Here I consider an extremely low read depth of an average depth of approximately 2-fold (ideally each allele is covered once by the read). By randomly selecting a proportion of reads from each individual based on its average read depth (i.e. the higher the average depth, the lower the proportion of reads that are sampled), the average read depth per sample in the resulting low-depth GBS data has a mean of 2.469-fold and a median of 2.519-fold (with a minimum of 2.312- and a maximum of 3.65-fold), and the reduced average read depth for each kākāpō is shown in the top figure of Figure 3.1.

Fig. 3.1 Comparison of average depth and reference genome coverage between extremely low-depth (2-fold) and complete kākāpō GBS data. Average depth refers to the average number of times that a nucleotide base is covered by unique reads over the genome of the target inidividual, and reference genome coverage refers to the proportion coverage of the reference genome by sequencing reads. Note that male kākāpō have lower proportion of coverage because they have two Z chromosomes but no W chromosome (females have ZW).

It is expected that the error rate will be much higher in the variants called from low-depth GBS data than those from the complete data. However, low-depth GBS is much cheaper and more practical in wildlife conservation considering the current cost. For populations with a low recombination rate (e.g. the chance of a recombination event is only 1% in every million base pairs on average per generation for humans), the low-depth GBS data should also be enough to infer an offspring inheritance from its parents regardless of the error rate.

It is also possible to obtain low-density genotype data without reducing sequencing depth. This can be done by picking a grid of short sequences and sequencing them with good quality in all individuals. The size of the marker region on each chromosome should be short enough such that a marker of the offspring can be found in at least one parent, and long enough such that the marker is different between founders. In other words, two markers are identical only if they are IBD. However, the similarity between many founders is not much less than the similarity between first-degree relatives (e.g., parents/offspring, siblings) as shown in Figure

3.2. This is because of the incomplete pedigree information and the high level of inbreeding in the kākāpō population (see Figure 3.3). Consequently, it is difficult to pick a subset of markers that are useful for IBD tracking, thus this approach was not pursued in this study.

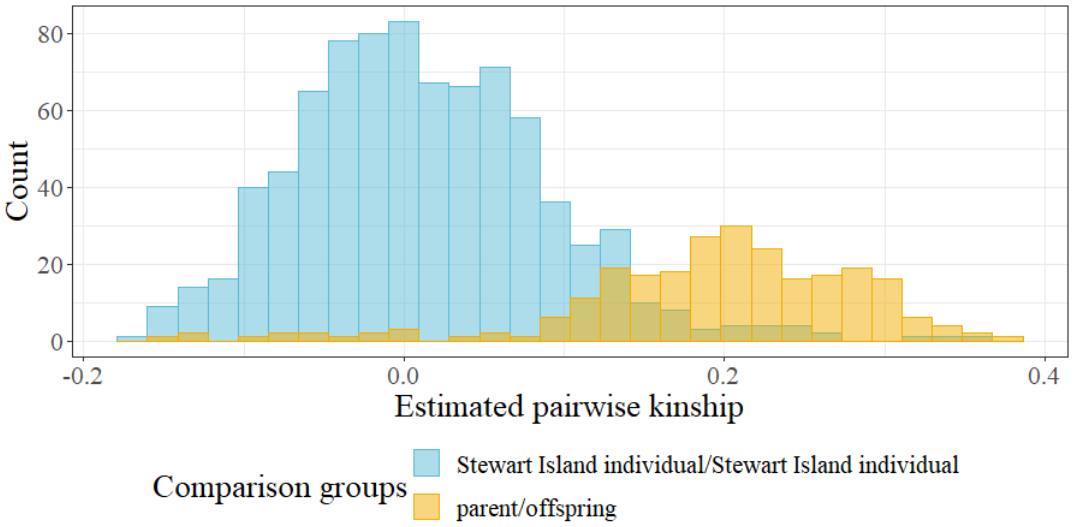

Fig. 3.2 Pairwise kinships inferred by GBS data using a marker-based approach proposed by Weir & Goudet [154].

Alternatively, all or a subset of the loci where the reference kākāpō, Jane, is heterozygous can be taken and genotyped in all individuals. To avoid confusion, Jane is the reference kākāpō for sequence mapping instead of the reference individuals for genotype imputation. This approach gives a random set of markers with varying density along the genome. Although Jane died without any offspring, its genome was sequenced with very high depth (approximately 100-fold). Therefore, genotyping Jane's variants in all individuals can produce a set of variants that are very likely to be true variants rather than genotyping errors. Using this approach, I obtained a set of low-density genotypes which contains 16,000 heterozygotes from Jane and an additional 4,000 heterozygous from Richard Henry. The reason for including heterozygotes from Richard Henry is discussed in Section 3.4. In later parts of the chapter, the low-density genotypes obtained this way are referred to as reference SNPs.

3.3 Variant calling and quality control

Variant calling is the process of mapping the GBS reads to a reference genome and identifying the variation in the alleles (i.e. SNPs) relative to the reference genome. For the kākāpō population, the Vertebrate Genome Project [44, 116] provided a high-quality genome assembly of a kākāpō (Jane), which is used as the reference genome. Initially, I

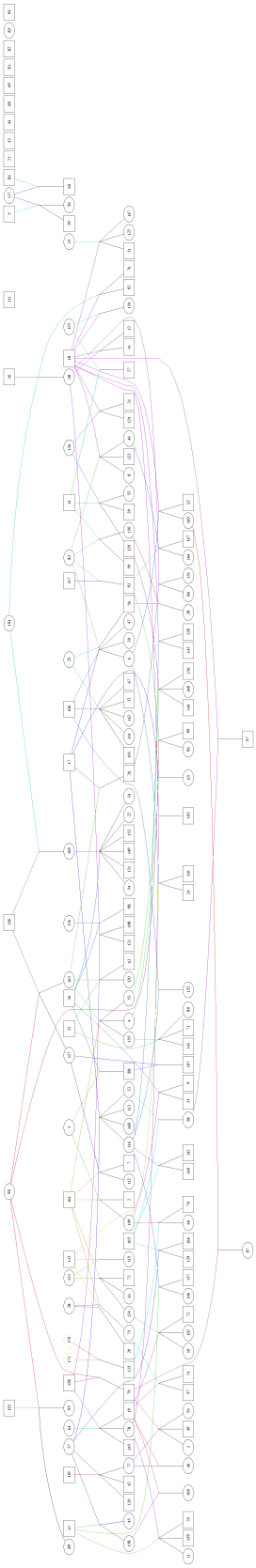

Fig. 3.3 The kākāpō pedigree, where circles represent females and squares represent males.

used the software `FreeBayes` to call variants from the low-density kākāpō GBS data, but the major issue of `FreeBayes` is the high error rate. Instead, I use the learning-based variant caller `DeepVariant` with a model trained using the kākāpō genomics data, and the code is adapted from Guhlin et al. [58]. The details of model training and examination of the performance of the trained `DeepVariant` model are described in Guhlin et al. [58]. Since `DeepVariant` is not a sex-aware caller, the downstream analysis focuses on the autosomes (non-sex chromosomes) only.

The number of SNPs identified for each individual before quality control is shown by the red dots in Figure 3.4. The number of SNPs called from extremely low-depth GBS data has a mean of 250,289, a median of 246,734, a minimum of 178,057 and a maximum of 502,095. The number of SNPs identified for each individual in the reference SNPs data has a mean of 10,212, a median of 10,202, a minimum of 8,792 and a maximum of 12,826. False or uncertain genotype calls can be removed using the filtering function in `BCFtools`.

After removing the loci with a fraction of missing genotypes that are larger than 20%, minor allele frequency less than 5% or major allele frequency larger than 95%, and Mendelian inconsistencies, the number of SNPs identified for each individual is shown by the green dots in Figure 3.4. The number of SNPs identified from extremely low-depth GBS data was greatly reduced after quality control, with a mean of 18,620, a median of 18,637, a minimum of 9,618 and a maximum of 34,614. In contrast, most of the reference SNPs are likely to be real variants. The number of filtered SNPs in the reference SNPs data has a mean of 9,631, a median of 9,729, a minimum of 7,397 and a maximum of 11,008.

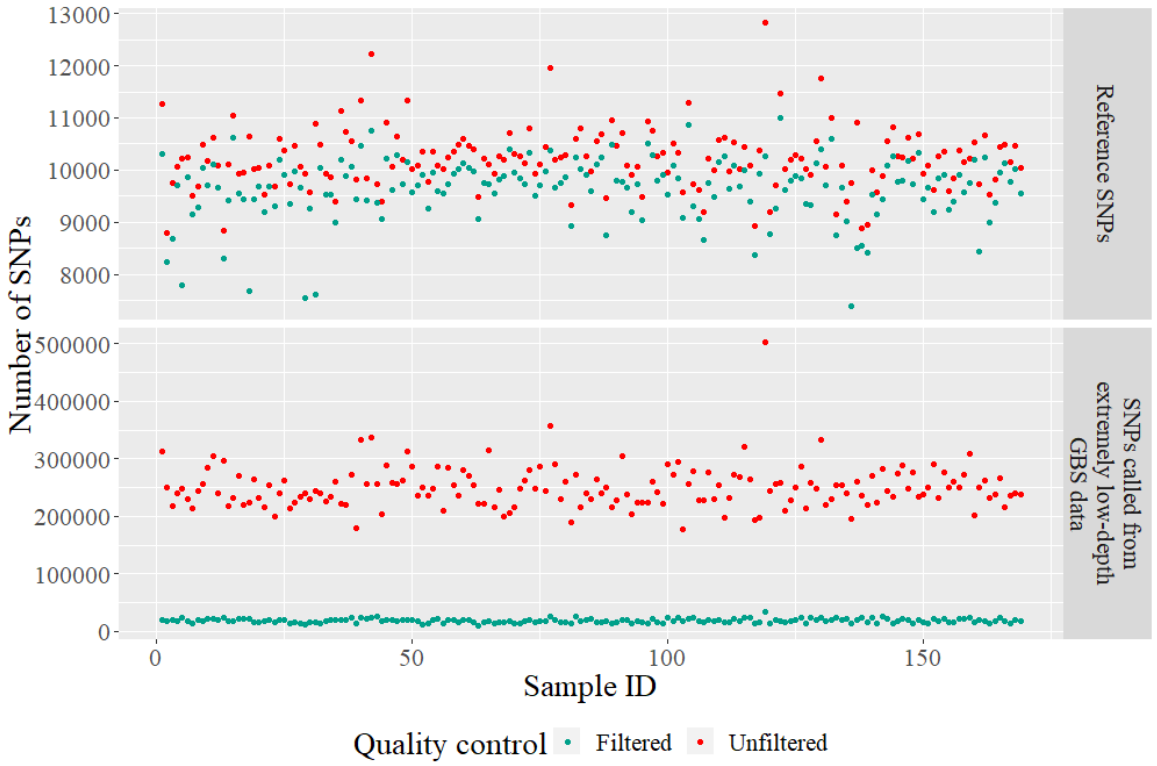

Fig. 3.4 Number of SNPs before and after quality control.

Furthermore, genetic variants with either too low depth or too high depth should also be removed because they are likely to be the consequence of aligning errors in sequence mapping. This is not a concern in reference SNPs because they are either heterozygotes of Jane, who was sequenced with very high read depth, or heterozygotes of Richard Henry with very high quality (QUAL=99). Note that QUAL is the quality score of SNP that is defined as $-10\log_{10}P(\text{genotype call is wrong})$, i.e., the higher the QUAL, the the lower the probability that the genotype call is wrong.

On the other hand, when the variants are called from extremely low-depth GBS data, simply excluding all the variants with very low depth is infeasible as the majority have depth below 5. To rule out potentially false-positive calls in variants called from extremely low-depth GBS data, I also removed variants with quality lower than the threshold (QUAL=15, indicated by the black vertical line in Figure 3.5), where the quality threshold was chosen by comparing the quality distribution between variants existing in both extremely low-depth data and complete data and variants existing in extremely low-depth data only.

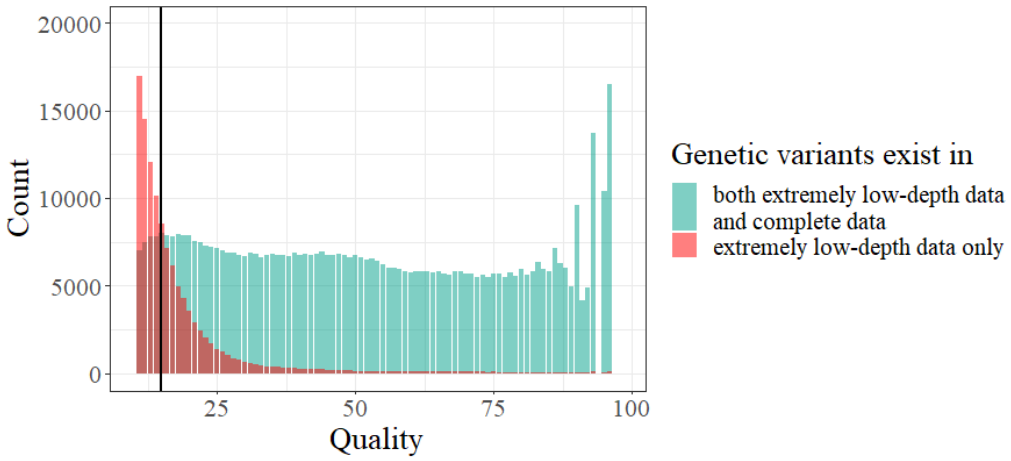

Fig. 3.5 The quality distribution of variants exist in both extremely low-depth data and complete data and variants exist in extremely low-depth data only. The black line (QUAL=15) is chosen to be the threshold and variants with quality lower than the threshold are removed.

3.4 Subject selection and kinship estimation

There are several options for subject selection, including pedigree-based approaches `ExomePicks` and `GIGI-Pick`, and relatedness-based approach `PRIMUS`. I applied these approaches to both low-depth GBS and reference SNP data, and compared the performance of different subject selection strategies based on genotype imputation accuracy in Section 3.5. I found that `ExomePicks` and `GIGI-Pick` work for general pedigrees, whereas selection strategies such as `PRIMUS` that rely on pairwise kinships and can be sensitive to the pedigree structure.

For outbred populations (e.g. humans), pairwise kinships of first-degree relatives should be distinguishable from pairwise kinships of unrelated individuals. Given low-depth GBS data with the average read depth ranges from 2-fold to 10-fold (1-fold to 5-fold per allele), I estimated pairwise kinships of 169 kākāpō using a marker-based approach proposed by Weir & Goudet [154]. In Weir & Goudet's approach, the kinship coefficient for a pair of individuals is estimated by the proportion of alleles carried by the pair of individuals that are identical in states, relative to the average matching for all pairs of distinct individuals. As this is a relative estimate, a negative pairwise kinship means that the pair of individuals shares less alleles than the population average.

As shown in Figure 3.6, pairwise kinship estimation becomes more reliable as the average read depth increases. However, regardless of the read depth, there is always an overlap between the estimated pairwise kinships of first-degree relatives and the estimated pairwise kinships of Stewart Island founders. The overlapping suggests that the Stewart Island

founders are related, resulting from the sharp historic reduction in the kākāpō population size. For truly unrelated individuals, the depth of sequencing reads should not lead to a different conclusion of their relationship. For example, the Fiordland kākāpō named Richard Henry is genetically different to the Stewart Island kākāpō because they were separated by the sea. In Figure 3.6, the relatedness between Richard Henry and Stewart Island founders is clearly lower than the relatedness between parents and offsprings even in the inference from extremely low-depth GBS data (2-fold). Unfortunately, Richard Henry is a special case and the pairwise kinships estimated by extremely low-depth GBS data provides no information on the relationships between the rest of the kākāpō. Therefore, it is very challenging to reconstruct the pedigree structure based on the pairwise kinships estimated from extremely low-depth GBS data.

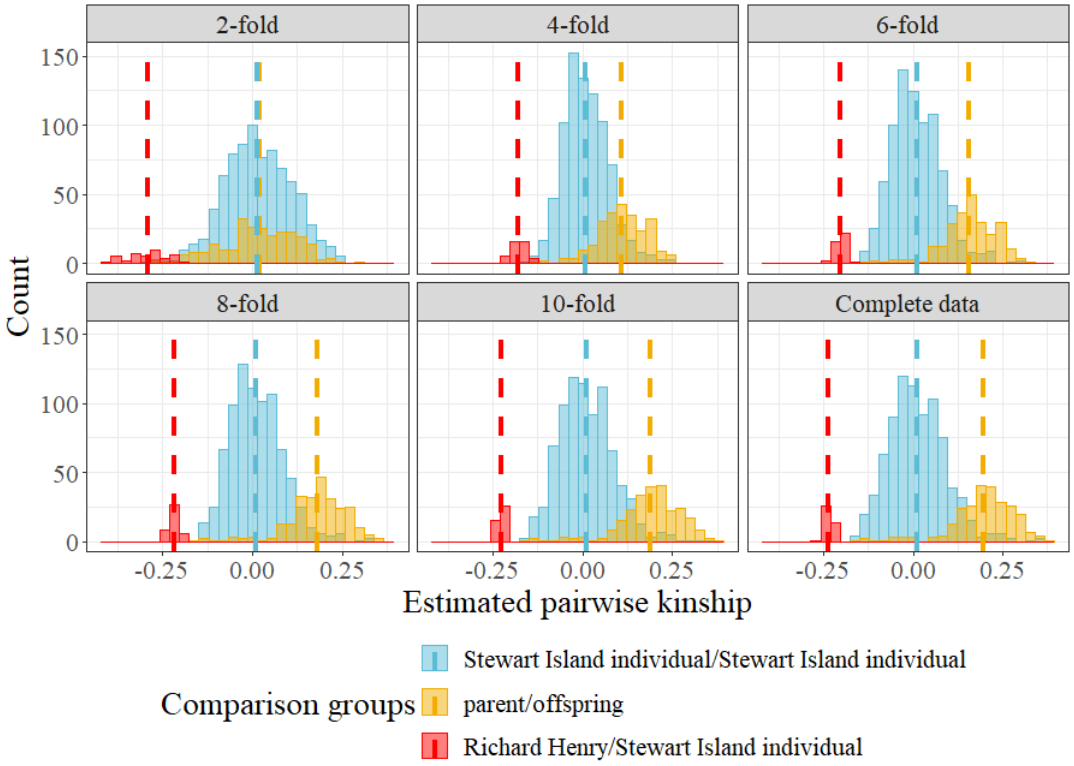

Fig. 3.6 Pairwise kinships inferred by GBS data with different depths.

I also estimated pairwise kinships from the reference SNPs data. In contrast to extremely low-depth GBS data, estimated pairwise kinships based on 20k reference SNPs are very close to that based on the complete data (Figure 3.7). For kākāpō, the reference SNPs from Jane alone do not capture the unique features of Richard Henry (Figure 3.7). After including genetic markers from Richard Henry in the reference SNPs, the relatedness inference based on low-density genotype is almost as good as the inference based on the complete data.

For kākāpō or other endangered inbred species, relatedness-based selection strategies may not be ideal regardless of the quality of sequencing data, as it is very difficult to reconstruct the pedigree correctly. When this is not the case, reference SNPs could be a better choice than low-depth GBS data for relatedness inference.

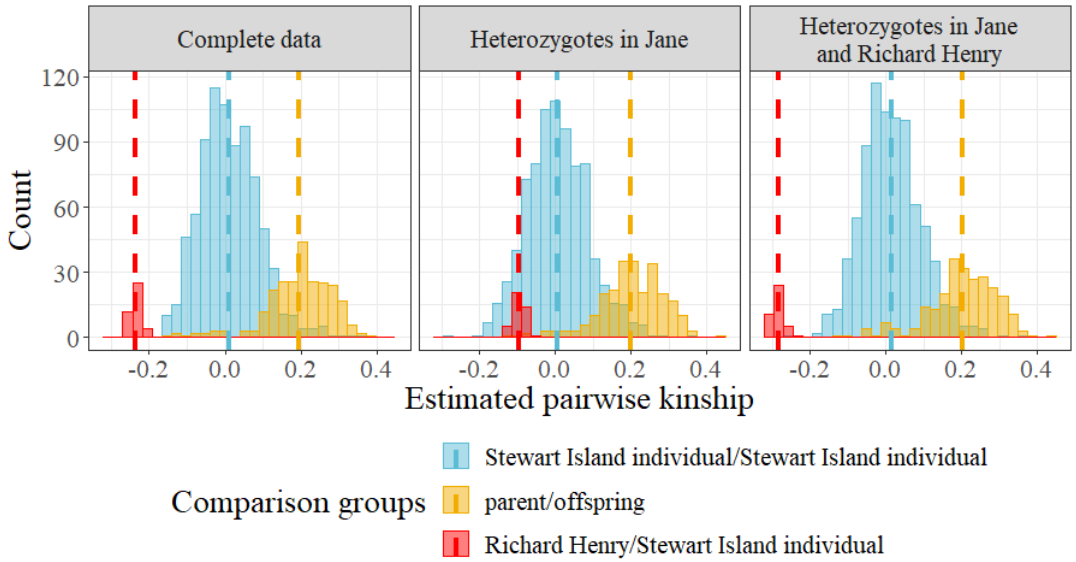

Fig. 3.7 Pairwise kinships inferred by reference SNPs (heterozygotes in Jane only and 16000 heterozygotes in Jane and 4000 heterozygotes in Richard Henry)

3.5 Genotype imputation

To the best of my knowledge, `GIGI` is the only well-known family-based imputation program that handles large pedigrees, and `GIGI2` [146] is a new program that implements `GIGI`'s imputation approach but much faster and using less memory. Considering the complexity of the kākāpō pedigree, I used `GIGI2` for family-based genotype imputation in this chapter. As another option, I used the population-based method `BEAGLE` to impute the missing genotype in the low-density kākāpō genotype data. In this case, kākāpō were treated as unrelated. The two subsections in this section discuss family-based approaches first and then population-based approaches in phasing and genotype imputation respectively, for the kākāpō study.

For the kākāpō study, I performed genotype imputation on chromosomes S1, S9 and S26, which were selected as representations of macrochromosomes, intermediate chromosomes and microchromosomes respectively. Microchromosomes are tiny chromosomes (less than 20 Mb) that are typically observed in the karyotype of birds, fish, reptiles and amphibians. The

chromosomes that are larger than microchromosomes are called macrochromosomes (greater than 40 Mb) and intermediate chromosomes (between 20 Mb and 40 Mb). In contrast to macrochromosomes, bird microchromosomes have a higher and unevenly distributed recombination rate (higher near chromosome ends), higher substitution (mutation due to substitution of nucleotides) rate, and consequently a higher gene density [9, 10, 28, 46, 133]. The small size of the bird microchromosome also imposes challenges to genome assembly (the reconstruction of the original DNA sequence from short DNA segments) [46], which may result in a high error rate in the downstream analyses.

3.5.1 Phasing

In this study, haplotyping in the family-based approach used by `GIGI2` is done by `gl_auto`, which locates genetic recombinations by sampling IVs given low-density genotypes and a pedigree. The quality of the low-density genotypes that are used to sample IVs has a direct impact on the imputation performance. More specifically, it is essential to have consistency between IVs and the observed genotypes at the positions of dense markers. Even with strict quality control, low-depth GBS data is not ideal for IV sampling because of the high error rate, and the genotyping errors lead to shorter haplotypes compared to the actual haplotypes. For example, when I use reference SNPs instead of SNPs called from low-depth GBS data, `GIGI2` is able to impute an average (over chromosomes) of 34% more loci, and the accuracy increases by an average of 7% among the imputed genotype given the dense genotype of 54 kākāpō selected by `GIGI-Picks`. Note that this result is not influenced by subject selection strategy as shown in the next section.

Reference haplotypes such as HapMap for the human population are necessary for genotype imputation of unrelated individuals. Since there are no reference haplotypes of the kākāpō population due to its small population size, the genotypes of reference kākāpō need to be phased. Among the common phasing tools, `WhatsHap` gives the best haplotypes for the kākāpō genomics data [57]. The only drawback of `WhatsHap` is that the algorithm is very time-consuming. In particular, for pedigree-awared phasing, the required computational resources increase exponentially as pedigree size increases. For example, `WhatsHap` can take 6 to 7 days to phase the genotypes of the simplest 3-generation family in the kākāpō pedigree (using a single CPU on the Broadwell E5-2695v4, dual socket 18 cores per socket nodes, 128GB RAM, with the individual processors being Intel Xeon E5, 2.1 GHz).

Since all selections involve more than two generations and a high level of inbreeding, the genotypes of reference kākāpō selected by `ExomePicks`, `GIGI-Pick` and `PRIMUS` are phased using reads only in this study to avoid extremely complex computation. In order to incorporate pedigree in phasing and complete the computation in a reasonable amount of time,

I selected 27 kākāpō from 9 trio families as reference individuals because parents-offspring is the simplest and the most informative relationship for phasing. Moreover, the parents from the chosen trio families are unrelated in the pedigree so that there can be as many distinct haplotypes as possible in the reference panel.

Besides achieving high accuracy, we also want to phase and impute as many variants as possible. For `gl_auto`, this only depends on the error rate of variants calling, whereas for `WhatsHap`, it also depends on the length of sequencing reads. `WhatsHap` is only able to phase heterozygotes that are covered by reads with another heterozygote, and kākāpō sequenced with short-read sequencing technology, which is less ideal than long reads. For the kākāpō data, this greatly limits the number of variants that can be phased. Consequently, `WhatsHap` cannot guarantee that a genetic variant is phased for all reference individuals, and such variants are removed from the reference set as required by `BEAGLE`.

Taking the largest chromosome S1 as an example, more than 99% of the variants can be imputed using the `GIGI` in combination with `gl_auto`, approximately 20% of the variants can be imputed using `BEAGLE` in combination with `WhatsHap` (reads only) and 46% of the variants can be imputed using `BEAGLE` in combination with `WhatsHap` (reads and pedigree), based on the genotypes of 27 reference kākāpō.

As a conclusion, the family-based approach `GIGI` in combination with `gl_auto` has a clear advantage in situations where the population has no reference haplotypes available and is not sequenced with long-read sequencing technologies. The following section provides more details on the accuracy of family-based imputation and population-based imputation with different selection strategies and different numbers of reference kākāpō.

3.5.2 Imputation accuracy

For family-based imputation, I use both the proportion of correctly imputed genotypes and Pearson's squared correlation between observed and imputed genotypes (R^2) to assess the imputation accuracy as they are two common metrics for imputation accuracy evaluation. Previous studies have shown that the proportion of correctly imputed genotypes overestimates accuracy for rare variants [113], and R^2 is the best measure of accuracy [147]. However, it might not be a good idea to rely only on R^2 because kākāpō is a highly inbred species (i.e., no variation between individuals at a large number of loci). Moreover, the set of individuals with observed genotypes is not always the same as the set of individuals whose genotypes can be imputed.

Figure 3.8 shows the positive relationship between the proportion of correctly imputed genotypes and the number of individuals with dense genotypes on the three chromosomes. It is also not surprising to see that the proportion of correctly imputed genotypes tends to

increase as the chromosome size increases since the genotyping error rate is higher on the microchromosomes. Regardless of chromosomes and selection strategy, roughly 80% of the genotypes can be imputed correctly given 14 kākāpō with dense genotype, and roughly 92% of the genotypes can be imputed correctly when dense genotypes are available for half of the kākāpō population.

In family-based imputation, kākāpō with dense genotypes are selected by "per nuclear family" output in `ExomePicks`, as recommended by the author [36], and the genome-wide coverage metric in `GIGI-Pick`. Cheung et al. [33] showed that `GIGI-Pick` outperforms `ExomePicks` by leveraging the uncertainty in the pedigree's inheritance pattern on both rare variants and SNPs for a pedigree with no inbreeding. However, the same result does not seem to hold for inbred populations with low genetic heterogeneity, such as kākāpō. In terms of proportion of correctly imputed genotypes, `GIGI-Pick` and `ExomePicks` have similar performance to random selection overall, but pedigree-based selections are slightly more informative than random selection when a large number of individuals are selected (Figure 3.8). By looking into R^2 at different MAF intervals (Figure 3.9), `GIGI-Pick` and `ExomePicks` lead to higher proportions of correctly imputed genotypes compared to random selection because they have higher R^2 for common variants (Figure 3.9).

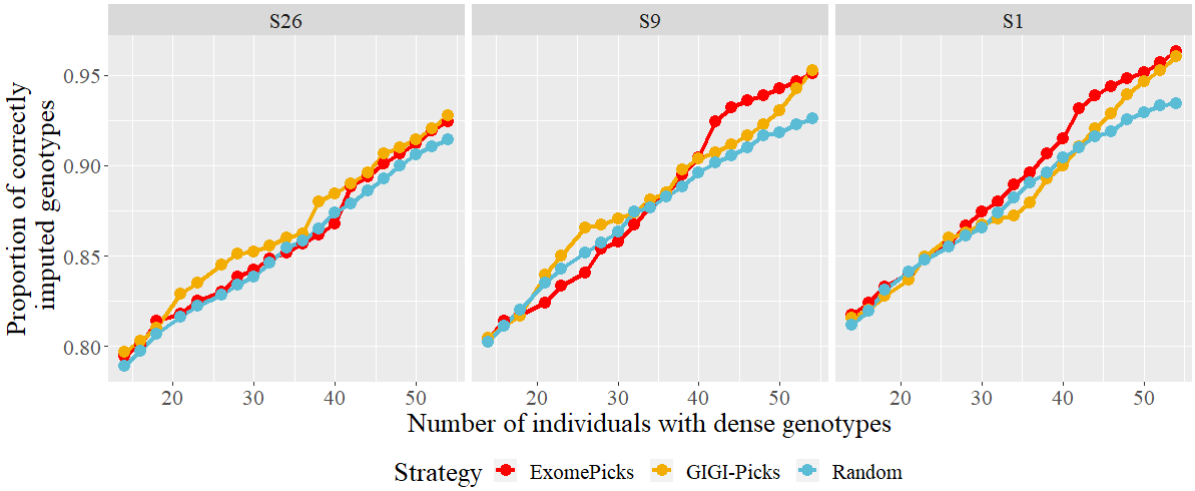

Fig. 3.8 Difference in the imputation performance of `GIGI2` with different selection strategies: the proportion of correctly imputed genotypes on chromosome S1, S9 and S26 given low-density genotype data (reference SNPs). The proportion of correctly imputed genotypes for random selection is the average proportion of correctly imputed genotypes over ten different random selections.

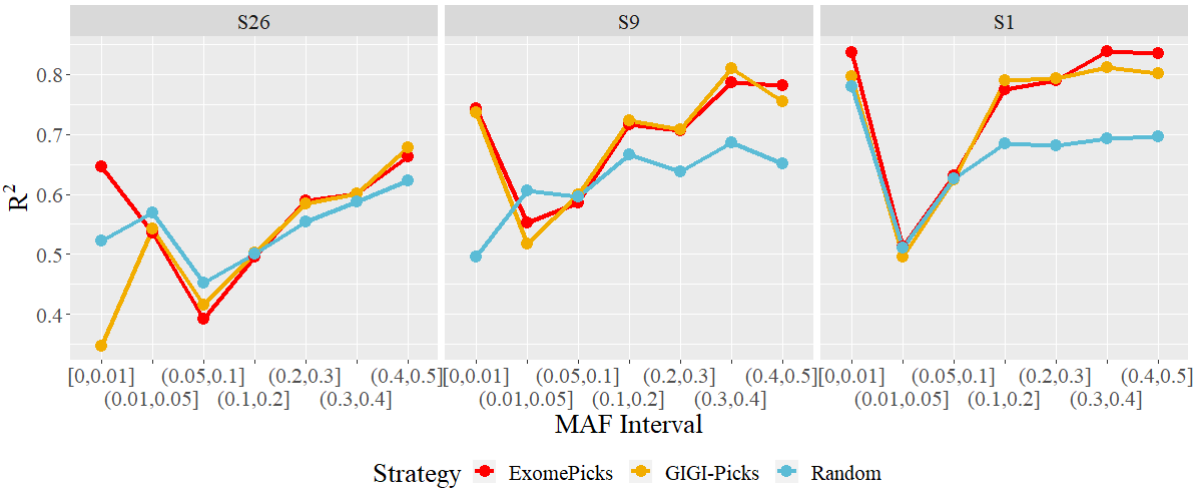

Fig. 3.9 Difference in the imputation performance of `GIGI2` with different selection strategies: Pearson's squared correlation between observed and imputed genotypes (R^2) on chromosome S1, S9 and S26 given 54 kākāpō with low-density genotype data (reference SNPs). R^2 for random selection is taken to be the average value over ten different random selections.

As mentioned at the beginning of section 3.5.1, the error rate of low-density genotypes greatly affects the accuracy of family-based imputation. This is also true for population-based imputation. For example, when reference SNPs are used instead of SNPs called from low-depth GBS data, `BEAGLE` is able to correctly impute 22% more missing genotypes on chromosome S1 and 11% more missing genotypes on chromosome S9 and S26 given the dense genotypes of 27 kākāpō selected by `PRIMUS`.

Another main factor influencing the accuracy of population-based imputation is the phasing method. In Figure 3.10, the imputation accuracy increases by 13% on average when pedigree information is incorporated in phasing, whereas the imputation accuracy improves less than 5% by doubling the number of reference kākāpō when genotypes of the reference kākāpō are phased using linkage disequilibrium information only. `PRIMUS` is an exception in the selection strategy where increasing number of reference individuals does not always result in higher imputation accuracy, because a smaller set of maximumly unrelated individuals is not necessarily a subset of a lager set of maximumly unrelated individuals.

The effect of reference individuals selection is also relatively small for the kākāpō data as different selections of reference kākāpō makes little differences in imputation accuracy. Although `PRIMUS` is designed to select the maximumly unrelated subset of individuals so that the reference genotype contains the most variants in the population, it does not show any advantages compared to random selection or even family-based approaches `ExomePicks` or `GIGI-Pick` in the population-based imputation of kākāpō genotype. This is likely to be the same problem as in family-based imputation: a different selection of reference kākāpō would

result in far fewer changes in reference haplotypes compared to the human population due to its low genetic heterogeneity.

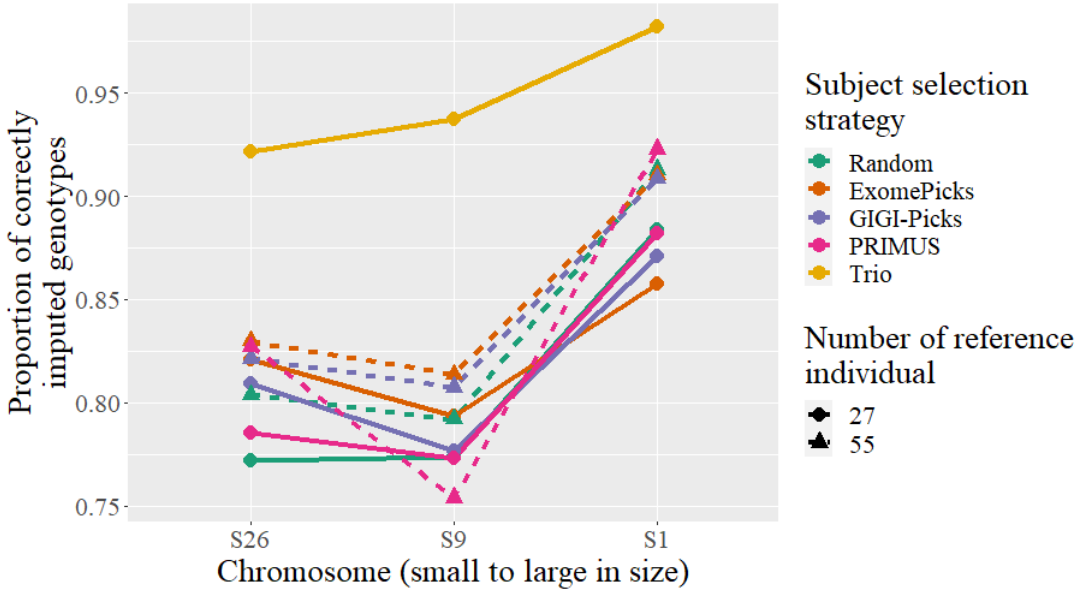

Fig. 3.10 Difference in the imputation performance of `BEAGLE` with different selection strategies and different number of reference individuals: the proportion of correctly imputed genotypes on chromosome S1, S9 and S26 given low-density genotype data (reference SNPs) and dense genotype of the $\sim$ 27 (represented by dot and solid line)/55 reference individuals (represented by triangle and dashed line). The proportion of correctly imputed genotypes for random selection is the average proportion of correctly imputed genotypes over ten different random selections. Note that the 27 reference individuals (composed of 9 trio families, and the results are represented by yellow solid line at the top of the figure) are phased using linkage disequilibrium (LD) and pedigree information, whereas the reference individuals selected by other strategies are phased using LD information only.

3.6 Summary

This chapter demonstrated the process of genotype imputation for endangered species kākāpō and investigated the factors that affect imputation performance. The imputation performance on the kākāpō genome data is affected the most by two factors.

First, when the target population has no reference haplotypes available and is not sequenced with long-read sequencing technologies, family-based genotype imputation that utilizes the pedigree information allows more missing genotypes to be imputed and correctly imputed than population-based genotype imputation that ignores the pedigree information (as discussed in Section 3.5.1). For population-based genotype imputation that requires

pre-phasing of the dense genotype, the use of pedigree information in phasing also leads to substantial improvement in the number of missing genotypes that can be imputed and the proportion of correctly imputed genotypes. For homogeneous species with complex population structures, using the relationship between individuals greatly reduce the uncertainty in phasing and imputation, hence allowing more missing genotype to be imputed with higher accuracy.

Second, among the potential factors that could affect imputation accuracy, the quality of low-density genotypes is the most important factor that affects the imputation quality (as discussed in Section 3.5.1 and Section 3.5.2). For the family-based imputation method `GIGI`, the quality of low-density genotypes affects the number of sampled IVs that are consistent with the observed genotypes, hence affecting the number of markers with missing genotypes that can be imputed and the proportion of correctly imputed genotypes. For the population-based imputation `BEAGLE`, the quality of low-density genotypes also affects the haplotype phase inference and hence the imputation accuracy.

The kākāpō study also found that the selection of reference individuals is unimportant in the genotype imputation for inbred species such as kākāpō. This is true for both family-based and population-based imputation (as discussed in Section 3.5.2). In family-based imputation for kākāpō, `GIGI-Pick` no longer has an advantage compared to `ExomePicks` by leveraging the uncertainty in the pedigree's inheritance pattern as it does for pedigrees without inbreeding. In population-based imputation for kākāpō, `PRIMUS` does not lead to higher imputation accuracy than other strategies by selecting the maximumly unrelated subset of individuals (hence enriching the reference haplotypes). A sensible explanation for this is the low genetic heterogeneity of inbred species, in other words, different selections provide a similar amount of genome information.

In summary, family-based genotype imputation can impute most missing genotypes in the low-density kākāpō genotype data with high accuracy. While imputed genotypes enable GWA studies with a reduced cost, the use of pedigree information greatly increases the computation time, and the computation time grows exponentially as the pedigree size increases. Furthermore, genotype imputation is more complicated for endangered species compared to humans, livestock and other model organisms because data required for imputation, such as the genetic map and reference haplotypes, are not available like the well-studied species. Since no genetic map is available, I assumed the recombination rate varies between chromosomes but remains constant within each chromosome in this study. This may cause a loss in phasing and imputation quality because many birds including kākāpō have low recombination rates at the middle of the chromosomes but high recombination rates close

to the end of the chromosomes. Therefore, the phasing accuracy could be improved once a fine-scaled genetic map of kākāpō becomes available.

In the next chapter, I will explore the possibility of inferring the population parameter required in GWA studies using the dense genotypes only available for a subset of individuals, and propose a method that is relatively fast and straightforward to implement.

Chapter 4

Two-phase sampling

Chapter 3 demonstrated the process of predicting the missing genotype in the low-density kākāpō genotype data using the dense genotype of a subset of individuals. An alternative approach to deal with the problem that only a small fraction of the sample can be resequenced is to treat it as a missing data problem, and use the subsample data to estimate the same parameters in mixed-effects models as would be estimated with the complete sample data. This chapter starts with an introduction on the incomplete data problem in section 4.1, followed by the common two-phase sampling design in section 4.2 and methods for model inference under two-phase sampling in section 4.3.

4.1 Missing data problem

Rubin [121] classified missing data problems into three categories according to their missingness mechanism.

If there is no relationship between the missingness and both observed and unobserved variables of the data, then every data point has the same probability of being missing, and the data are said to be missing completely at random (MCAR). In this case, the missing data can be simply removed from the data, and the parameters estimated with maximum likelihood methods are unbiased. However, data are rarely MCAR in practice.

A more general and realistic assumption in maximum likelihood estimation (MLE) is the data are missing at random (MAR), which allows the missingness to be depending on observed variables. In other words, data points in the same group defined by an observed variable have the same probability of being missing. Hence, a random subset can be obtained from the underlying population if we can control for the observed variables. MAR is said to be ignorable because maximum likelihood inference can ignore the missingness mechanism

when the missingness distribution is independent of the outcome distribution conditional on the observed variables [88].

When neither MCAR or MAR holds, the data are not missing at random (NMAR) which means the missingness depends on unobserved variables. Such missingness is nonignorable and the MLEs of parameters of interest are very likely to be biased if the missing data mechanism is not modeled. NMAR is the most complex case in the three missingness mechanisms, especially when the probability of being missing varies for unknown reasons.

If the missingness is induced by selecting a particular sample from the population, then missingness mechanism depends on the sampling design. Two-phase samples may be MAR given only the observed data when the missingness depends on either the the explanatory variable or the response variable [110], but NMAR when missingness depends on both of them. Another example of NMAR data is when individuals are selected by non-probability sampling from the population using non-random methods based on various criteria such as convenience. While non-probability sampling allows easy and cheap data collection, it is impossible to estimate the sampling probability and identify possible bias. On the contrary, probability sampling is usually more complex and expensive, but all individuals have known non-zero sampling probabilities as they are randomly selected from the population. For a probability sample, the sampling weights, which are the inverse of sampling probabilities, can be calculated, hence we are able to correct the bias caused by the missingness.

4.2 Two-phase sampling

In genetic association studies, the goal is to identify the genetic variants that are associated with the diseases or traits. It is usually relatively easy and affordable to obtain the disease status or measure the trait values such as blood pressure, but cost-prohibitive to obtain high-density genetic variant data at sample sizes sufficiently large for association studies, despite the decreasing cost of DNA sequencing. Consequently, association studies typically find nearby genetic markers rather than the functional variants that are responsible for the trait. One approach to find the functional variants is to obtain the complete sequencing data of the genome region that shows a signal of association. However, resequencing with high density is usually limited to only a subsample by the budget constraint.

Neyman [103] proposed two-phase sampling or double sampling, which is a cost-efficient strategy for data collection. In phase I, a large sample is collected from the target population and relatively cheap information is obtained for all individuals. In phase II, a subsample is selected based on the phase I information, and the expensive variable is measured for the subsample.

A two-phase sampling design that depends on the response variable, also referred to as outcome-dependent sampling, is a primary strategy to increase the power of an association study. For binary diseases with low prevalence, simple random sampling is not efficient as it will only include a small number of cases. Given the disease status of all individuals in phase I, outcome-dependent design (case-control design) samples all the cases with the disease and a small fraction of the controls (individuals without disease), and the variable of interest is measured on the phase II sample. By oversampling the rare outcomes, outcome-dependent sampling results in a subsample in which more individuals carry the functional genetic variants and hence greatly increases the power of the association study. It has also been shown for continuous outcomes that individuals with extreme trait values provide more information than randomly sampled individuals [149, 161].

Since the aim of resequencing is to find the functional variants associated with the trait of interest and nearby variants tend to be inherited together, it may be helpful to select the subsample based on outcome as well as the genetic markers. In orther words, the idea is to increase the number of possible carriers of the functional variants by oversampling cases or individuals with extreme trait values who carry the highest-signal genetic marker. Such stratified case-control sampling was first proposed by White [155] in 1982, and it is more informative than the standard case-control design by sampling based on both outcome and exposure. For conservation study of kākāpō or other endangered species, the sample size may be too small to detect any association between low-density genetic variants and the trait of interest. In such scenarios, we can select a subsample based on phenotypes and pedigree, which are both available for all kākāpō. The pedigree information of endangered species is typically known by their recovery programme and it is closely related to the genetic data.

For the missingness mechanisms in section 4.1, consistent or asymptotically unbiased MLEs can be obtained using a subsample under the framework of two-phase sampling. The next section reviews several common methods for handling missing data under two-phase designs.

4.3 Estimation methods

Let Y denote the outcome variable, X denote the expensive covariate, S denote the inexpensive covariate that is correlated with X, and A denote the auxiliary variable. Under outcome-dependent sampling, Y and A are obtained for the entire phase I sample, whereas X is only measured for the phase II subsample. Let R be the sampling indicator given phase I data, with $R_i = 1$ if the i-th individual is in phase II and $R_i = 0$ otherwise. For parameter

vector θ, the observed-data likelihood is given by

$$L = \prod_{R_i=1} P(y_i|x_i,s_i,A;\theta)P(x_i|s_i,A) \prod_{R_i=0} P(y_i|s_i,A;\theta).$$

If Y is independent of A given X, or $Y|A$ can be integrated out over A (although this is often not the case), the observed-data likelihood can be written as

$$L = \prod_{R_i=1} P(y_i|x_i,s_i;\theta)P(x_i|s_i) \prod_{R_i=0} P(y_i|s_i;\theta),$$

where $P(Y|S;\theta) = \int P(Y|X,S;\theta)P(X|S)dx$ is the measurement error model. The full likelihood is usually complicated to derive and may lead to inconsistent estimates if $P(X|S)$ is not correctly specified. As the imputation approach for missing data problem is discussed in Chapter 2, this section focuses on the common methods for statistical inference under two-phase sampling with a focus on the maximum likelihood approach.

4.3.1 Weighted likelihood

The weighted likelihood is a Horvitz-Thompson (HT)-type estimator [65] that has been widely used because of its simplicity and robustness, and multiple methods have been developed from it [52, 73, 115, 120, 169]. An estimate of the population log-likelihood can be obtained using individuals with complete observed data using sampling weights, which are the inverse sampling probabilities:

$$\ell(\theta) = \sum_{R_i=1} \frac{1}{\pi_i} \log P(y_i|x_i,s_i;\theta),$$

where π_i is the phase II inclusion probability for the i-th individual given phase I data. In general, the HT-type estimator is robust to the bias caused by model misspecification when the sampling weights are correctly specified. However, for mixed models, which requires integration over the random effects, the HT-type estimator is no longer a sum of over all clusters when the sampling design involves within-cluster sampling.

Schildcrout et al. [125] considered linear mixed model with longitudinal data under outcome-dependent sampling, and developed a conditional complete data likelihood that uses phase II individuals only. For cluster-correlated data under two-phase designs, Rivera-Rodriguez et al. [118] proposed a generalized estimating equations approach based on inverse-probability weighting for inference of marginally specified generalized linear models, and a calibrated inverse-probability weighting estimator for improved estimation.

4.3.2 Stabilized weights and generalized raking

The HT-type estimator is known to be inefficient as the phase I information of non-phase II individuals is completely ignored. In order to improve efficiency, several methods have been developed for weight adjustments. When the mean model is correctly specified, Magee [93], Robins et al. [119], Pfeffermann and Sverchkov [108], Skinner and Mason [132] proposed estimates of a function that is chosen to minimise the variation of the sampling weights, and weights divided by the function are referred to as stabilized weights [119].

Generalized raking, also known as calibration of the weights, is a survey sampling technique that improves the efficiency of estimation of population means by adjusting sampling weights based on auxiliary variables [42]. In contrast to stabilized weights, calibration does not make any model assumptions. Under two-phase sampling, calibration works when auxiliary variables are linearly correlated with estimators for the regression parameter of interest [20, 91, 117]. Breslow et al. [21] described a strategy for obtaining an estimation of the correlated auxiliary variable using fully observed variables. Støer and Samuelsen [140] and Rivera and Lumley [117] used a similar procedure for different sampling designs. While a poor choice of the function in stabilized weights can lead to larger variance, the calibrated estimator is always asymptotically more efficient than the uncalibrated estimator [90].

4.3.3 Pseudolikelihood

Pseudolikelihood is an approximation to the full likelihood function in a more tractable form. Breslow and Cain [15] developed a pseudolikelihood method to estimate the parameters in logistic models under two-phase design, with the phase I sample selected by a case-control sampling. Let $Y = \{0,1\}$ be the binary outcome and $S = \{1,\cdots,J\}$ be the stratification variable, N_{ij} be the number of phase I subjects with $Y = i$ and $S = j$, x_{ijk} be the covariate vector that is only measured for the n_{ij} phase II subjects randomly selected from N_{ij}, and logit^{-1} denotes the standard logistic distribution function. Breslow and Cain [15] first obatined an estimate of the log-odds for stratum j by maximizing the pseudolikelihood of the phase I data

$$L_1 = \prod_{i=0}^{1}\prod_{j=1}^{J} P_{ij}^{N_{ij}},$$

with

$$P_{1j} = \Pr(Y = i|S = j) = \text{logit}^{-1}\left(\log\frac{N_{1j}}{N_{0j}}\right) \text{ and } P_{0j} = 1 - P_{1j},$$

and then substituted it into the pseudolikelihood of phase II data

$$L_2 = \prod_{i=0}^{1} \prod_{j=1}^{J} \prod_{k=1}^{n_{ij}} p_{ijk}^{n_{ijk}},$$

with

$$p_{1jk} = \text{logit}^{-1}\left(\log \frac{N_{0j} n_{1j}}{N_{1j} n_{0j}} + \beta_0 + x_{1jk}^T \beta_1\right) \text{ and } p_{0jk} = 1 - p_{1jk},$$

where $\log \frac{N_{0j} n_{1j}}{N_{1j} n_{0j}}$ is an offset term for correction of sampling bias. Schill et al. [126] assumed the same model and gave an alternative estimator by jointly fitting logistic models to phase I and phase II data.

When the missingness is independent of the discrete outcome Y, Pepe and Fleming [106] and Carroll and Wand [30] developed estimated likelihood methods by replacing the unspecified marginal distribution function $P(X)$ with an consistent estimator. The method proposed by Pepe and Fleming [106] requires a discrete covariate, however that of Carroll and Wand [30] allows continuous covariate using kernel density estimators. Weaver and Zhou [152] extended their work and proposed a maximum estimated likelihood estimator for continuous outcome Y under outcome-dependent sampling design. Chatterjee et al. [32] also developed an efficient pseudoscore estimator that accepts aither a discrete or a continuous outcome, and is similar in nature to the estimated likelihood proposed by Weaver and Zhou [152].

In order to accommodate the potential correlation for inividuals and unknown covariates within a cluster, Xu and Zhou [162] proposed a semiparametric estimated likelihood estimator for linear mixed model with cluster random effects. They considered an outcome-auxiliary-dependent sampling design, and accounted for the sampling bias through a nonparametric estimator of the conditional cumulative distribution function of $F(X|S,A)$. Most of the other pseudolikelihood approachs for mixed model inference use sampling weights to adjust the bias caused by complex sampling. Pfeffermann et al. [107] developed a pseudolikelihood for linear mixed model that employed a probability-weighted iterative generalized least squares algorithm. Rabe-Hesketh and Skrondal [112] proposed a sample weighted pseudolikelihood for generalized linear mixed models. Let i be the subscript for the sampled clusters s, π_i be the sampling probability for cluster $s(i)$, $\mathbf{u}_i$ be the cluster random effect, j (and k) be the subscript for individuals sampled within a cluster, and $\theta = (\theta_y, \theta_u)$ be the set of model

parameters. The sample weighted log-likelihood is given by

$$\hat{\ell}(\theta) = \sum_{i \in s} \frac{1}{\pi_i} \ell_i(\theta), \tag{4.1}$$

where

$$\ell_i(\theta) = \log \int \exp\left(\sum_{j \in s(i)} \frac{1}{\pi_{j|i}} f(y_{ij}|x_{ij}, \mathbf{u}_i; \theta_y) \right) g(\mathbf{u}_i|\theta_u) d\mathbf{u}_i, \tag{4.2}$$

with $f(\cdot)$ and $g(\cdot)$ being the density functions of the response variable y and random effect $\mathbf{u}$.

As shown in Eq 4.2, the sample weighted log-likelihood (Eq 4.1) is not design-unbiased because the sampling weights appear in the weighted estimating equation in a non-linear form. Rao et al. [114] and Yi et al. [165] overcame this problem by considering the pairwise composite likelihood that is constructed from the sum of the pairwise likelihood. With the same notation, the idea of pairwise log-likelihood is to replace the log-likelihood for each cluster in Eq 4.2 by the sum of all possible pairwise log-likelihood:

$$\hat{\ell}(\theta) = \sum_{i \in s} \frac{1}{\pi_i} \sum_{\substack{j<k \\ j,k \in s(i)}} \frac{1}{\pi_{jk|i}} \ell_{jk|i}(\theta), \tag{4.3}$$

where π_i is the sampling probability for cluster i, and $\pi_{jk|i}$ is the probability of both individual j and k are sampled given cluster i is sampled, and

$$\ell_{jk|i}(\theta) = \log \int f(y_{ij}|x_{ij}, \mathbf{u}_i, \theta_y) f(y_{ik}|x_{ik}, \mathbf{u}_i, \theta_y) g(\mathbf{u}_i|\theta_u) d\mathbf{u}_i. \tag{4.4}$$

Huang's PhD thesis [68] extended their work by establishing consistency and asymptotic normality of the weighted pairwise likelihood estimator under two situations: (1) when the sampling clusters (primary sampling units in the sampling design) are not the same as the model clusters that define the random effect; (2) when the random effects are correlated.

4.3.4 Full likelihood

Maximum likelihood is generally more efficient than weighted likelihood and pseudo-likelihood because it makes more assumptions, however it is more complicated to implement because the nuisance parameter is often treated nonparametrically and may involve a high-dimensional integral.

Under two-phase sampling with a simple random sampling in phase I, Scott and Wild [128] proposed a maximum likelihood estimation method for discrete outcome Y and explanatory variable X, which can be substantially more efficient than pseudolikelihood. However, the number of nuisance parameters increases rapidly as the number of levels in X increases, and the method becomes unfeasible. To solve this problem, Scott and Wild [129] developed a semiparametic maximum likelihood method that obtains the maximum likelihood estimates by iterating the pseudolikelihood procedure and updating the offset parameter. When the phase I sample is selected by a case-control sampling, Breslow and Holubkov [18] derived a full maximum likelihood estimator, and showed their method is equivalent to Scott and Wild's approach [129] with phase I data being a simple random sample.

For continuous outcomes with standard outcome-dependent sampling, Zhou et al. [171] derived a semiparametric likelihood as a product over a number of mutually exclusive intervals of the outcome. Then they profile out the likelihood to obtain the empirical likelihood and maximize the profile likelihood over all distributions whose support contains the observed X values.

Whittemore [156] and Zhao et al. [168] proposed maximum likelihood methods for case-control family design, which is also referred to proband design, that identifies the relatives of each proband (the first person in a family who had diagnosis of diseases) selected in the case-control study and records their disease status and covariates. They allow correlated binary data by considering the conditional distribution of relatives' outcome given the proband outcomes. Neuhaus et al. [100] considered an extension to the case-control family design, and developed a semiparametric maximum likelihood method that applies to any designs if the population of families can be divided into a finite number of strata.

For the semiparametric maximum likelihood approach discussed above, the difficulty in implementation is that the nuisance parameter is treated nonparametrically. Neuhaus et al. [101] extended their work in [100] to fit generalized linear mixed models that allows family-specific random effects. For mixed models, the computation of the marginal probabilities requires integration of the conditional probabilities over the distribution of the random effect. Therefore, the full likelihood becomes substantially more complicated when the clusters are large or when the sampling probabilities depend on outcomes within clusters [101].

Chapter 5

Linear mixed models under two-phase sampling

The most common approach for fitting linear mixed models is maximum likelihood estimation (MLE) and it is implemented in a number of software such as the R [111] package `lme4` [11] and its extension `lme4qtl` [173] which allows user-defined variance-covariance matrices for random effects. An underlying assumption for MLE using the sample data is that the observations are sampled in a way that they are representative of the whole population. However, random sampling is generally less efficient than outcome- or covariate-dependent sampling designs for model inferences, and MLE methods that do not incorporate the informative sampling design and can therefore lead to biased estimation.

This chapter develops a weighted MLE approach that takes advantage of the fact that the kinship matrix is known for the whole kākāpō population, allowing us to model the population covariance matrix rather than the sample covariance matrix. Since the population kinship structure is often known for endangered species, the proposed weighted MLE approach provides a general solution for fitting linear mixed models in conservation genetics. The proposed method is written as an R package `WLMM` and available on GitHub (https://github.com/zoeluo15/WLMM).

The aim of this chapter is to describe the proposed weighted MLE approach and evaluate its performance under two-phase sampling designs using two case studies. The kākāpō study should be a good example for fitting linear mixed models in conservation study of small inbred populations with complex pedigree. Access to the kākāpō data is via application to the Department of Conservation, who assess requests in partnership with Ngāi Tahu, the kaitiaki (guardians) for this data set. A simulated nuclear family dataset is used to extend the results to larger populations with simple pedigree structures, and it should provide a basis for analysis in large outbred populations, such as humans.

The chapter starts with description of the methods in section 5.1, with the proposed weighted MLE approach in section 5.1.3. Then, section 5.2 provides the model inference under two-phase sampling. Lastly, section 5.4 proves the consistency of the proposed likelihood estimator.

5.1 Methods

5.1.1 The linear mixed model

Consider the following model that describes a continuous phenotype y (e.g. length, height),

$$y = X\beta + Zu + e,\ u \sim \mathcal{N}(0, \sigma_g^2 \Phi) \text{ and } e \sim \mathcal{N}(0, \sigma_e^2 I)$$

where N is the population size, y is a $N \times 1$ vector for the phenotype values, X is a $N \times 2$ design matrix made up of a intercept column of 1 and a column for genotype, containing the number of copies of alternative alleles for all individual at a particular locus, i.e.,

$$X_{i2} = \begin{cases} 0, & \text{if the genotype of } i\text{-th kākāpō is } aa, \\ 1, & \text{if the genotype of } i\text{-th kākāpō is } Aa, \\ 2, & \text{if the genotype of } i\text{-th kākāpō is } AA, \end{cases}$$

Here the genetic markers are assumed to be biallelic (one variant allele relative to the reference allele), but the approach would also work for multiallelic markers (markers with more than one variant allele relative to the reference allele). The parameter vector $\beta = (\beta_0, \beta_1)$ comprises the population mean and the coefficient for the genetic fixed effect.

Z is a $N \times N$ design matrix that specifies the structure of the random effect u. The genetic random effect u and the environmental random effect e are $N \times 1$ vectors, independent of one another and follow multivariate normal distributions with σ_g^2 being the genetic variance, $\Phi_{N \times N}$ being the kinship matrix, σ_e^2 being the environmental variance and I being an $N \times N$ identity matrix. In `lme4` [11], the covariance matrix of the random effect is defined by its grouping structure which can be inappropriate in the situation where individuals are genetically related to each other. The R package `lme4qtl` [173] is developed for such situations that allows a customized covariance matrix for the random effect.

In this study, I consider a single genetic random effect with one individual per level of the random effect. That is, the design matrix Z is simply an identity matrix. Therefore, the

model is reduced to the simplest form with a different multivariate normal error termε.

$$y = X\beta + \varepsilon,$$

and the covariance matrix of ε can be written as

$$\begin{aligned}\Xi &= \sigma_g^2\Phi + \sigma_e^2 I \\ &= \sigma^2(h^2\Phi + (1-h^2)I)\end{aligned}$$

where $\sigma^2 = \sigma_g^2 + \sigma_e^2$ is the total phenotypic variance and $h^2 = \frac{\sigma_g^2}{\sigma^2}$ is the heritability which measures the proportion of variation in phenotype y can be explained by genetic variation.

For a linear mixed model, the estimated population log-likelihood implemented in `lme4` and `lme4qtl` [11, 173] is composed of a log-determinant term and a residual sum of squares (RSS) term,

$$\ell(\theta) = -\frac{1}{2}\log|\Xi| - \frac{1}{2}(y - X\beta)^T\Xi^{-1}(y - X\beta), \tag{5.1}$$

where $\theta = \{\beta, \sigma^2, h^2\}$ denotes the parameters of interest. The log-likelihood in Eq 5.1 works when the sample is representative of the population, but it is likely to provide biased estimates otherwise.

5.1.2 Full likelihood

Full likelihood properly accounts for covariance structure and the missing mechanism, however it is more complicated to construct. In this case, we need to deal with the fact that individuals in the pedigree who were not sampled have unobserved genotypes. One solution is to set up a Bayesian model that uses the full likelihood, and samples the unobserved genotypes under the constraint of consistency with observed genotypes. In practice, this is feasible for simple pedigrees, but becomes computationally intensive and time-consuming as the pedigree size and complexity increase.

For the simulated nuclear family data with the outcome-dependent sampling in Design 1, it is relatively straightforward to construct the full likelihood. Code is included below to fit the model using NIMBLE [40]. The model of missing mechanism is simple because $\mathrm{P}(R_i = 1)$ is either 1 or follows a binomial distribution with $p = \frac{n_2}{N - n_1}$ (See Design 1 for the definition of N, n_1 and n_2). The only complicated part is modeling the probability distribution of missing genotypes because individuals are related.

Design 1. *Let N be the population size and n be the sample size, consider the following outcome-dependent sampling.*

Step 1. Always sample the n_1 individuals from the two a% tails of the phenotype distribution;

Step 2. Randomly sample $n_2 = n - n_1$ individuals from the remaining $N - n_1$ individuals, where n is the sample size.

```
code <- nimbleCode({
  beta_0 ~ dunif(-1000, 1000)
  beta_1 ~ dunif(-1000, 1000)
  sigma_g2 ~ dunif(0, 1000)
  sigma_e2 ~ dunif(0, 1000)

  for (i in 1:n_middle) { # sampling
    R[middle[i]] ~ dbinom(size=1, prob=frac) # Sampling indicator of the non-extreme individuals
  }

  for (i in 1:n_rest) { # Impute missing genotype
    indicator[i,1:4] <- c(R[FID[row_idx[i],1]],R[FID[row_idx[i],2]],R[FID[row_idx[i],3]],R[FID[row_idx[i],4]])
    out[i,1:3] <- imputation(geno=x_raw[],FID=FID[,],indicator=indicator[i,1:4],
                             rowx=row_idx[i],colx=col_idx[i],maf=maf)
    x[rest[i]] ~ dcat(prob=out[i,1:3])
  }

  for (j in 1:4) {
    for (k in 1:4) {
      Xi[j,k] <- sigma_g2*kinship[j,k]+sigma_e2*(j==k) # Construct the covariance matrix
    }
  }

  for (i in 1:nf) {
    for (j in 1:4) {
      mu[i,j] <- beta_0 + beta_1*x[FID[i,j]]
    }
    y[i,1:4] ~ dmnorm(mu[i,1:4],cov=Xi[1:4,1:4]) # Fitting linear mixed model
  }
})
```

Nevertheless, it is workable because the nuclear families are independent of one another, so the missing genotypes of individuals from different families can be sampled independently. Since the simulated nuclear families all have the same size and structure, it is possible to list all the possibilities of missingness in the family (as shown in the if-else loop in the following code). Then we can model the probability distribution of missing genotypes based on parent-offspring, sibling relationship and population allele frequency. When the genotype of any relatives is observed, parent and offspring share one of the two alleles, and siblings either share one allele with a probability of 50% or share two alleles or no alleles with a probability of 25%. When none of the relatives has the observed genotype, the genotype frequencies p_a^2, $2p_ap_A$ and p_A^2 are assumed to satisfy the Hardy-Weinberg equilibrium. That

is,

$$p_a^2 + 2p_a p_A + p_A^2 = 1,$$

where p_a and p_A are the frequencies of reference allele and alternative allele in the population.

```
imputation <- nimbleFunction(
  run = function(geno = double(1), FID = double(2), indicator = double(1),
                 rowx = double(), colx = double(), maf = double()) {
    # geno: observed genotypes
    # FID: identies sorted by family
    # indicator: sampling indicator
    # rowx: row index of the individual with missing genotype
    # colx: column index of the individual with missing genotype
    # maf: minor allele frequency
    returnType(double(1))
    p <- numeric(length = 3)
    if (colx==3|colx==4) {
      if (indicator[1]==1&indicator[2]==0) { # Genotype of the first child is observed
        p[1] <- ((2-(geno[FID[rowx,1]]-1))/2)*(1-maf)
        p[2] <- ((2-(geno[FID[rowx,1]]-1))/2)*maf+((geno[FID[rowx,1]]-1)/2)*(1-maf)
        p[3] <- ((geno[FID[rowx,1]]-1)/2)*maf
        return(p)
      } else if (indicator[1]==0&indicator[2]==1) { # Genotype of the second child is observed
        p[1] <- ((2-(geno[FID[rowx,2]]-1))/2)*(1-maf)
        p[2] <- ((2-(geno[FID[rowx,2]]-1))/2)*maf+((geno[FID[rowx,2]]-1)/2)*(1-maf)
        p[3] <- ((geno[FID[rowx,2]]-1)/2)*maf
        return(p)
      } else if (indicator[1]==1&indicator[2]==1) { # Genotypes of both children are observed
        p[1] <- 0.5*((2-(geno[FID[rowx,1]]-1))/2)*(1-maf)+0.5*((2-(geno[FID[rowx,2]]-1))/2)*(1-maf)
        p[2] <- 0.5*(((2-(geno[FID[rowx,1]]-1))/2)*maf+((geno[FID[rowx,1]]-1)/2)*(1-maf))+
          0.5*(((2-(geno[FID[rowx,2]]-1))/2)*maf+((geno[FID[rowx,2]]-1)/2)*(1-maf))
        p[3] <- 0.5*((geno[FID[rowx,1]]-1)/2)*maf+0.5*((geno[FID[rowx,2]]-1)/2)*maf
        return(p)
      } else { # Genotypes of both children are missing
        p[1] <- (1-maf)^2
        p[2] <- 2*maf*(1-maf)
        p[3] <- maf^2
        return(p)
      }
    } else {
      if (indicator[3]==1&indicator[4]==0) { # Genotype of the father is observed
        p[1] <- ((2-(geno[FID[rowx,3]]-1))/2)*(1-maf)
        p[2] <- ((2-(geno[FID[rowx,3]]-1))/2)*maf+((geno[FID[rowx,3]]-1)/2)*(1-maf)
        p[3] <- ((geno[FID[rowx,3]]-1)/2)*maf
        return(p)
      } else if (indicator[3]==0&indicator[4]==1) { # Genotype of the mother is observed
        p[1] <- ((2-(geno[FID[rowx,4]]-1))/2)*(1-maf)
        p[2] <- ((2-(geno[FID[rowx,4]]-1))/2)*maf+((geno[FID[rowx,4]]-1)/2)*(1-maf)
        p[3] <- ((geno[FID[rowx,4]]-1)/2)*maf
        return(p)
      } else if (indicator[3]==1&indicator[4]==1) { # Genotypes of both parents are observed
        p[1] <- ((2-(geno[FID[rowx,3]]-1))/2)*((2-(geno[FID[rowx,4]]-1))/2)
        p[2] <- ((2-(geno[FID[rowx,3]]-1))/2)*((geno[FID[rowx,4]]-1)/2)+((geno[FID[rowx,3]]-1)/2)*((2-(geno[FID[
              rowx,4]]-1))/2)
        p[3] <- ((geno[FID[rowx,3]]-1)/2)*((geno[FID[rowx,4]]-1)/2)
        return(p)
      } else if ((colx==1&indicator[2]==1&indicator[3]==0&indicator[4]==0)|
                 (colx==2&indicator[1]==1&indicator[3]==0&indicator[4]==0)) {
        # Genotypes of both parents are missing but genotype of the sibling is observed
        p[1] <- 0.25
        p[2] <- 0.5
        p[3] <- 0.25
        return(p)
      } else { # Genotypes of the all relatives are missing
        p[1] <- (1-maf)^2
        p[2] <- 2*maf*(1-maf)
        p[3] <- maf^2
        return(p)
      }
    }
  })
```

For complex pedigree structures, computation of the missing genotype probabilities would require the Lander-Green algorithm [81]. In brief, the Lander-Green algorithm calculates: (1) the likelihood of observed genotype data given the pedigree; and (2) the updated conditional likelihood with the missing genotype for an individual set to a specific value. Then, the posterior probability of the missing genotype being that specific value conditional on observed genotype data equals to the ratio of the two likelihoods.

For large complex pedigrees such as kākāpō that cannot be split into smaller pedigrees, the computational time of the Lander-Green algorithm increases exponentially as the pedigree size increases. The Elston-Stewart algorithm [48] may be a better option as its computational time is linear in the pedigree size. However, the computational time of the Elston-Stewart algorithm is exponential in the number of genetic markers. Therefore, full-likelihood methods are not feasible for complicated pedigree, and a different approach is required.

5.1.3 Weighted maximum likelihood estimation

Assuming the kinship matrix is known for the whole population, we are then able to calculate the log-determinant term. Let R be a $N \times N$ indicator matrix with $R_{ij} = 1$ if both the i- and j-th individual are in the subsample and $R_{ij} = 0$ otherwise, and π be a $N \times N$ matrix with the (i, j)-th entry equals to the joint sampling probability of the i- and j-th individual, the residual term in Eq 5.1 can be written as a pairwise sum. Then, the weighted log-likelihood is

$$\hat{\ell}(\theta) = -\frac{1}{2}\log|\Xi| - \frac{1}{2}\sum_{i=1,j=1}^{N}\frac{R_{ij}}{\pi_{ij}}(y - X\beta)_i\left(\Xi^{-1}\right)_{ij}(y - X\beta)_j. \tag{5.2}$$

Note that X_i and y_i are available if $R_i = 1$ for the i-th individual, and Ξ^{-1} is defined and available for all pairs of individuals in the population.

Eq 5.2 is implemented in the R package `WLMM`, and the parameter vector $\hat{\theta}$ is obtained by numerical optimization using the BOBYQA (Bound Optimization by Quadratic Approximation) algorithm [109].

The proposed weighted log-likelihood is different to the pairwise composite likelihood in Rao et al. [114] and Yi et al. [165] in Eq 4.3, which only requires the covariance matrix for *observed* pairs of individuals jk within the i-th cluster.

The RSS estimator

For small datasets, it can be difficult to maximize the log-likelihood in Eq 5.2 over all the model parameters, and the estimation is likely to be inaccurate. Profile likelihood

is a standard approach to reduce the dimension of the likelihood function. Suppose the heritability h^2 is given. We can obtain the MLE of β as it does not depend on σ^2. Let $\widehat{C} = \hat{h^2}\Phi + (1-\hat{h^2})I$ and then the log-likelihood is maximized by

$$\hat{\sigma^2} = \frac{\sum_{i=1,j=1}^{N} \frac{R_{ij}}{\pi_{ij}} (y - X\hat{\beta})_i \left(\widehat{C}^{-1}\right)_{ij} (y - X\hat{\beta})_j}{N}, \tag{5.3}$$

where the numerator is a summation over all pairs of individual i and individual j. Since the numerator in Eq 5.3 is a weighted pairwise sum of squared residuals, it is referred to as a Horvitz-Thompson-type (HT-type) RSS estimator in the remainder of this chapter.

The HT-type RSS estimator can be problematic for small datasets. The diagonal elements of the pairwise sampling probability matrix are large because they are the first-order sampling probabilities, so the diagonal elements of the sampling weight matrix $1/\pi$ are small. Consequently, the HT-type RSS is not guaranteed to be positive because the sum of the negative off-diagonal weighted terms could be larger than the diagonal positive terms. When the RSS is negative, the profile log-likelihood

$$\begin{aligned}
\hat{\ell_p}(h^2) &= -\frac{1}{2}\left(\log|C| + N\log\left(\sum_{i=1,j=1}^{N} \frac{R_{ij}}{\pi_{ij}} (y - X\beta)_i \left(C^{-1}\right)_{ij} (y - X\beta)_j\right)\right) \\
&= -\frac{1}{2}\left(\log|C| + N\log\left((y - X\beta)^T \left(\frac{R}{\pi} \odot C^{-1}\right)(y - X\beta)\right)\right)
\end{aligned} \tag{5.4}$$

is not defined. An alternative RSS estimator is inspired by the Sen-Yates-Grundy (SYG) variance estimator (Sen 1953, Yates and Grundy 1953) of the population total,

$$\mathbb{V}\left[\hat{Y}_{\text{SYG}}\right] = \frac{1}{2}\sum_{i,j\in s} \frac{\text{Cov}(R_i, R_j)}{\pi_{ij}} \left(\frac{y_i}{\pi_i} - \frac{y_j}{\pi_j}\right)^2,$$

where s denote the sample. The log-likelihood with the SYG-type RSS estimator can be written as

$$\hat{\ell}(\beta, \sigma^2, h^2) = -\frac{1}{2}\log|\Xi| - \frac{1}{4(N-1)} \sum_{i=1,j=1}^{N} \frac{R_{ij}}{\pi_{ij}} \left[\left(\Xi^{-\frac{1}{2}}(y - X\beta)\right)_i - \left(\Xi^{-\frac{1}{2}}(y - X\beta)\right)_j\right]^2, \tag{5.5}$$

where $\Xi = \sigma^2(h^2\Phi + (1-h^2)I)$. The SYG-type RSS estimator is guaranteed to be positive as it is a sum of squares and the log-likelihood in Eq 5.5 is always a concave function,

whereas the profile log-likelihood in Eq 5.4 may not be a concave function as $\frac{R}{\pi} \odot C^{-1}$ is not necessarily positive definite for small datasets (e.g. see Figure 5.1). Note that the peak in the top subplot of Figure 5.1 actually goes to positive and negative infinity, because that matrix is not positive definite at that heritability value (i.e., σ^2 is negative, and the logarithm of negative number is undefined).

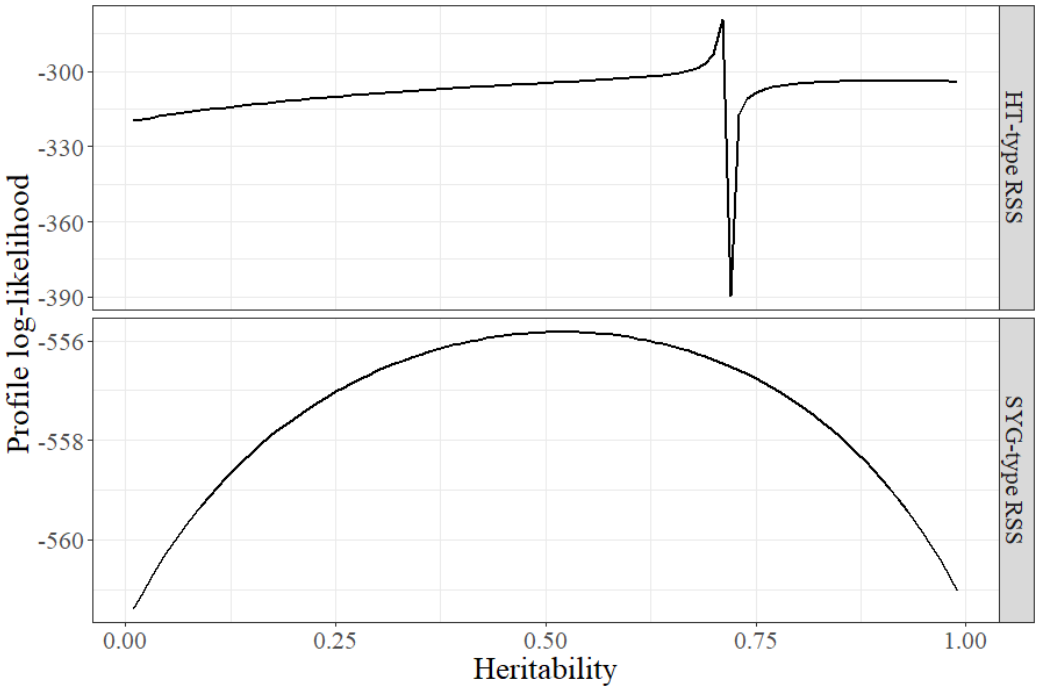

Fig. 5.1 The profile log-likelihood function evaluated over heritability for a particular sample generated from the kākāpō egg length data (see Section 5.2) by outcome-dependent sampling. Note that the peak in the top subplot actually goes to positive and negative infinity, because that matrix is not positive definite at that heritability value.

Let $d_{ijk} = (\widehat{C}^{-\frac{1}{2}})_{ik} - (\widehat{C}^{-\frac{1}{2}})_{jk}$. The MLE of β is given by

$$\hat{\beta} = \begin{bmatrix} \hat{\beta}_0 \\ \hat{\beta}_1 \end{bmatrix} = \begin{bmatrix} t_1 & t_2 \\ t_2 & t_3 \end{bmatrix}^{-1} \begin{bmatrix} t_4 \\ t_5 \end{bmatrix},$$

where

$$\begin{cases} t_1 = \sum\limits_{R_{ij}=1} \dfrac{1}{\pi_{ij}} \left(\sum\limits_{R_k=1} d_{ijk} X_{k1} \right)^2, \\ t_2 = \sum\limits_{R_{ij}=1} \dfrac{1}{\pi_{ij}} \left(\sum\limits_{R_k=1} d_{ijk} X_{k1} \right) \left(\sum\limits_{R_k=1} d_{ijk} X_{k2} \right), \\ t_3 = \sum\limits_{R_{ij}=1} \dfrac{1}{\pi_{ij}} \left(\sum\limits_{R_k=1} d_{ijk} X_{k2} \right)^2, \\ t_4 = \sum\limits_{R_{ij}=1} \dfrac{1}{\pi_{ij}} \left(\sum\limits_{R_k=1} d_{ijk} X_{k1} \right) \left(\sum\limits_{R_k=1} d_{ijk} Y_k \right), \\ t_5 = \sum\limits_{R_{ij}=1} \dfrac{1}{\pi_{ij}} \left(\sum\limits_{R_k=1} d_{ijk} X_{k2} \right) \left(\sum\limits_{R_k=1} d_{ijk} Y_k \right). \end{cases}$$

Since the estimated log-likelihood in Eq 5.5 is based on the difference between pairs of individuals, it provides almost no information about the population mean β_0. This can be fixed by replacing $\hat{\beta}_0$ with a standard estimator for the population mean which is calculated as the weighted mean of the residuals:

$$\hat{\beta}_0 = \frac{\sum_i^n \frac{1}{\pi_i}(Y_i - X_i \hat{\beta}_1)}{\sum_i^n \frac{1}{\pi_i}}.$$

The new $\hat{\beta}_0$ is then used for estimation of the other parameters. The MLE of σ^2 is given by

$$\hat{\sigma^2} = \frac{1}{N(N-1)} \sum_{R_{ij}=1} \frac{1}{\pi_{ij}} \left[\left(\widehat{C}^{-\frac{1}{2}} (y - X\hat{\beta}) \right)_i - \left(\widehat{C}^{-\frac{1}{2}} (y - X\hat{\beta}) \right)_j \right]^2.$$

And the heritability can be estimated by linear optimization over $[0,1]$,

$$\hat{h^2} = \underset{h^2 \in [0,1]}{\operatorname{argmax}}\ \hat{\ell}_p(h^2),$$

where

$$\hat{\ell}_p(h^2) = -\frac{1}{2} \left(\log |C| + N \log \left(\sum_{i=1, j=1}^{N} \frac{R_{ij}}{\pi_{ij}} \left[\left(C^{-\frac{1}{2}} (Y - X\beta) \right)_i - \left(C^{-\frac{1}{2}} (Y - X\beta) \right)_j \right]^2 \right) \right).$$

Parametric bootstrap confidence interval

This section describes a parametric bootstrap method for computing confidence intervals of model parameters under two-phase sampling designs.

Let $\mathbb{P}$ be the population data with true parameters $\theta(\mathbb{P}) = \{\beta_0, \beta_1, \sigma^2, h^2\}$, and $\hat{\theta} = \{\hat{\beta}_0, \hat{\beta}_1, \hat{\sigma^2}, \hat{h^2}\}$ be the sample weighted MLE of $\theta(\mathbb{P})$. The 90% bootstrap confidence interval of $\theta(\mathbb{P})$ can be obtained by the following procedure.

Step 1. Simulate M new populations based on $\hat{\theta}(\mathbb{P}) = \{\hat{\beta}_0, \hat{\beta}_1, \hat{\sigma}, \hat{h^2}\}$.

Step 2. For each $m \in \{1, 2, \cdots, M\}$, let $\mathbb{P}_m$ be the m-th population. Draw a sample from $\mathbb{P}_m$ under the outcome-dependent design, and let $\theta^*(\mathbb{P}_m) = \{\beta_{0m}^*, \beta_{1m}^*, \sigma^{2*}{}_m, h^{2*}{}_m\}$ be the sample weighted MLE of $\hat{\theta}(\mathbb{P})$;

Step 3. For each $m \in \{1, 2, \cdots, M\}$, let $\delta_m^* = \theta^*(\mathbb{P}_m) - \hat{\theta}(\mathbb{P})$, $Q_{0.05}$ denote the 95th percentile, and $Q_{0.95}$ denote the 5th percentile of δ^*, then the 90% confidence interval is given by $\big(\hat{\theta}(\mathbb{P}_m) - Q_{0.05}, \hat{\theta}(\mathbb{P}_m) - Q_{0.95}\big)$.

If a large number of populations $\mathbb{P}$ are generated based on the true parameters $\theta(\mathbb{P})$, the probability of the true parameter fall into the bootstrap confidence interval approaches 90% as the population size and sample size increase. The performance of this parametric bootstrap method is investigated by simulation. For a thousand nuclear family datasets ($N = 1200$) simulated given the true parameters with samples drawn by the outcome-dependent design described in section 5.2, 91.4% of the bootstrap confidence intervals contain β_0, 91.6% of the bootstrap confidence intervals contain β_1, 90.4% of the bootstrap confidence intervals contain σ^2, and 92.5% of the bootstrap confidence intervals contain h^2.

5.2 Weighted MLE inference under two-phase sampling

This section compares the performance of four methods: 1) linear regression; 2) linear mixed model with sample unweighted MLE; 3) linear mixed model with sample weighted MLE; and 4) linear mixed model with Bayesian inference using the full likelihood, under two different two-phase sampling designs. The comparison is carried out using a real kākāpō dataset and a simulated nuclear family dataset, generating a large number of samples from each of the two populations and fitting models to each sample.

The kākāpō phenotypic data was collected as part of regular monitoring activities by the Kākāpō Recovery Team, and access was provided by the New Zealand Department of Conservation. The phenotype used here is egg length, which is one of the two continuous characteristics that is measured for the most kākāpō (the other one is egg width, see Figure 5.2 for the number of phenotyped individuals for other continuous traits). Among the 104 kākāpō whose egg length was measured, besides the three kākāpō from the right-hand-side

of Figure 5.3 who are siblings or half-siblings, the rest of the kākāpō all belonged to the same family due to inbreeding.

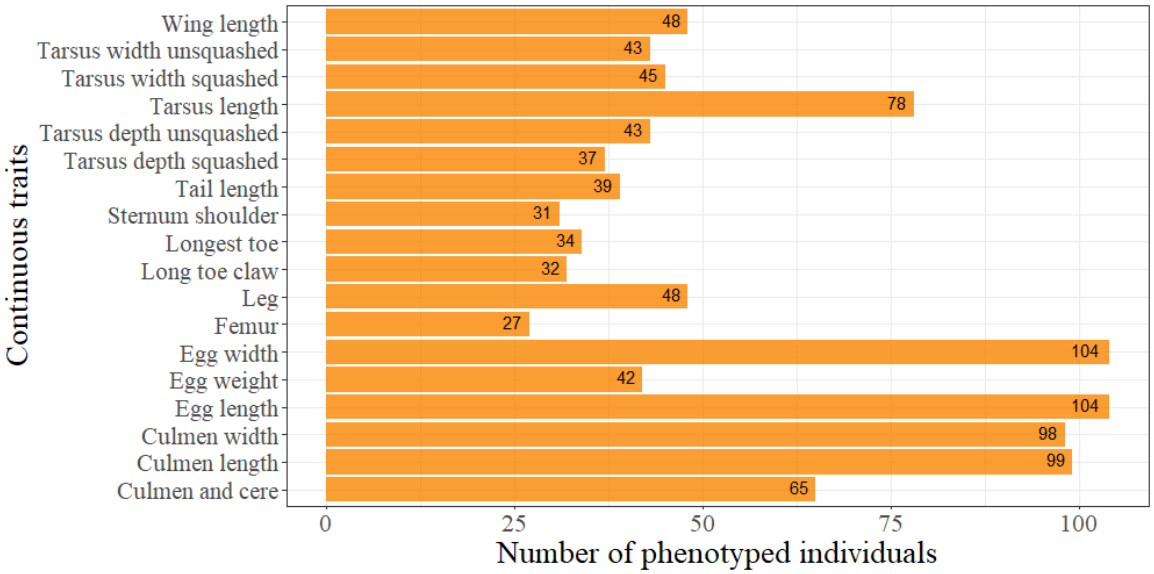

Fig. 5.2 Number of phenotyped kākāpō for each continuous trait (by the 17th March 2021).

Since no genetic markers or regions are known as or carry potential causal genes, the genotype data used here is the genotype of a single locus randomly selected on chromosome S1, where all phenotyped kākāpō are genotyped at this locus and the genotypes have no Mendelian inconsistency. Choosing a different locus would change the true parameters and their estimation. Later in this section, I will show that the same conclusion can be obtained from the proposed approach regardless of the chosen locus by varying the true parameters in the simulated data.

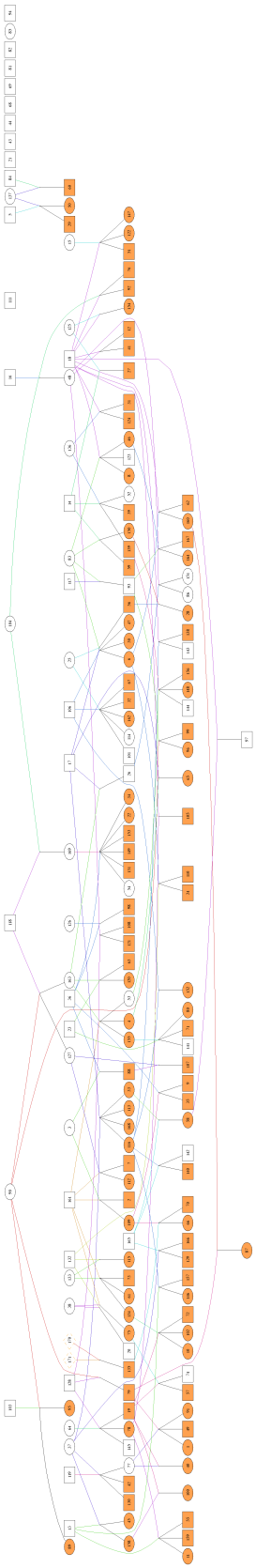

Fig. 5.3 The kākāpō pedigree, where circles represent females, squares represent males, and colored ones are those kākāpō whose egg length are measured.

5.2.1 Outcome-dependent sampling design

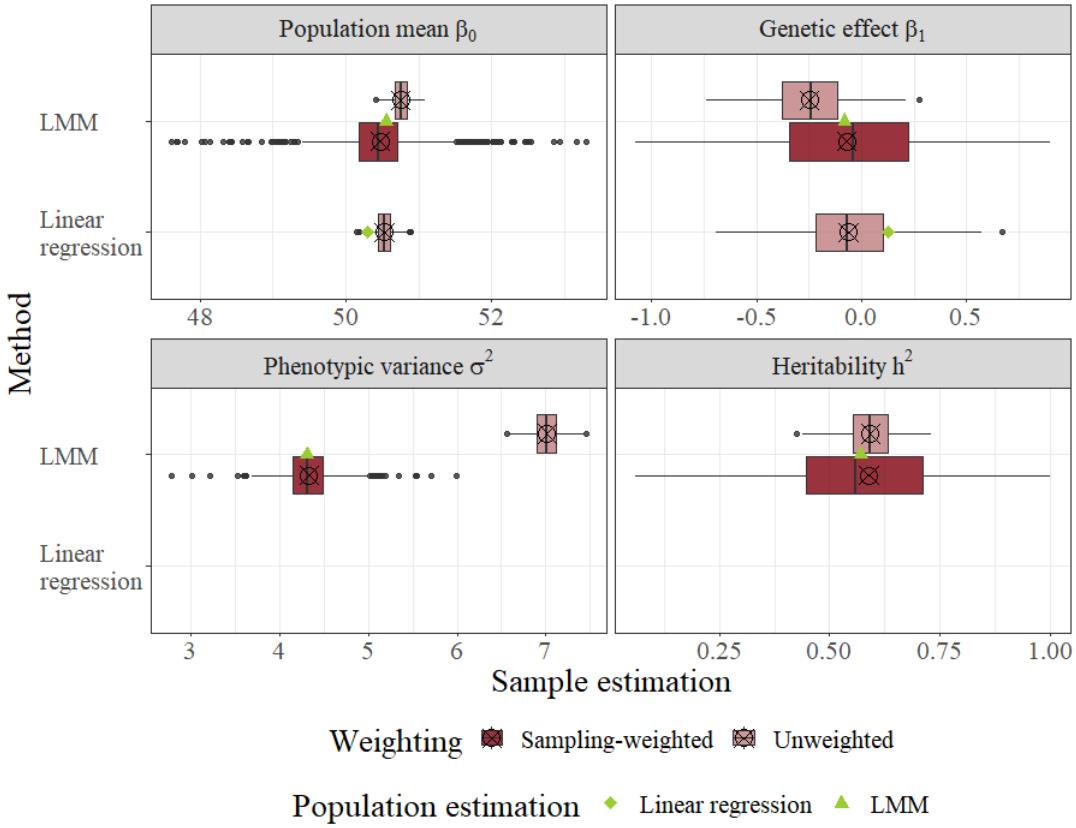

Fig. 5.4 Inference of model parameters under outcome-dependent sampling for kākāpō egg length data.

Consider the outcome-dependent sampling in Design 1, and generate a thousand samples each contains half of the population ($N = 104$, $n = 52$) with individuals from the two 15% tails are always selected. In Figure 5.4, the MLEs from `lme4qtl` are systematically biased under the outcome-dependent design, but the weighted MLE is able to correct the sampling bias by re-weighting the samples. In particular, the MLE for phenotypic variance is clearly overestimated as the sample over-represents the proportion with extreme trait values. In contrast to unweighted methods, the median of weighted MLEs are pretty close to the parameter values estimated using the complete data after discarding some extreme estimates.

However, the variability of the sample estimation using the log-likelihood with HT-type RSS estimator is very large due to the small sample size. Figure 5.5 shows the sample estimations for the kākāpō dataset using the log-likelihood with SYG-type RSS estimator described in section 5.1.3. Although the MLE of the log-likelihood in Eq 5.5 is biased and

the bias in $\hat{\sigma}^2$ tends to increase as the data size increases (Figure 5.6 and Figure 5.7), it seems to provide reasonable estimates for small datasets (Figure 5.5 and Figure 5.6).

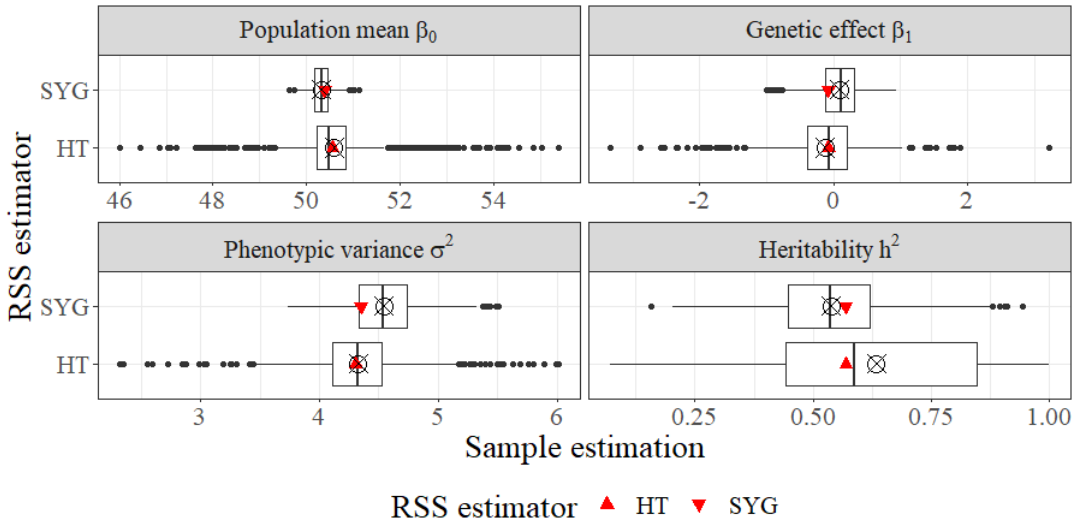

Fig. 5.5 Inference of linear mixed model parameters under outcome-dependent sampling using log-likelihood with HT-type RSS estimator and log-likelihood with SYG-type RSS estimator for the kākāpō egg length data.

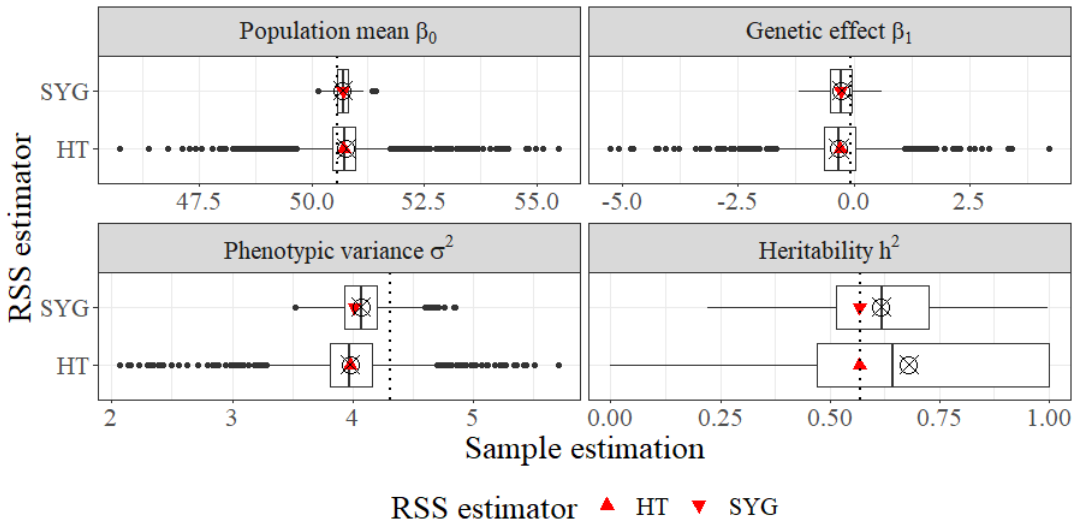

Fig. 5.6 Inference of linear mixed model parameters under outcome-dependent sampling using log-likelihood with HT-type RSS estimator and log-likelihood with SYG-type RSS estimator for the simulated nuclear family data ($N = 120$). The vertical dotted lines represent the true parameters of the simulated data.

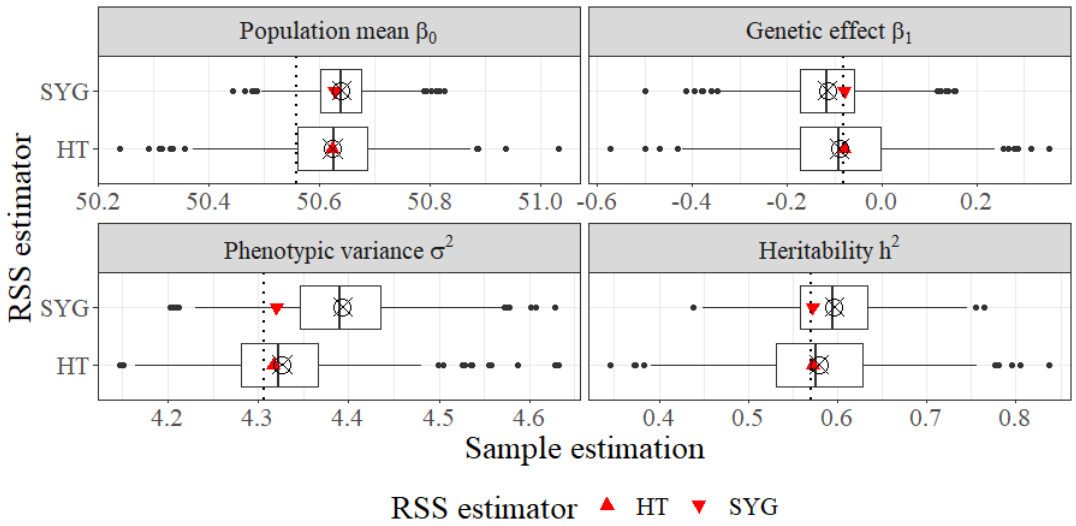

Fig. 5.7 Inference of linear mixed model parameters under outcome-dependent sampling using log-likelihood with HT-type RSS estimator and log-likelihood with SYG-type RSS estimator for the simulated nuclear family data ($N = 1200$). The vertical dotted lines represent the true parameters of the simulated data.

To extend the results of the HT-type RSS estimator to larger populations, I consider a simulated dataset with $N = 1200$ individuals from 300 independent nuclear families, each consisting of two unrelated parents and two offspring. As opposed to the kākāpō case, where the majority of the individuals are related to each other, individuals in the simulated data are only related to their family members, hence the covariance matrix Ξ is simply a block diagonal matrix.

Figure 5.8 shows the parameter inference of a thousand samples generated by the same outcome-dependent sampling. The results are mostly consistent with the small kākāpō dataset, but less variable and both mean and median of the weighted MLEs agree with the population estimates. As data size increases, the bias in heritability estimator becomes more obvious than in 5.4, which implies that the proportion of genetic variance in total phenotypic variance will also be overestimated under outcome-dependent sampling without weights adjustment.

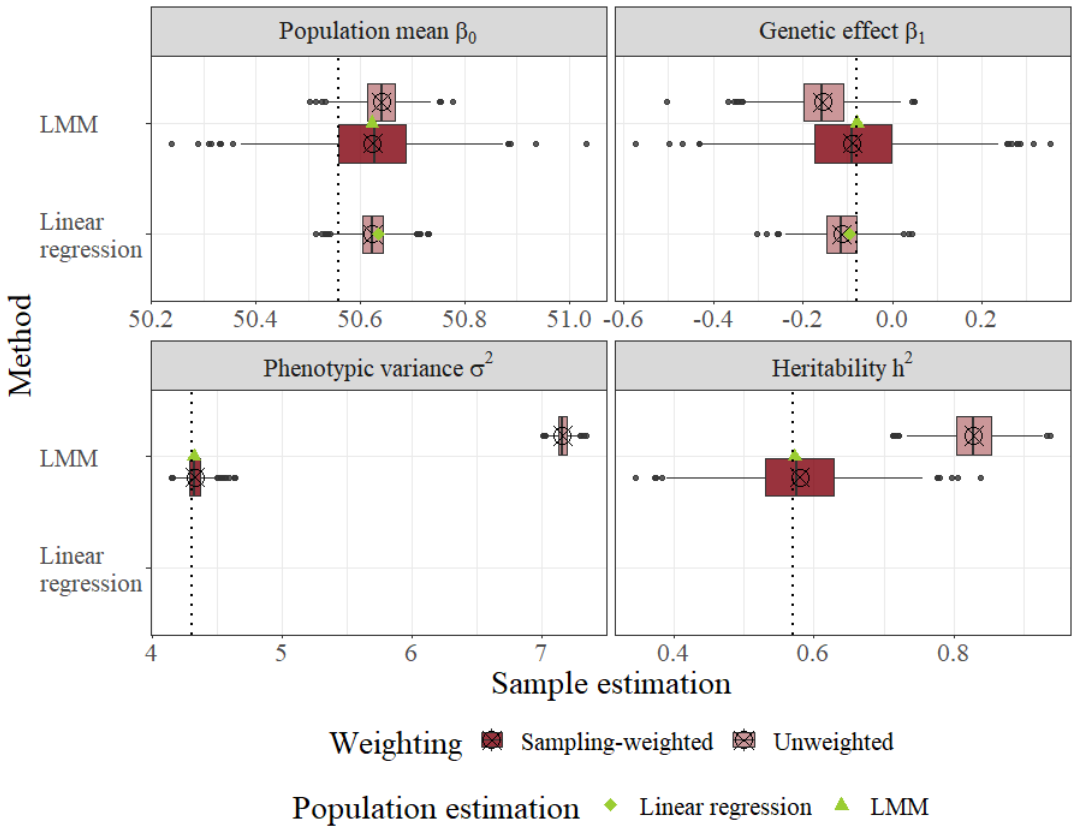

Fig. 5.8 Inference of model parameters under outcome-dependent sampling for simulated nuclear family data ($N = 1200$). The vertical dotted lines represent the true parameters of the simulated data.

To see whether varying model parameters affects the conclusions on the proposed weighted approach, I simulated eight nuclear family datasets with different parameter values ($\beta_1 = -0.8$ or $\beta_1 = 8$, $\sigma^2 = 1$ or $\sigma^2 = 5$, $h^2 = 0$ or $h^2 = 0.8$) and the results are shown in Figure 5.9. Each pair of datasets investigates the possible effect of varying one or more model parameters (e.g. Dataset 1 and Dataset 2 demonstrate the effect of varying β_1 only, Dataset 1 and Dataset 8 demonstrate that the effect of varying β_1, σ^2, h^2). Figure 5.9 shows that varying model parameters has no effect on the conclusions, i.e., the sampling bias can be corrected by re-weighting the samples, even when $h^2 = 0$. On the other hand, when the parameters are estimated without weights adjustments, $\boldsymbol{\beta} = (\beta_0, \beta_1)$ and σ^2 are always biased, the bias in h^2 seems to be the worst when β_1 is small and σ^2 is large, and almost no bias in h^2 when β_1 is large and σ^2 is small.

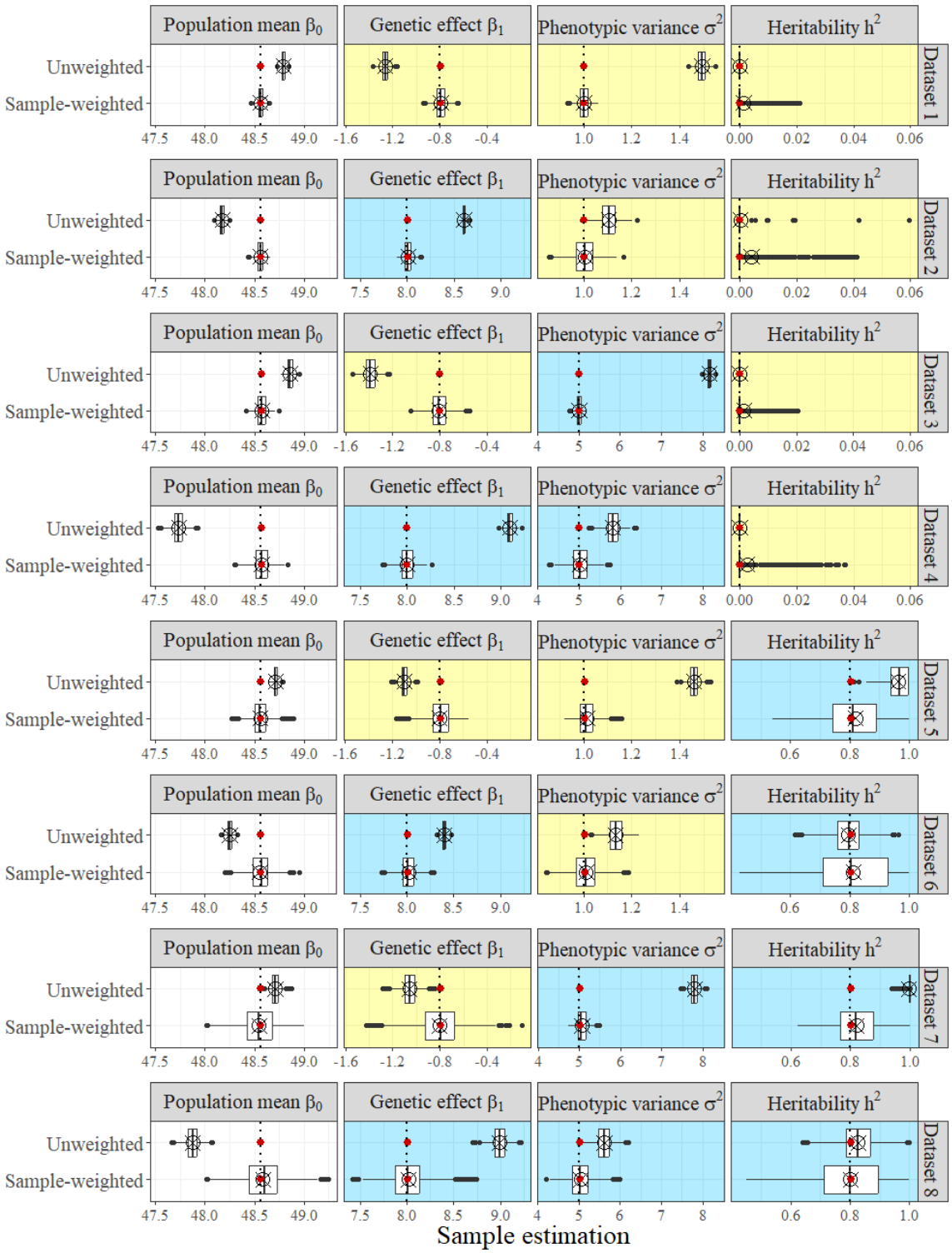

Fig. 5.9 The effect of varying model parameters (β_1,σ^2,h^2). The eight simulated nuclear family datasets ($N = 1200$) are generated with $\beta_1 = -0.8$ or $\beta_1 = 8$, $\sigma^2 = 1$ or $\sigma^2 = 5$, $h^2 = 0$ or $h^2 = 0.8$. For each column, panels with the same colour have the same true value and the same x-axis range. The samples are selected under outcome-dependent sampling. The vertical dotted lines represent the true parameters of the simulated data and the red dots represent the population estimates of the simulated data.

As mentioned in section 5.1.2, the Bayesian approach with the full likelihood is workable in the simple scenario of independent nuclear families in the simulated dataset. The boxplot in Figure 5.10 shows the comparison between weighted likelihood and the Bayesian inference using the full likelihood, where the latter is carried out by the MCMC algorithm implemented in NIMBLE [40]. When half of the population is sampled, there is a large variance in the weighted MLE, but both the mean and median of the sample weighted MLE agree with the population estimates. In contrast to the sample weighted MLE, the posterior mean of the MCMC samples is less variable but biased, particularly for the genetic effect and heritability. As 25% more individuals are sampled from the population, the variability in the sample weighted MLE is roughly halved, and the posterior mean of the MCMC samples approaches the population estimate.

The histogram in Figure 5.10 shows the distribution of the computation time of sample weighted MLE and the Bayesian approach with the full likelihood. In this simple example, the average computation time for the Bayesian approach using full likelihood is about 2.7 times as long as weighted likelihood. The Bayesian approach using full likelihood may be even slower for varying family sizes, more complicated sampling designs or pedigree structures, but it would be too complicated to implement.

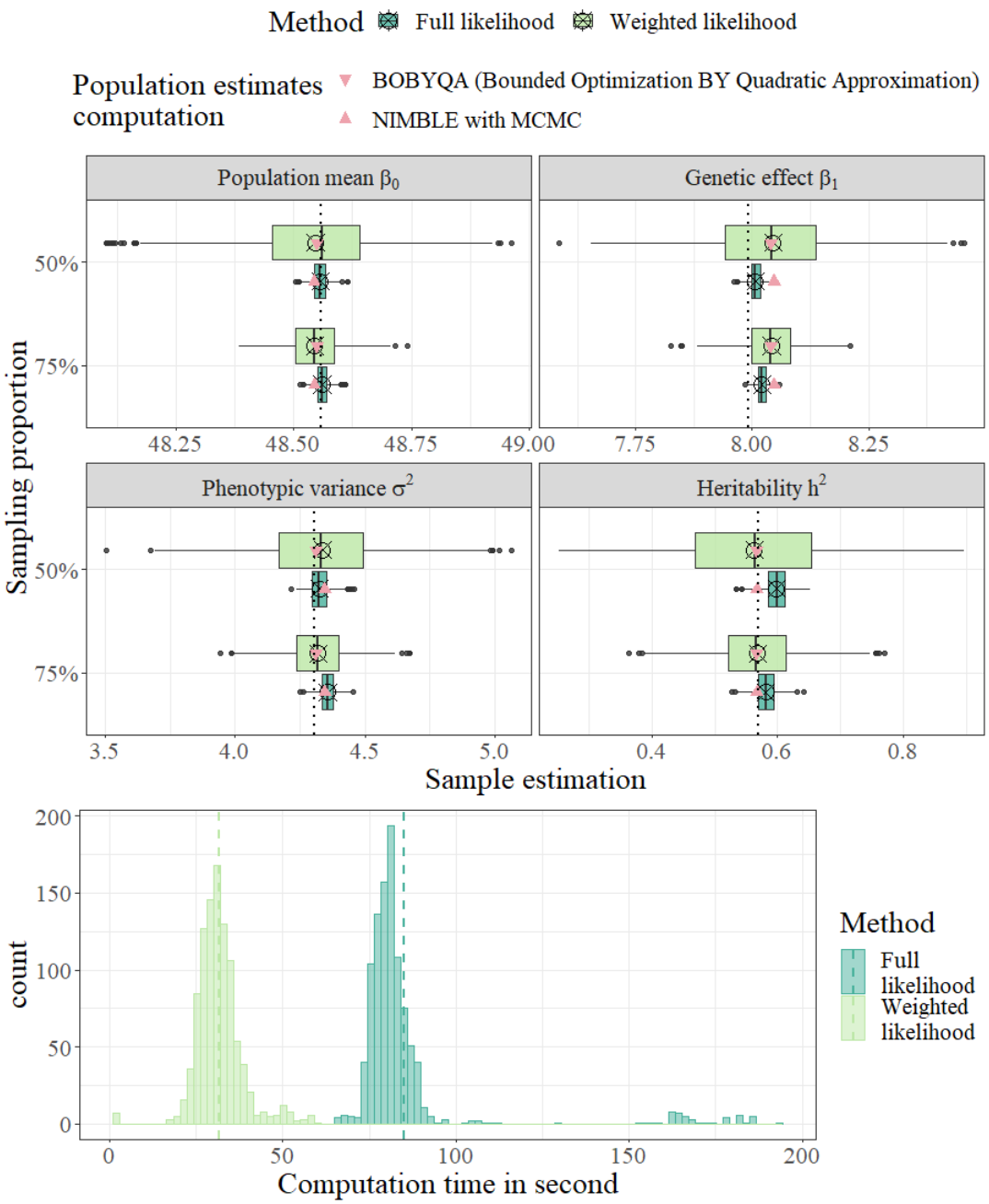

Fig. 5.10 The weighted likelihood versus the Bayesian inference using full likelihood: (1) the box plot shows the inference of model parameters under outcome-dependent sampling for simulated nuclear family data ($N = 1200$), where the vertical dotted lines represent the true parameters of the simulated data; (2) the histogram shows the computation time of the two methods when half the population are sampled (increasing the sampling proportion does not make a visible difference in the computation time), and the vertical dotted line are the mean computation time.

I also compared the proposed weighted likelihood to the pairwise pseudolikelihood, and the code for the pairwise pseudolikelihood can be found on GitHub (https://gist.github.com/tslumley/39b154317d6e0726ac4d138164d38a24). In contrast to the naive likelihood which is not adjusted by sampling weights, both the weighted likelihood and the pseudolikelihood are able to correct the sampling bias (see Figure 5.11). However, using the sample covariance matrix does not lead to any improvement in the efficiency compared to using the pair covariance matrix, particularly in the estimation of the fixed effect. Figure 5.12 investigates the reason for the difference between the weighted likelihood and the pairwise pseudolikelihood. The relatively low correlation between estimators of the same parameter using different methods shows that the two methods are not extracting exactly the same information from the data.

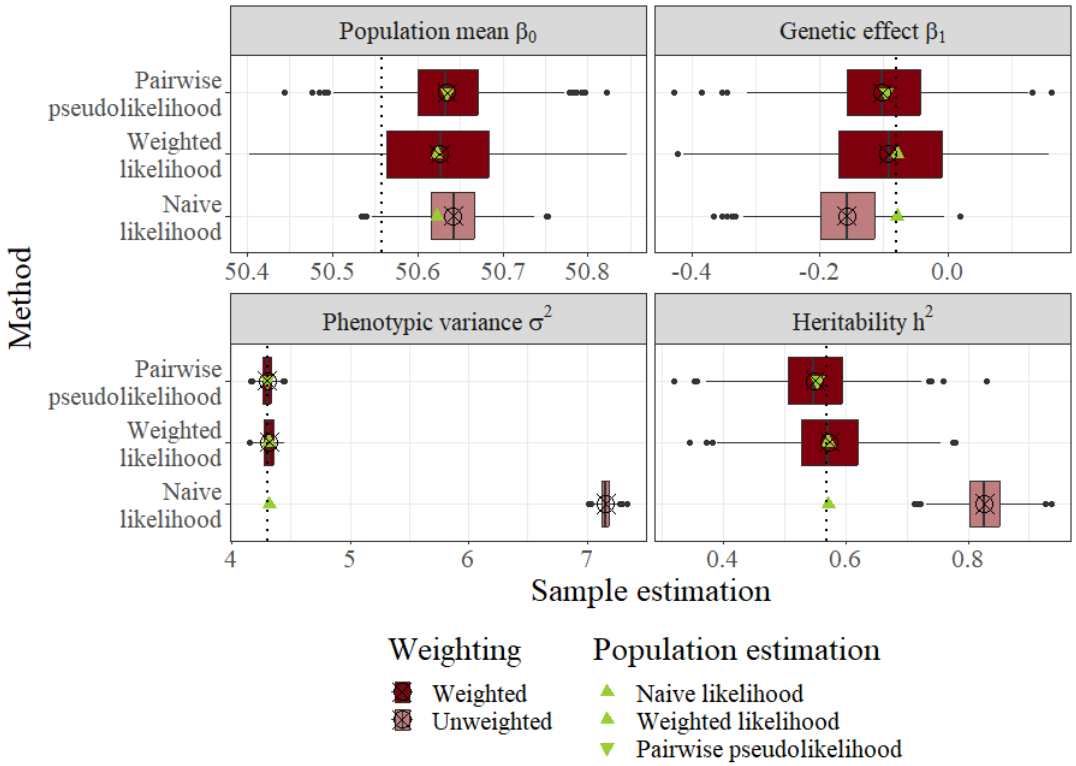

Fig. 5.11 The weighted likelihood versus the pairwise pseudolikelihood. The box plot shows the inference of model parameters under outcome-dependent sampling for simulated nuclear family data ($N = 1200$), where the vertical dotted lines represent the true parameters of the simulated data.

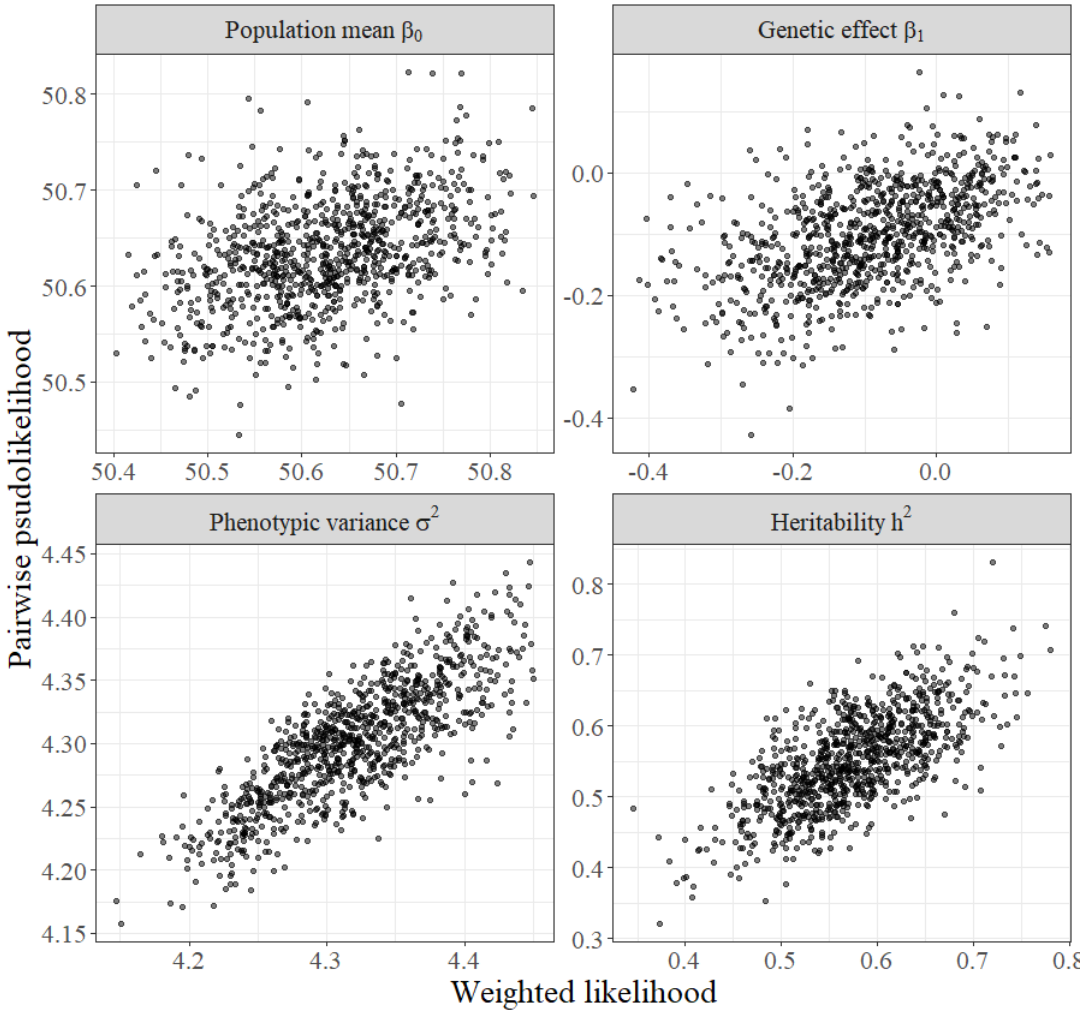

Fig. 5.12 Correlation between estimations of the same parameter using different method.

Here I use the parametric bootstrap method described in section 5.1.3 to compute the 90% bootstrap confidence interval of model parameters under outcome-dependent sampling for the kākāpō data and the simulated nuclear family data.

To obtain expectations of $Q_{0.05}$ and $Q_{0.95}$, we can repeat the above procedure a thousand times using the kākāpō data and the simulated nuclear family data. For a sample weighted MLE $\hat{\theta}$, the 90% bootstrap confidence interval of θ is shown in Table 5.1 and Table 5.2. The weighted likelihood with HT-type RSS estimator gives very wide confidence intervals for all parameters because it is difficult to maximize over multiple parameters for the small kākāpō data. On the other hand, the weighted likelihood with SYG-type RSS estimator gives much narrower confidence intervals for β and σ^2, but the 90% bootstrap confidence interval for h^2 remains uninformative as it almost contains the entire range of h^2. For the simulated data,

which is approximately ten times larger than the kākāpō data, the sample weighted MLE is more precise as the 90% bootstrap confidence intervals are relatively narrow.

Parameter	90% Bootstrap confidence interval	
	The HT-type RSS estimator	The SYG-type RSS estimator
β_0	$\left[\hat{\beta}_0 - 15.86, \hat{\beta}_0 + 15.23\right]$	$\left[\hat{\beta}_0 - 0.76, \hat{\beta}_0 + 0.75\right]$
β_1	$\left[\hat{\beta}_1 - 4.98, \hat{\beta}_1 + 4.89\right]$	$\left[\hat{\beta}_1 - 0.91, \hat{\beta}_1 + 0.92\right]$
σ^2	$\left[\hat{\sigma^2} - 1.11, \hat{\sigma^2} + 4.31\right]$	$\left[\hat{\sigma^2} - 1.42, \hat{\sigma^2} + 1.15\right]$
h^2	$\left[\hat{h^2} - 0.43, \hat{h^2} + 0.47\right]$	$\left[\hat{h^2} - 0.50, \hat{h^2} + 0.46\right]$

Table 5.1 The 90% bootstrap confidence interval of model parameters under outcome-dependent sampling for the kākāpō data ($N = 104$), where $\hat{\theta}$ is the sample weighted MLE.

Parameter	90% Bootstrap confidence interval
β_0	$\left[\hat{\beta}_0 - 0.24, \hat{\beta}_0 + 0.24\right]$
β_1	$\left[\hat{\beta}_1 - 0.29, \hat{\beta}_1 + 0.29\right]$
σ^2	$\left[\hat{\sigma^2} - 0.36, \hat{\sigma^2} + 0.33\right]$
h^2	$\left[\hat{h^2} - 0.16, \hat{h^2} + 0.13\right]$

Table 5.2 The 90% bootstrap confidence interval of model parameters under outcome-dependent sampling for the simulated nuclear family data ($N = 1200$), where $\hat{\theta}$ is the sample weighted MLE.

Bootstrap can also be used to obtain the standard errors of the weighted estimation. The variance of the weighted estimation is the sum of phase I variance and phase II variance. While the standard error at phase I can be estimated by the square roots of the diagonal elements of the negative Hessian, the standard error at phase II can be estimated by bootstrapping. For example, we can generate K bootstrap samples from the same dataset by repeating the outcome-dependent sampling K times, and obtain the weighted MLE $\hat{\theta}_k$ for each bootstrap sample. Then, the bootstrap standard error for the weighted MLE $\hat{\theta}$ (of the

original sample) is given by

$$SE(\hat{\theta}) = \sqrt{\frac{1}{K-1}\sum_{k=1}^{K}(\hat{\theta}_k - \bar{\theta})^2},$$

where $\bar{\theta} = \frac{1}{K}\sum_{k=1}^{K}\hat{\theta}_k$ denotes the mean of the weighted MLE across the K bootstrap samples.

5.2.2 Outcome-pedigree-dependent sampling

Design 2. *Let N be the population size and n be the sample size, consider the following utcome-pedigree-dependent sampling.*

Step 1. Always sample the n_1 individuals from the two a% tails who has at least two relatives that are also in two a% tails of the phenotype distribution;

Step 2. Randomly sample $n_2 = n - n_1$ individuals from the rest $N - n_1$ individuals, where n is the sample size.

Since we are interested in estimating the proportion of genetic variance in total phenotypic variance, it may be helpful to sample individuals based on the pedigree as well as the outcome. Consider the outcome-pedigree-dependent sampling in Design 2, and generate a thousand samples each contains half of the population ($N = 1200$, $n = 600$) with individuals whose phenotype and at least two relatives' phenotypes are from the two 25% tails are always selected.

From Figure 5.13, the same conclusion can be made under outcome-pedigree-dependent sampling as in outcome-dependent sampling; that is, the sampling bias can be corrected by re-weighting the samples. However, the weighted MLE seems to be more variable than under sampling designs based on outcome only, despite including more related individuals, whereas MLE tends to do better under outcome-pedigree-dependent sampling. For both two-phase designs, the weighted MLE is consistent as shown in Figure 5.14, and formal proof is provided in Section 5.4.

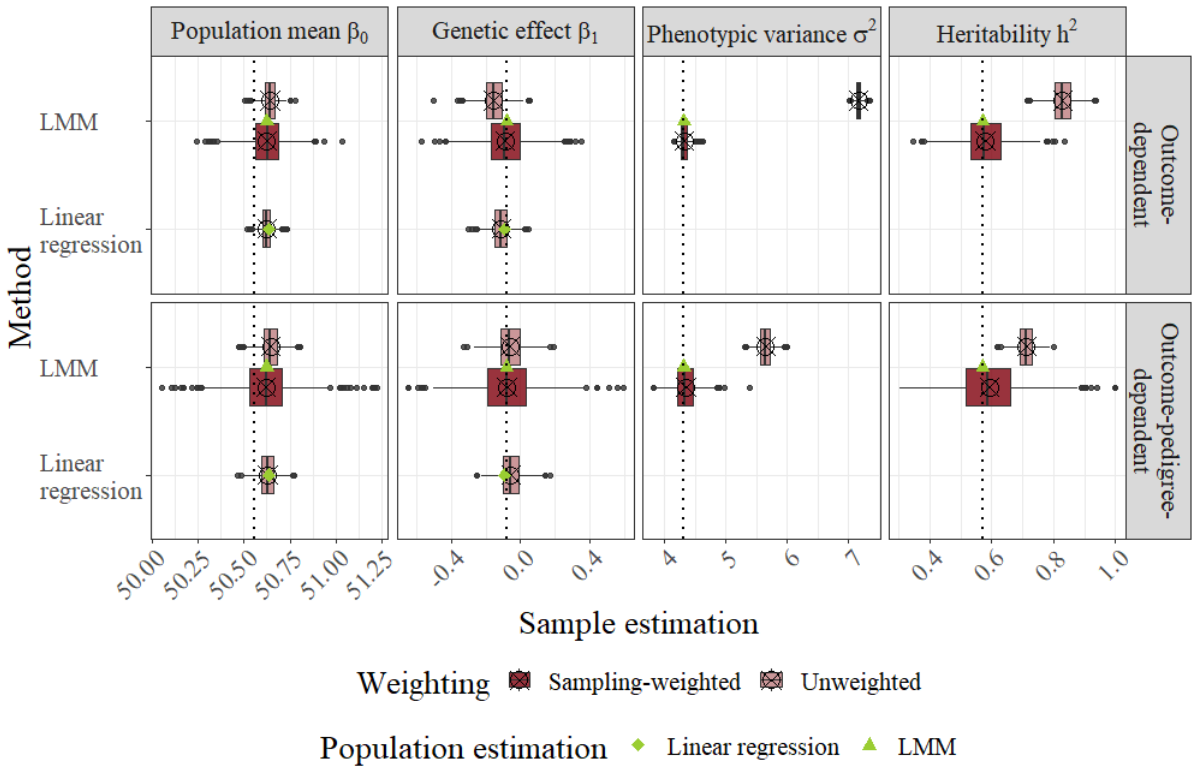

Fig. 5.13 Inference of model parameters under outcome-dependent sampling and outcome-pedigree-dependent sampling for simulated nuclear family data ($N = 1200$). The vertical dotted lines represent the true parameters of the simulated data. The top row is the same as Figure 5.8, and it is included here for comparsion.

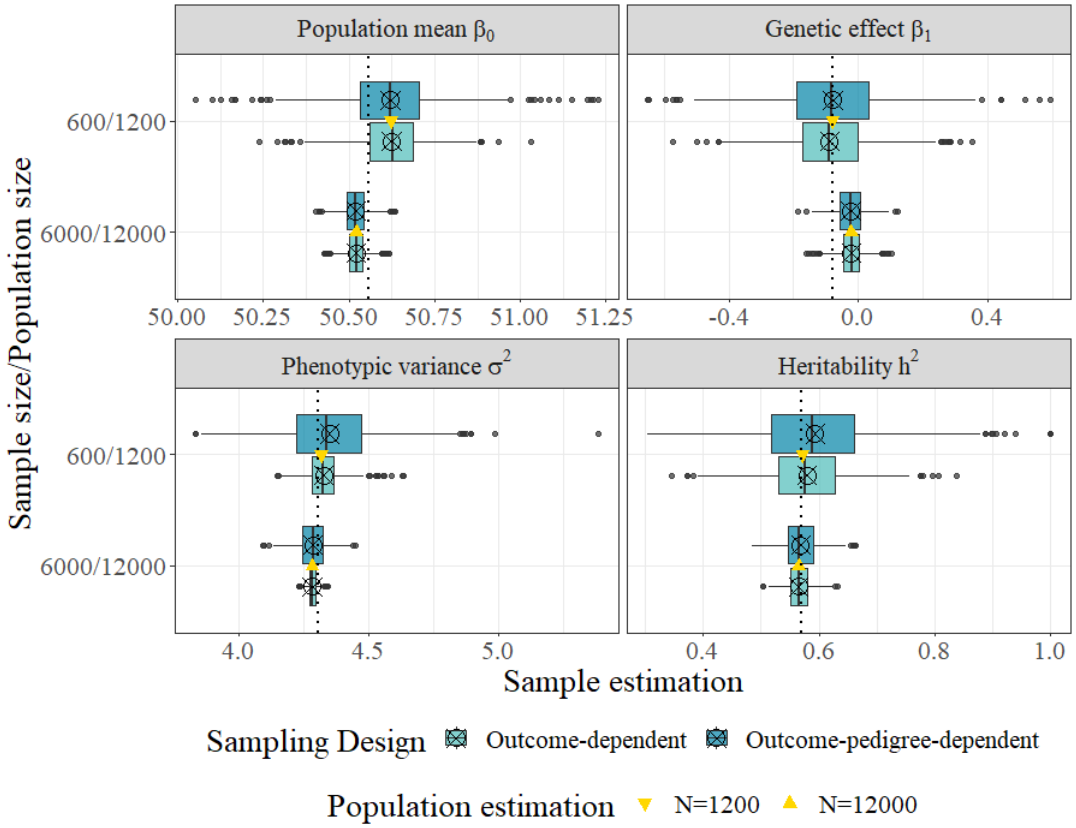

Fig. 5.14 Inference of linear mixed model parameters under two-phase sampling for simulated nuclear family data with increase data size. The vertical dotted lines represent the true parameters of the simulated data.

5.3 Single-locus mixed models versus multi-locus mixed models

For complex traits controlled by several loci with moderate to large effect and numerous loci with small effect, single-locus mixed models may lead to a loss in the power of detecting association. In contrast to single-locus mixed models, multi-locus mixed models account for the possible confounding effects of the background loci across the genome by including multiple loci with large effect in the model. For multiple closely linked loci, multi-locus models can be used to estimate the effect of one locus while conditioning on the others [37].

Although multi-locus mixed models enable a power gain, fitting such a model is challenging, particularly in GWAS, because there are hundreds of thousands of genetic variants but the sample size is typically less than tens of thousands. Furthermore, the presence of linkage disequilibrium posed by population structure makes it more complicated to identify

the causal variants. The multi-locus mixed model can be extended for Bayesian analysis, which better accommodates the genetic architecture of complex traits via a flexible prior on SNP effect sizes [89, 130, 139, 172]. However, it remains a problem to fit Bayesian mixed models under two-phase designs.

For the kākāpō egg length data, the proposed weighted method takes approximately 5.56 hours (using an laptop with Intel(R) Core(TM) i7-8550U CPU and 16GB RAM) to fit a single-locus linear mixed model under a two phase design to 100K loci. If the trait is assumed to be affected by many variants with small effect, then the model can be fitted with fixed heritability as its estimation is unlikely to vary much for a different locus [25, 74, 141, 167]. Consequently, the computation time is reduced to approximately 1.94 hours. Due to the difficulties with multi-locus mixed models, and the fact that single-locus mixed models have successfully identified thousands of variants in GWAS of humans, animals and plants [7, 55, 75, 95, 136, 163], it is still worth fitting a single-locus mixed model in the kākāpō study. Then, a multi-locus model can be used to fit the top hits in the follow-up study.

Becasue of the small size of the kākāpō data, it is difficult to fit a linear mixed model with many loci. Another consequence is that, there may be no variation or little variation between the sets of genotypes of a kākāpō sample at some loci. While rank deficiency is not necessarily a problem in model fitting, it is inconvenient for comparison between sample estimation and population estimation of the model parameters. Therefore, I will fit a two-locus model to the kākāpō data and a ten-locus model to the simulated nuclear family data (N=1200) to show that it is also possible to fit a multi-locus mixed model and obtain sample weighted estimation of the model parameters. Figure 5.15 and Figure 5.16 demonstrate that the proposed weighted approach can be applied for multi-locus linear mixed model as long as there is enough data.

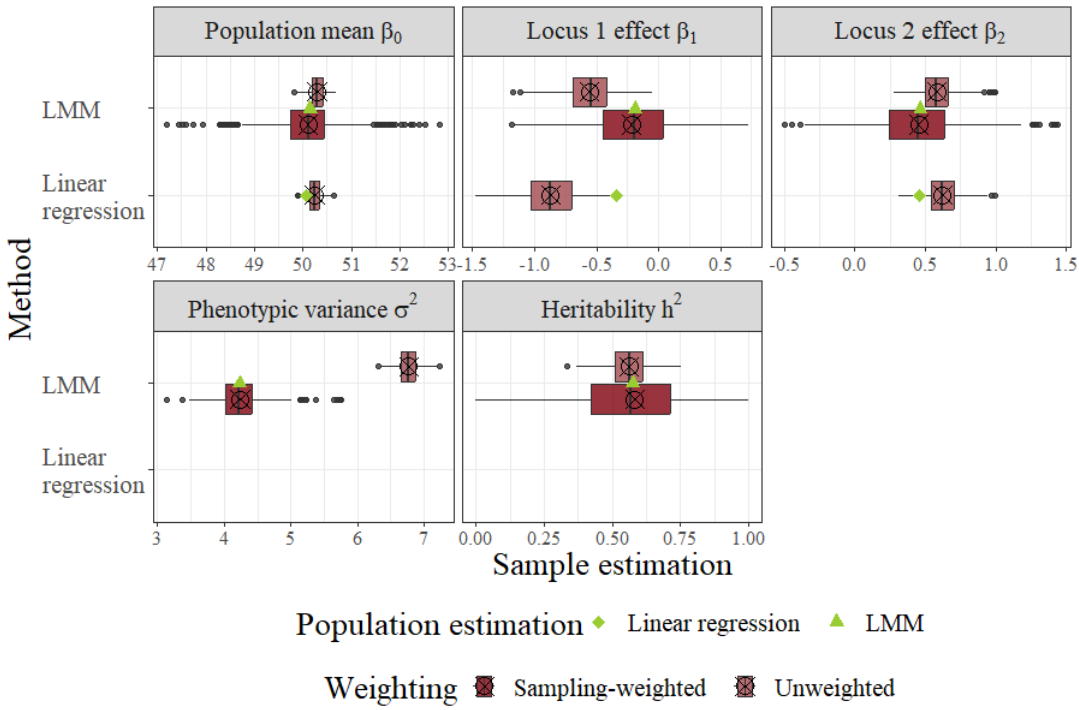

Fig. 5.15 Inference of multi-locus model parameters under outcome-dependent sampling for the kākāpō egg length data.

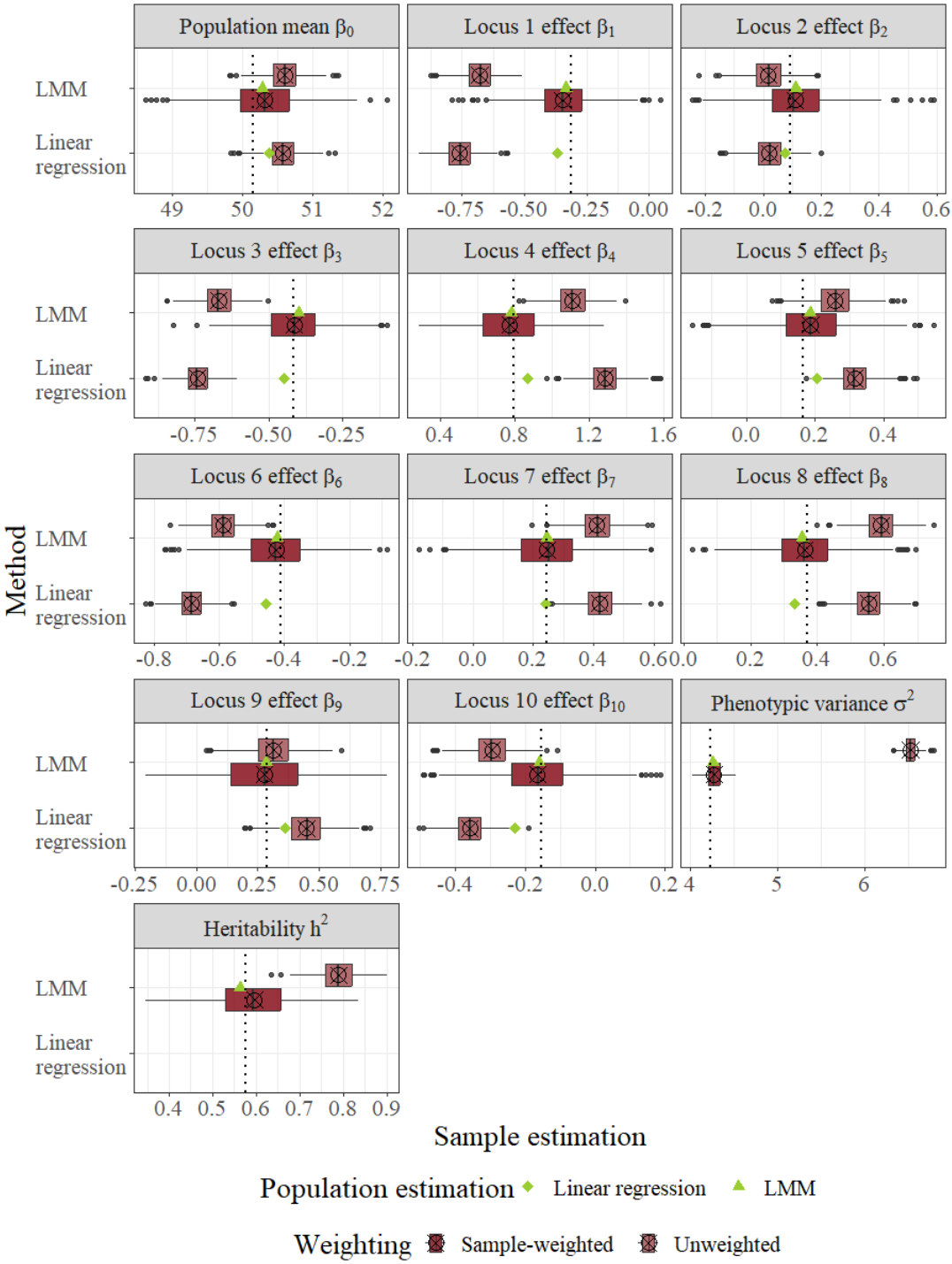

Fig. 5.16 Inference of multi-locus model parameters under outcome-dependent sampling for the simulated nuclear family data ($N = 1200$). The vertical dotted lines represent the true parameters of the simulated data.

5.4 Consistency of the sample weighted likelihood estimator

Theorem 5.4.1. *(Law of Large Numbers) Let $\ell_N(\theta)$ be the population log-likelihood estimator*

$$\ell_N(\theta) = -\frac{1}{2}\left(\log|\Xi| + \sum_{i=1,j=1}^{N} V_{ij}(\theta)\right), \tag{5.6}$$

and $\hat{\ell}_n(\theta)$ be the sample weighted log-likelihood estimator

$$\hat{\ell}_n(\theta) = -\frac{1}{2}\left(\log|\Xi| + \sum_{i=1,j=1}^{N} \frac{R_{ij}}{\pi_{ij}} V_{ij}(\theta)\right), \tag{5.7}$$

where

$$V_{ij}(\theta) = (X\beta - y)_i \left(\Xi^{-1}\right)_{ij} (X\beta - y)_j. \tag{5.8}$$

Given a sequence of finite populations $\{\mathcal{F}_N\}$ and an associated sequence of sample designs, $\hat{\ell}_n(\theta)$ is design consistent for $\ell_N(\theta)$ if for every $\varepsilon > 0$,

$$\lim_{\substack{N\to\infty,\\ \frac{n}{N}\to c}} P\left\{\frac{1}{N}\left|\hat{\ell}_n(\theta) - \ell_N(\theta)\right| > \varepsilon | \mathcal{F}_N\right\} = 0, \tag{5.9}$$

where c is a constant and $c \in (0,1]$.

Proof. Let $\mathbb{V}_\pi$ and Cov_π denote the variance and covariance with respect to the sampling design. Recall that π_{ij} and π_{kl} are the probabilities of both individual i and j, both individual k and l are sampled respectively. Similarly, π_{ijkl} is the probabilities of individuals i, j, k, l are sampled.

$$\begin{aligned}
\mathbb{V}_\pi\left[\hat{\ell}_n(\theta)-\ell_N(\theta)\right] &= \mathbb{V}_\pi\left[-\frac{1}{2N^2}\left(\log|\Xi|+\sum_{i=1,j=1}^{N}\frac{R_{ij}}{\pi_{ij}}V_{ij}(\theta)\right)+\frac{1}{2N^2}\left(\log|\Xi|+\sum_{i=1,j=1}^{N}V_{ij}(\theta)\right)\right] \\
&= \frac{1}{4N^4}\mathbb{V}_\pi\left[\sum_{i=1,j=1}^{N}\frac{R_{ij}}{\pi_{ij}}V_{ij}(\theta)\right] \\
&= \frac{1}{4N^4}\sum_{i=1,j=1}^{N}\sum_{k=1,l=1}^{N}\mathrm{Cov}_\pi\left(\frac{R_{ij}}{\pi_{ij}}V_{ij}(\theta),\frac{R_{kl}}{\pi_{kl}}V_{kl}(\theta)\right) \\
&= \frac{1}{4N^4}\sum_{i=1,j=1}^{N}\sum_{k=1,l=1}^{N}\frac{1}{\pi_{ij}}V_{ij}(\theta)\frac{1}{\pi_{kl}}V_{kl}(\theta)\mathrm{Cov}_\pi(R_{ij},R_{kl}) \\
&= \frac{1}{4N^4}\sum_{i=1,j=1}^{N}\sum_{k=1,l=1}^{N}\frac{1}{\pi_{ij}}V_{ij}(\theta)\frac{1}{\pi_{kl}}V_{kl}(\theta)(\pi_{ijkl}-\pi_{ij}\pi_{kl}) \\
&= \frac{1}{4N^4}\sum_{i=1,j=1}^{N}\sum_{k=1,l=1}^{N}\frac{\pi_{ijkl}}{\pi_{ij}\pi_{kl}}V_{ij}(\theta)V_{kl}(\theta)-V_{ij}(\theta)V_{kl}(\theta) \\
&= \frac{1}{4N^4}\sum_{i=1,j=1}^{N}\sum_{k=1,l=1}^{N}\left(\frac{\pi_{ijkl}}{\pi_{ij}\pi_{kl}}-1\right)V_{ij}(\theta)V_{kl}(\theta). \qquad (5.10)
\end{aligned}$$

For any sampling design that satisfies

$$\sum_{i=1,j=1}^{N}\sum_{k=1,l=1}^{N}\left(\frac{\pi_{ijkl}}{\pi_{ij}\pi_{kl}}-1\right)V_{ij}(\theta)V_{kl}(\theta)\to 0$$

as $N\to\infty$ and $\frac{n}{N}\to c$ with $c\in(0,1]$, $\hat{\ell}_n(\theta)$ is design consistent for $\ell_N(\theta)$. □

For example, in the outcome-dependent sampling described in section 5.1.2, $\pi_{ij}=\pi_i\pi_j$, where π_i and π_j equal to 1 if individual i and j are from the two extreme tails or $\frac{n_2}{N-n_1}$ otherwise. Similarly, $\pi_{ijkl}=\pi_i\pi_j\pi_k\pi_l$. Hence, Theorem 5.4.1, which is also the regularity condition A.3 of Theorem 5.4.2, is true, and the following Theorem 5.4.2 holds.

Theorem 5.4.2. *(Uniform Law of Large Numbers) Under the following regularity conditions,*

A.1 *The parameter $\Theta\subset\mathbb{R}^k$ is compact;*

A.2 *$\hat{\ell}_n(\theta)$ is differentiable at each $\theta\in\Theta$ with bounded derivative;*

A.3 *$\frac{1}{N}\left(\hat{\ell}_n(\theta)-\ell_N(\theta)\right)\xrightarrow{P}0$ as $N\to\infty$ and $\frac{n}{N}\to c$ with $c\in(0,1]$ with respect to design probability π (Theorem 5.4.1).*

Then, $\sup_{\theta\in\Theta}\frac{1}{N}\left|\hat{\ell}_n(\theta)-\mathbb{E}_{\theta_0}\left[\ell_N(\theta)\right]\right|\xrightarrow{P}0$ *as $N\to\infty$ and $\frac{n}{N}\to c$ with $c\in(0,1]$.*

Proof. See the proof of Corollary 2.2 in [102]. □

Definition 5.4.1. *The pairwise sampling probability π_{ij} is defined as*

$$\pi_{ij} = \mathbb{E}_\pi[R_{ij}|Y,\Phi],$$

where $\mathbb{E}_\pi$ is the expectation with respect to the sampling design, R_{ij} is the pairwise sampling indicator function, response Y follows a multivariate normal distribution $\mathcal{N}(X\beta, \sigma^2(h^2\Phi + (1-h^2)I))$ and Φ is the kinship matrix. Both Y and Φ are phase I information which is available for all individuals in phase I sample.

Lemma 5.4.3. *The true parameter θ_0 of θ is a solution of*

$$\mathbb{E}_{Y\pi}[U_n(\theta)] = 0,$$

where $\mathbb{E}_{Y\pi}$ is the expectation with respect to the model and the sampling design, U_n is the sample weighted likelihood score.

Proof.

$$\begin{aligned}
\mathbb{E}_{Y\pi}[U_n(\theta_0)] &= \mathbb{E}_{Y\pi}\left[\frac{\partial}{\partial\theta}\hat{\ell}_n(\theta)\Big|_{\theta_0}\right] \\
&= -\frac{1}{2}\mathbb{E}_{Y\pi}\left[\frac{\partial}{\partial\theta}\log|\Xi(\theta)|\Big|_{\theta_0} + \frac{\partial}{\partial\theta}\sum_{i=1,j=1}^{N}\frac{R_{ij}}{\pi_{ij}}V_{ij}(\theta)\Big|_{\theta_0}\right] \\
&= -\frac{1}{2}\mathbb{E}_{Y}\left[\frac{\partial}{\partial\theta}\log|\Xi(\theta)|\Big|_{\theta_0} + \frac{\partial}{\partial\theta}\sum_{i=1,j=1}^{N}\frac{\mathbb{E}_{\pi|Y}[R_{ij}]}{\pi_{ij}}V_{ij}(\theta)\Big|_{\theta_0}\right] \\
&= -\frac{1}{2}\mathbb{E}_{Y}\left[\frac{\partial}{\partial\theta}\log|\Xi(\theta)|\Big|_{\theta_0} + \frac{\partial}{\partial\theta}\sum_{i=1,j=1}^{N}V_{ij}(\theta)\Big|_{\theta_0}\right] \\
&= \mathbb{E}_{Y}\left[\frac{\partial}{\partial\theta}\ell_N(\theta)\Big|_{\theta_0}\right] \\
&= 0
\end{aligned}$$

Note that $\mathbb{E}_{\pi|Y}$ is the expectation with respect to the sampling design given the model. □

Theorem 5.4.4. *(Consistency) Under the following regularity conditions,*

A.1 *Identifiability of the model, i.e. $\theta \neq \theta_0 \Leftrightarrow f(Y|\theta) \neq f(Y|\theta_0)$;*

A.2 *The parameter $\Theta \subset \mathbb{R}^k$ is compact;*

A.3 *$\hat{\ell}_n(\theta)$ is differentiable with respect to $\theta \in \Theta$.*

The sample weighted MLE $\hat{\theta}_n$ such that $\mathbb{E}_{Y\pi}[U_n(\hat{\theta}_n)] = 0$ is consistent, i.e. $\hat{\theta}_n \xrightarrow{P} \theta_0$.

Proof. We have

$$\frac{1}{N^2}\hat{\ell}_n(\theta) = -\frac{1}{2N^2}\left(\log|\Xi(\theta)| + \sum_{i=1,j=1}^{N}\frac{R_{ij}}{\pi_{ij}}V_{ij}(\theta)\right),$$

where R_{ij} is the sampling indicator function. By the Uniform Law of Large Numbers in Theorem 5.4.2 for correlated random variables, for each θ, we have

$$\sup_{\theta\in\Theta}\frac{1}{N}\left|\hat{\ell}_n(\theta) - \mathbb{E}_{\theta_0}\left[\ell_N(\theta)\right]\right| \xrightarrow{\mathrm{P}} 0,$$

where $\mathbb{E}_{\theta_0}$ denotes the expectation with respect to a distribution parameterized by θ_0. Therefore,

$$\hat{\theta}_n = \operatorname*{argmax}_{\theta\in\Theta}\hat{\ell}_n(\theta) \xrightarrow{\mathrm{P}} \operatorname*{argmax}_{\theta\in\Theta}\mathbb{E}_{\theta_0}\left[\ell_N(\theta)\right].$$

Since $\log(x)$ is concave, and by Jensen's inequality, we have

$$\begin{aligned}
\mathbb{E}_{\theta_0}[\ell_N(\theta)] - \mathbb{E}_{\theta_0}[\ell_N(\theta_0)] &= \mathbb{E}_{\theta_0}\left[\log\frac{f(Y|\theta)}{f(Y|\theta_0)}\right] \\
&\leq \log\mathbb{E}_{\theta_0}\left[\frac{f(Y|\theta)}{f(Y|\theta_0)}\right] \\
&= \log\int\frac{f(y|\theta)}{f(y|\theta_0)}f(y|\theta_0)dy \\
&= \log\int f(y|\theta)dy \\
&= 0.
\end{aligned}$$

Therefore, $\mathbb{E}_{\theta_0}\left[\ell_N(\theta)\right]$ is maximized at $\theta = \theta_0$, and $\hat{\theta}_n \xrightarrow{\mathrm{P}} \theta_0$. □

The sample weighted likelihood estimator $\hat{\ell}_n(\theta)$ is consistent under the Design 1 and Design 2 if the true parameter θ_0 of θ is a solution of $\mathbb{E}_{Y\pi}[U_n(\theta)] = 0$ (as shown in Figure 5.17) and the numerator in Eq 5.10 goes to 0, i.e.,

$$\sum_{i=1,j=1}^{N}\sum_{k=1,l=1}^{N}\left(\frac{\pi_{ijkl}}{\pi_{ij}\pi_{kl}} - 1\right)V_{ij}(\theta)V_{kl}(\theta) \to 0. \tag{5.11}$$

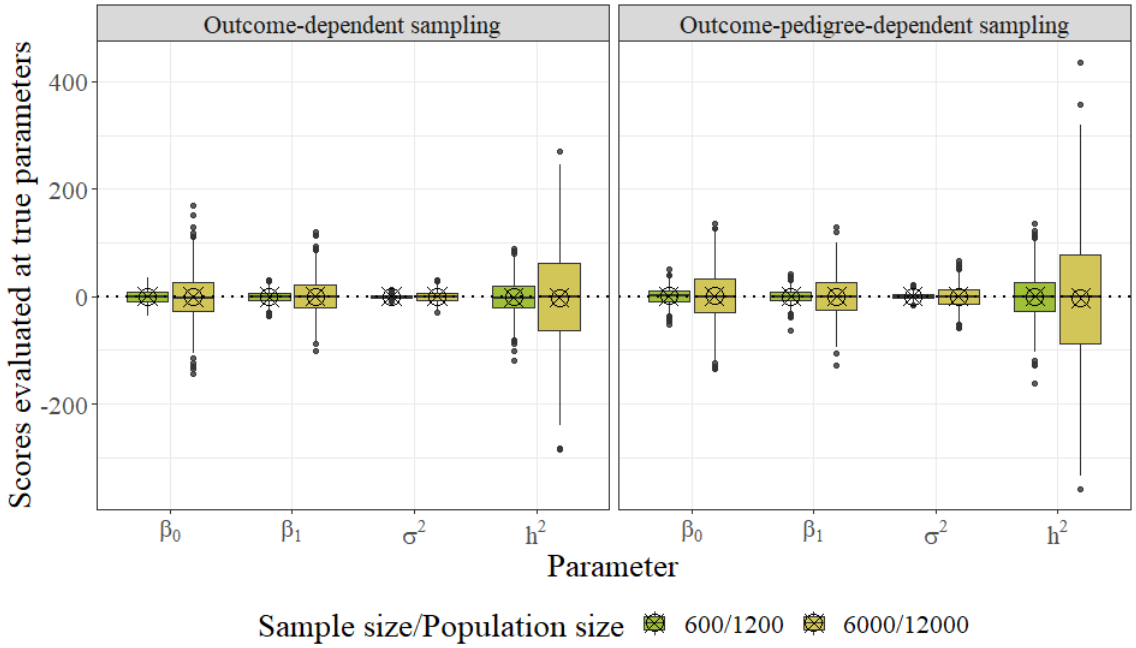

Fig. 5.17 The distribution of evaluated score function of a thousand samples generated from the simulated nuclear family datasets under the outcome-dependent and outcome-pedigree-dependent sampling design in section 5.2.

Let S_1 be the set of individuals sampled in *Step 1* and S_2 be the set of individuals sampled in *Step 2* in Design 1 or Design 2. Recall that n_1 is the number of individuals who are always sampled, and n_2 is the number of randomly sampled individuals from the rest of the unselected individuals. Then, we have

$$\pi_{ij} = \begin{cases} 1 & \text{if } i,j \in S_1 \\ \dfrac{n_2}{N-n_1} & \text{if } i \in S_1, j \in S_2 \text{ or } i \in S_2, j \in S_1 \text{ or } i = j \in S_2 \\ \dfrac{n_2}{N-n_1}\dfrac{n_2-1}{N-n_1-1} & \text{if } i,j \in S_2 \text{ and } i \neq j \end{cases}$$

The calculation of the left-hand-side of Eq 5.11 can be devided into six parts.

C.1 All subjects belong to S_1, i.e., $\{i,j,k,l\} \in S_1$:

$$\sum_{i=1,j=1}^{N}\sum_{k=1,l=1}^{N}\left(\frac{\pi_{ijkl}}{\pi_{ij}\pi_{kl}}-1\right)V_{ij}(\theta)V_{kl}(\theta) = \sum_{i=1,j=1}^{N}\sum_{k=1,l=1}^{N}(1-1)V_{ij}(\theta)V_{kl}(\theta) = 0$$

C.2 One subject belongs to S_2 with the others belong to S_1, i.e., $i \in S_2, \{j,k,l\} \in S_1$:

$$\sum_{i=1,j=1}^{N}\sum_{k=1,l=1}^{N}\left(\frac{\pi_{ijkl}}{\pi_{ij}\pi_{kl}}-1\right)V_{ij}(\theta)V_{kl}(\theta)=\sum_{i=1,j=1}^{N}\sum_{k=1,l=1}^{N}\left(\frac{\pi_i}{\pi_i}-1\right)V_{ij}(\theta)V_{kl}(\theta)=0,$$

similarly for $j \in S_2, \{i,k,l\} \in S_1$, $k \in S_2, \{i,j,l\} \in S_1$, $l \in S_2, \{i,j,k\} \in S_1$.

C.3 One pair belongs to S_1 and one pair belongs to S_2, i.e., $\{i,j\} \in S_1, \{k,l\} \in S_2$:

$$\sum_{i=1,j=1}^{N}\sum_{k=1,l=1}^{N}\left(\frac{\pi_{ijkl}}{\pi_{ij}\pi_{kl}}-1\right)V_{ij}(\theta)V_{kl}(\theta)=\sum_{i=1,j=1}^{N}\sum_{k=1,l=1}^{N}\left(\frac{\pi_{kl}}{\pi_{kl}}-1\right)V_{ij}(\theta)V_{kl}(\theta)=0,$$

similarly for $\{k,l\} \in S_1, \{i,j\} \in S_2$.

C.4 Subjects in a pair belong to different groups, i.e., $\{i,k\} \in S_1, \{j,l\} \in S_2$:

$$\begin{aligned}
&\sum_{i=1,j=1}^{N}\sum_{k=1,l=1}^{N}\left(\frac{\pi_{ijkl}}{\pi_{ij}\pi_{kl}}-1\right)V_{ij}(\theta)V_{kl}(\theta)\\
&=\sum_{i=1,j=1}^{N}\sum_{k=1,l=1}^{N}\left(\frac{\pi_{jl}}{\pi_j\pi_l}-1\right)V_{ij}(\theta)V_{kl}(\theta)\\
&=\sum_{i=1,j=1}^{N}\sum_{k=1,l=1}^{N}\left(\frac{n_2}{N-n_1}\frac{n_2-1}{N-n_1-1}\left(\frac{N-n_1}{n_2}\right)^2-1\right)V_{ij}(\theta)V_{kl}(\theta)\\
&=\sum_{i=1,j=1}^{N}\sum_{k=1,l=1}^{N}\left(\frac{n_2-1}{N-n_1-1}\frac{N-n_1}{n_2}-1\right)V_{ij}(\theta)V_{kl}(\theta)\\
&=\sum_{i=1,j=1}^{N}\sum_{k=1,l=1}^{N}\left(\frac{n_2-1}{n_2}\frac{N-n_1}{N-n_1-1}-1\right)V_{ij}(\theta)V_{kl}(\theta)\\
&=\sum_{i=1,j=1}^{N}\sum_{k=1,l=1}^{N}\left(\left(1+O\left(\frac{1}{n_2}\right)\right)\left(1+O\left(\frac{1}{N-n_1}\right)\right)-1\right)V_{ij}(\theta)V_{kl}(\theta)\\
&=\sum_{i=1,j=1}^{N}\sum_{k=1,l=1}^{N}\left(1+O\left(\frac{1}{n_2}\right)+O\left(\frac{1}{N-n_1}\right)+O\left(\frac{1}{n_2(N-n_1)}\right)-1\right)V_{ij}(\theta)V_{kl}(\theta)\\
&=\sum_{i=1,j=1}^{N}\sum_{k=1,l=1}^{N}\left(O\left(\frac{1}{n_2}\right)+O\left(\frac{1}{N-n_1}\right)\right)V_{ij}(\theta)V_{kl}(\theta)
\end{aligned}$$

$\to 0$ as $n_1 \to \infty, n_2 \to \infty, N \to \infty$.

Similarly for $\{j,l\} \in S_1, \{i,k\} \in S_2$, $\{j,k\} \in S_1, \{i,l\} \in S_2$, $\{i,l\} \in S_1, \{j,k\} \in S_2$.

C.5 One subject belongs to S_1 with the others belong to S_2, i.e., $i \in S_1, \{j,k,l\} \in S_2$:

$$
\begin{aligned}
&\sum_{i=1,j=1}^{N}\sum_{k=1,l=1}^{N}\left(\frac{\pi_{ijkl}}{\pi_{ij}\pi_{kl}}-1\right)V_{ij}(\theta)V_{kl}(\theta)\\
&=\sum_{i=1,j=1}^{N}\sum_{k=1,l=1}^{N}\left(\frac{\pi_{jkl}}{\pi_{j}\pi_{kl}}-1\right)V_{ij}(\theta)V_{kl}(\theta)\\
&=\sum_{i=1,j=1}^{N}\sum_{k=1,l=1}^{N}\left(\frac{n_2}{N-n_1}\frac{n_2-1}{N-n_1-1}\frac{n_2-2}{N-n_1-2}\left(\frac{N-n_1}{n_2}\frac{N-n_1}{n_2}\frac{N-n_1-1}{n_2-1}\right)-1\right)\\
&\quad\times V_{ij}(\theta)V_{kl}(\theta)\\
&=\sum_{i=1,j=1}^{N}\sum_{k=1,l=1}^{N}\left(\frac{n_2-2}{N-n_1-2}\frac{N-n_1}{n_2}-1\right)V_{ij}(\theta)V_{kl}(\theta)\\
&=\sum_{i=1,j=1}^{N}\sum_{k=1,l=1}^{N}\left(\frac{n_2-2}{n_2}\frac{N-n_1}{N-n_1-2}-1\right)V_{ij}(\theta)V_{kl}(\theta)\\
&=\sum_{i=1,j=1}^{N}\sum_{k=1,l=1}^{N}\left(O\left(\frac{1}{n_2}\right)+O\left(\frac{1}{N-n_1}\right)\right)V_{ij}(\theta)V_{kl}(\theta)\\
&\to 0 \text{ as } n_1\to\infty, n_2\to\infty, N\to\infty.
\end{aligned}
$$

Similarly for $j \in S_1, \{i,k,l\} \in S_2$, $k \in S_1, \{i,j,l\} \in S_2$, $l \in S_1, \{i,j,k\} \in S_2$.

C.6 All subjects belong to S_2, i.e., $\{i,j,k,l\} \in S_2$:

$$
\begin{aligned}
&\sum_{i=1,j=1}^{N}\sum_{k=1,l=1}^{N}\left(\frac{\pi_{ijkl}}{\pi_{ij}\pi_{kl}}-1\right)V_{ij}(\theta)V_{kl}(\theta)\\
&=\sum_{i=1,j=1}^{N}\sum_{k=1,l=1}^{N}\left(\frac{\pi_{ijkl}}{\pi_{ij}\pi_{kl}}-1\right)V_{ij}(\theta)V_{kl}(\theta)\\
&=\sum_{i=1,j=1}^{N}\sum_{k=1,l=1}^{N}\left(\frac{n_2}{N-n_1}\frac{n_2-1}{N-n_1-1}\frac{n_2-2}{N-n_1-2}\frac{n_2-3}{N-n_1-3}\right.\\
&\quad\left.\times\left(\frac{N-n_1}{n_2}\frac{N-n_1-1}{n_2-1}\right)^2-1\right)V_{ij}(\theta)V_{kl}(\theta)\\
&=\sum_{i=1,j=1}^{N}\sum_{k=1,l=1}^{N}\left(\frac{n_2-2}{N-n_1-2}\frac{n_2-3}{N-n_1-3}\frac{N-n_1}{n_2}\frac{N-n_1-1}{n_2-1}-1\right)V_{ij}(\theta)V_{kl}(\theta)\\
&=\sum_{i=1,j=1}^{N}\sum_{k=1,l=1}^{N}\left(\frac{n_2-2}{n_2}\frac{n_2-3}{n_2-1}\frac{N-n_1}{N-n_1-2}\frac{N-n_1-1}{N-n_1-3}-1\right)V_{ij}(\theta)V_{kl}(\theta)\\
&=\sum_{i=1,j=1}^{N}\sum_{k=1,l=1}^{N}\left[\left(1+O\left(\frac{1}{n_2}\right)\right)\left(1+O\left(\frac{1}{n_2}\right)\right)\left(1+O\left(\frac{1}{N-n_1}\right)\right)\right.\\
&\quad\left.\times\left(1+O\left(\frac{1}{N-n_1}\right)\right)-1\right]V_{ij}(\theta)V_{kl}(\theta)\\
&=\sum_{i=1,j=1}^{N}\sum_{k=1,l=1}^{N}\left[\left(1+O\left(\frac{1}{n_2}\right)\right)\left(1+O\left(\frac{1}{N-n_1}\right)\right)-1\right]V_{ij}(\theta)V_{kl}(\theta)\\
&=\sum_{i=1,j=1}^{N}\sum_{k=1,l=1}^{N}\left(O\left(\frac{1}{N-n_1}\right)+O\left(\frac{1}{n_2}\right)\right)V_{ij}(\theta)V_{kl}(\theta)\\
&\to 0 \text{ as } n_1\to\infty, n_2\to\infty, N\to\infty.
\end{aligned}
$$

Therefore,

$$\sum_{i=1,j=1}^{N}\sum_{k=1,l=1}^{N}\left(\frac{\pi_{ijkl}}{\pi_{ij}\pi_{kl}}-1\right)V_{ij}(\theta)V_{kl}(\theta)\to 0,$$

and the sample weighted likelihood estimator $\hat{\ell}_n(\theta)$ is consistent under the two sampling designs in section 5.2 by Theorem 5.4.4.

5.5 Summary

So far, there are limited number of methods developed for fitting linear mixed model with correlated individuals under complex design (e.g., [112, 114, 165]). In particular, the existing

methods assume the sampling clusters are the same as the clusters in the random effect. This is possible when the correlation structure is a block diagonal matrix (e.g. the right plot of Figure 5.18), but infeasible for complex correlation structure (e.g. the left plot of Figure 5.18). Huang [68] showed in his PhD thesis that the pairwise pseudolikelihood proposed by Rao et al. [114] and Yi et al. [165] is also valid when the sampling clusters are not the same as the model cluster. However, there has been no published papers or available packages for fitting linear mixed model with complex correlation structure under two-phasing sampling. In this chapter, a weighted maximum likelihood approach that takes advantage of knowing the population kinship structure was developed to fill in such gap in the literature.

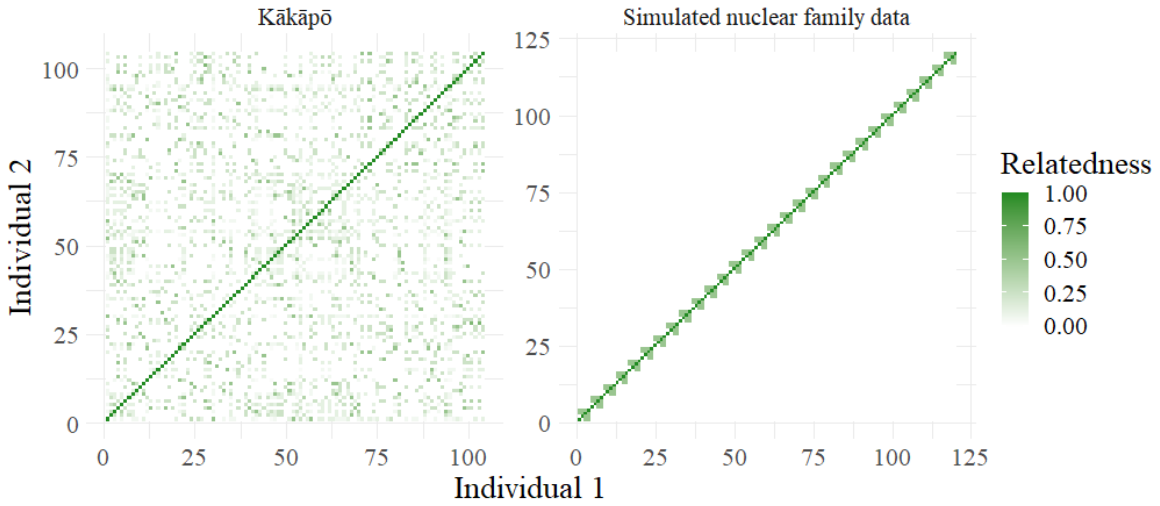

Fig. 5.18 Comparing the correlation structure between kākāpō and the simulated nuclear family data. As all the nuclear families have the same correlation structure, only 10% of the families are plotted for a better view.

The case studies of kākāpō and simulated nuclear family data demonstrated that the sampling bias can be corrected by re-weighting the samples regardless of correlation structure. As population size and sample size increase, the weighted sample estimation of the log-likelihood with HT-type RSS estimator converge to the population estimation, but the sample estimation is quite variable for small datasets. On the other hand, the the log-likelihood with SYG-type RSS estimator is not consistent but gives better inference on small datasets.

While the full likelihood approach properly accounts for covariance structure and the missing mechanism, it cannot straightforwardly be applied to general designs and correlation structures. In comparison to the pairwise likelihood that only account for the pairwise correlations, it was expected that there would be a gain in the efficiency by utilizing the sample covariance matrix in the proposed weighted likelihood. However, no improvement

was shown in the simulation study, and the two methods do not seem to be extracting the same information from the data. A possible future direction is to develop a likelihood estimator that combines the information used by the weighted likelihood and the pairwise pseudolikelihood and is more efficient than either.

Chapter 6

Generalized linear mixed models under two-phase sampling

The aim of the chapter is to extend the proposed weighted MLE approach for linear mixed models in Chapter 5 to fit generalized linear mixed models for binary traits (e.g. disease status). The conventional way to analyze a binary trait is to assume that it has a continuous normally-distributed liability that measures the susceptibility to the trait, and the binary phenotype is determined depending on whether the liability exceeds the threshold. This is called a liability threshold model and was first introduced into human genetics by Falconer [50]. In non-genetic contexts, the liability threshold model is often referred to as the probit-normal model as a class of the generalized linear mixed models. The probit link function has several advantages over the logit link function [98]. In particular, a probit-normal model with a single observational-level random effects term can be reduced to the usual probit model with a different residual term. More importantly, it allows us to incorporate the sampling information into parameter estimation.

For generalized linear mixed models, there is no closed-form solution available for the likelihood function. Hence some approximation methods must be used to obtain parameter estimation. Common classes of approaches include quasi-likelihood, numerical integral (Laplace approximation and Gaussian quadrature) and Monte Carlo methods (Newton-Raphson and the EM algorithm). Among these common approaches, it is relatively straightforward to re-weight the observations under informative sampling in the EM algorithm, and it takes a very similar form to the linear mixed models for generalized linear mixed models with a probit link function. McCulloch proposed a Monte Carlo EM algorithm with a Gibbs sampler to maximize the likelihood of a probit-normal model with independent random effects [98], and I will extend McCulloch's model to allow correlated random effects in this chapter, which is similar to the model proposed by Chan and Kuk [31].

This chapter starts with an introduction to the liability threshold model in section 6.1 and the Monte Carlo EM algorithm with Gibbs sampler in section 6.2. Then section 6.3 illustrates the proposed weighted maximum likelihood method for probit-linear mixed models with correlated random effects, and uses simulated data to show that the bias induced by informative sampling can be corrected.

6.1 Liability threshold model

Similar to quantitative traits, the trait liability y^* can be described by an additive model that is a sum of the fixed effect of a genetic covariate X, and an error term u, which comprises observation-level genetic random effects and random errors,

$$y^* = X\beta + u,\ y^* \sim \mathcal{N}(X\beta, \Xi),$$
$$y_i = \mathbb{1}_{y_i^* > t},\ i = 1, 2, \cdots, N,$$

where $u \sim \mathcal{N}(0, \Xi)$ and t is the liability threshold, which can be calculated by taking the inverse of the cumulative distribution function of the trait prevalence. Let Φ denote the kinship matrix that specifies the correlation between levels of the genetic random effect u, and I be an identity matrix. The variance-covariance matrix of the liability y^* can be written as

$$\Xi = \sigma_g^2 \Phi + \sigma_e^2 I.$$

For identifiability, we define the total phenotypic variance on the liability scale to be 1. Hence the heritability of liability is defined by

$$h^2 = \frac{\sigma_g^2}{\sigma_g^2 + \sigma_e^2} = \sigma_g^2,$$

and the variance-covariance matrix Ξ becomes

$$\Xi = h^2 \Phi + (1 - h^2) I.$$

The heritability of liability can be converted to the heritability of the observed binary trait with an assumption of the trait prevalence in the population. However, this chapter focus on estimating heritability on the liability scale and more mathematical details of the conversion between the heritability of liability and the heritability of the binary trait is described in [83].

6.2 The Monte Carlo EM algorithm for MLE

Since the continuous liability is unobserved, to apply the EM algorithm to estimate the model parameters, we need to take the expectation of the log-likelihood conditional on the observed binary data, which can be computed as follows

$$\begin{aligned}
\ell_N(\beta, h^2; y) =& \mathbb{E}[\ell(\beta, h^2; y^*)|y] \\
=& -\frac{1}{2}\left(\log|\Xi| + \mathbb{E}[(y^* - x\beta)^T \Xi^{-1}(y^* - x\beta)|y]\right) \\
=& -\frac{1}{2}\left(\log|\Xi| + \text{tr}\left(\mathbb{E}[\Xi^{-1}(y^* - x\beta)(y^* - x\beta)^T|y]\right)\right) \\
=& -\frac{1}{2}\left(\log|\Xi| + \text{tr}\left(\Xi^{-1}\left(\mathbb{V}(y^*|y) + (\mathbb{E}[y^*|y] - X\beta)(\mathbb{E}[y^*|y] - X\beta)^T\right)\right)\right) \\
=& -\frac{1}{2}\left(\log|\Xi| + \text{tr}\left(\Xi^{-1}\mathbb{V}(y^*|y)\right) + (\mathbb{E}[y^*|y] - X\beta)^T \Xi^{-1}(\mathbb{E}[y^*|y] - X\beta)\right).
\end{aligned} \tag{6.1}$$

Note that this equation is identical to McCulloch's approach [98] but allows correlated random effects. In equation 6.1, the conditional mean $\mathbb{E}[y^*|y]$ and variance $\mathbb{V}(y^*|y)$ have no closed-form expression available, so they require extra computations either by numerical integration or Gibbs sampling. The R package `tmvtnorm` [160] computes the mean vector and covariance matrix for the truncated multivariate normal distribution by numerical integration. Numerical integration is fast at low dimensions (e.g. the size of human families), but the computational time climbs exponentially as the dimension increases. Since the main pedigree in the kākāpō population includes 152 individuals, we use Gibbs sampling to compute the conditional mean and variance as numerical integration becomes computationally infeasible.

Let N be the total number of individuals and $y^{*(0)} = (y_1^{*(0)}, \ldots, y_N^{*(0)})$ be the initial values that are consistent with the observed phenotype y, we can generate $y^{*(k)} = (y_1^{*(k)}, \ldots, y_N^{*(k)})$ using the following procedure

$$\begin{gathered}
y_1^{*(k)} \text{ from } f(y_1^*|y_2^{*(k-1)}, y_3^{*(k-1)}, \ldots, y_N^{*(k-1)}, y; \hat{\beta}, \hat{h^2}) \\
\vdots \\
y_i^{*(k)} \text{ from } f(y_i^*|y_1^{*(k)}, y_2^{*(k)}, \ldots, y_{i-1}^{*(k)}, y_{i+1}^{*(k-1)}, \ldots, y_N^{*(k-1)}, y; \hat{\beta}, \hat{h^2}) \\
\vdots \\
y_N^{*(k)} \text{ from } f(y_N^*|y_1^{*(k)}, y_2^{*(k)}, \ldots, y_{N-1}^{*(k)}, y; \hat{\beta}, \hat{h^2})
\end{gathered}$$

where $\hat{\beta}$ and $\hat{h^2}$ are the estimates at a particular EM iteration. Since the liability y^* is distributed $\mathcal{N}(X\beta, \Xi)$, $y_i^*|y_{\mathcal{J}}^*$ where $\mathcal{J} = \{1, \ldots, i-1, i+1, \ldots, N\}$ also follows a normal distribution. It was shown in [27] that the conditional mean and standard deviation of $y_i^*|y_{\mathcal{J}}^*$

can be computed as

$$\mu_{i|\mathcal{J}} = y_i^* - \frac{\left[(\Xi)^{-1}(y^* - X\beta)\right]_i}{(\Xi^{-1})_{ii}}, \tag{6.2}$$

$$\sigma_{i|\mathcal{J}} = \frac{1}{(\Xi^{-1})_{ii}}. \tag{6.3}$$

For the kākāpō data with small population size, the complicated random effects structures does not greatly increase the computational time of the Gibbs sampling since the inverse of variance-covariance matrix Ξ is only calculated once at every EM iteration. But computation bottleneck becomes the matrix inversion as population size grows. We can then simulate y^* from a truncated normal distribution to ensure it is consistent with the observed y:

$$\begin{cases} y_i^* \sim \mathcal{N}(\mu_{i|\mathcal{J}}, \sigma^2_{i|\mathcal{J}}), -\infty \leq y_i^* \leq t \text{ if } y_i = 0, \\ y_i^* \sim \mathcal{N}(\mu_{i|\mathcal{J}}, \sigma^2_{i|\mathcal{J}}), t \leq y_i^* \leq \infty \text{ if } y_i = 1. \end{cases} \tag{6.4}$$

After discarding the burn-in period, we can approximate the conditional mean $\mathbb{E}[y^*|y]$ and variance $\mathbb{V}(y^*|y)$ from the realizations of y^* generated by the Gibbs sampler from $f(y^*|y)$. That is,

$$\hat{\mathbb{E}}[y^*|y] = \frac{1}{K}\sum_{k=B+1}^{B+K} y^{*(k)}, \tag{6.5}$$

$$\hat{\mathbb{V}}(y^*|y) = \frac{1}{K}\sum_{k=B+1}^{B+K} (y^{*(k)} - \hat{\mathbb{E}}[y^*|y])(y^{*(k)} - \hat{\mathbb{E}}[y^*|y])^T, \tag{6.6}$$

where B denotes the burn-in period, and K is the total number of iterations in the Gibbs sampling. We are now able to maximize the log-likelihood of the generalized linear mixed model with the probit link function using the Monte Carlo EM algorithm.

The MCEM algorithm

Let $\beta^{(0)}$ and $h^{2(0)}$ be the initial values. Set $t = 0$.

Step 1: (E-step) Given $\beta^{(t)}$ and $h^{2(t)}$, calculate $\hat{\mathbb{E}}[y^*|y]^{(t)}$ and $\hat{\mathbb{V}}(y^*|y)^{(t)}$ using Gibbs sampler:

Step 1.1: For $k = 1,\ldots,K$ and $i = 1,\ldots,N$, generate $y_i^{*(k)}$ from a truncated normal distribution

$$f(y_i^*|y_1^{*(k)}, y_2^{*(k)}, \ldots, y_{i-1}^{*(k)}, y_{i+1}^{*(k-1)}, \ldots, y_N^{*(k-1)}, y; \beta^{(t)}, h^{2(t)})$$

with $\mu_{i|\mathcal{J}}$ in Eq.6.2 and $\sigma_{i|\mathcal{J}}$ in Eq.6.3, where $\mathcal{J} = \{1,\ldots,i-1,i+1,\ldots,N\}$.

Step 1.2: Discard the burn-in period, and obtain the estimates of conditional mean $\hat{\mathbb{E}}[y^*|y]^{(t)}$ using Eq.6.5 and the conditional variance $\hat{\mathbb{V}}(y^*|y)^{(t)}$ using Eq.6.6.

Step 2: (M-step) Set

$$h^{2^{(t+1)}} = \underset{h^2\in[0,1]}{\operatorname{argmax}}\, \ell_N(\beta^{(t)}, h^{2^{(t)}}; y),$$

which is given in Eq.6.1, and

$$\beta^{(t+1)} = \left(X^T {\Xi^{(t+1)}}^{-1} X\right)^{-1} X^T {\Xi^{(t+1)}}^{-1} \hat{\mathbb{E}}[y^*|y]^{(t)}.$$

Step 3: Once convergence is reached, set $\hat{\beta} = \beta^{(t+1)}$, $\hat{h^2} = h^{2^{(t+1)}}$, otherwise increase t by 1 and return to *Step* 1.

6.2.1 Examples

This section demonstrates the application of the Monte Carlo EM algorithm described in section 6.2 to the Weil data that was analyzed by McCulloch [98], and a simulated dataset from Kim et al. [76]. The model for the Weil data only includes independent random effects, and the model for the simulated dataset allows correlated random effects.

Model with independent random effects

In the Weil data [153], 16 pregnant rats received a control diet and 16 received a chemically treated diet, and the survival of the rats are recorded after 4 and 21 days. McCulloch [98] assumes the following latent survival model

$$\begin{aligned} y^*_{ijk} &= X_{ijk}\beta_i + u_{ij} + \varepsilon_{ijk}, \\ y_{ijk} &= \mathbb{1}_{y^*_{ijk}\geq 0}, \end{aligned} \tag{6.7}$$

where i indexes treatment/control, j indexes litter, and k indexes rat within a litter, X_{ijk} indicate if the rats receives treatment or control diet, β_i are the treatment/control effects and

u_{ij} are the random litter effects. Putting Eq.6.7 in vector form, we have

$$y^* = X\beta + Zu + \varepsilon, \varepsilon \sim \mathcal{N}(0, \Xi)$$

where Z is the design matrix, $u \sim \mathcal{N}(0, \sigma^2 I)$ and $\Xi = \sigma^2 ZZ^T + I$. Since the random litter effects are independent, there exists a closed form solution for the variance estimator,

$$\hat{\sigma}^2 = \frac{\mathbb{E}[u^T u|y]}{n_l},$$

where n_l is the number of litters and

$$\begin{aligned}\mathbb{E}[u^T u|y] =& \sigma^4 \text{tr}\left(\Xi^{-1} ZZ^T \Xi^{-1} \mathbb{V}(y^*|y)\right) + \sigma^4 (\mathbb{E}[y^*|y] - X\beta)^T \Xi^{-1} ZZ^T \Xi^{-1} (\mathbb{E}[y^*|y] - X\beta) \\ &+ \text{tr}(\sigma^2 I - \sigma^4 Z^T \Xi^{-1} Z).\end{aligned}$$

The comparison of the method described in section 6.2 that uses linear optimization to McCulloch's approach that uses closed-form expression is shown in Figure 6.1. The plot shows that the two approaches are almost identical and converge to the same parameter values provided in [98]. As mentioned in McCulloch's paper, it is expected that the EM algorithm with a closed-form solution would take more iterations for the control group when the estimate is close to the boundary of the parameter space.

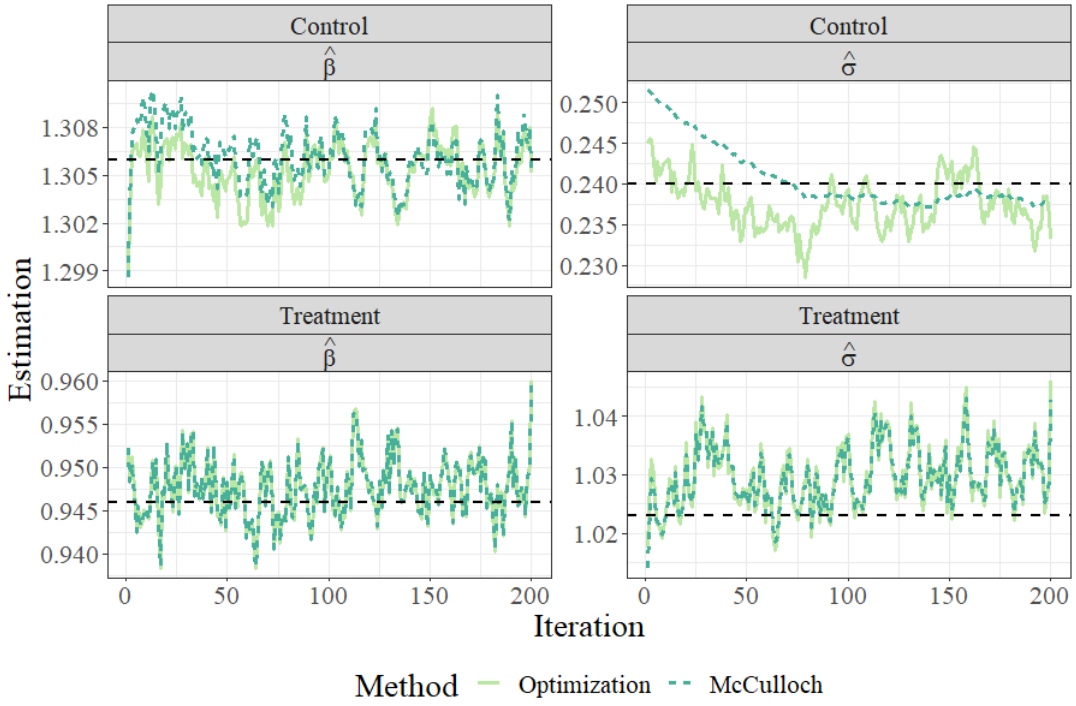

Fig. 6.1 The Weil data: The MCEM algorithm for MLE, where the dashed lines are the estimations in [98].

Model with correlated random effects

Kim et al. developed a method for heritability estimation based on Liability Threshold Model for binary traits (LTMH) [76]. The parameterization in LTMH is the same as in section 6.2, but the algorithms are slightly different, and this will be discussed later. We use a simulated dataset to show that the Monte Carlo EM algorithm described in section 6.2 converges to the true parameter values as LTMH (see Figure 6.2). It was found in simulation studies that the number of Gibbs samples is unimportant in early EM iterations, but increasing the number of Gibbs samples as estimation gets closer to the true parameter value helps to capture the correlation between samples better. As opposed to a model with correlated random effects, McCulloch found that a larger number of Gibbs samples does not result in any improvement for models with independent random effect [98].

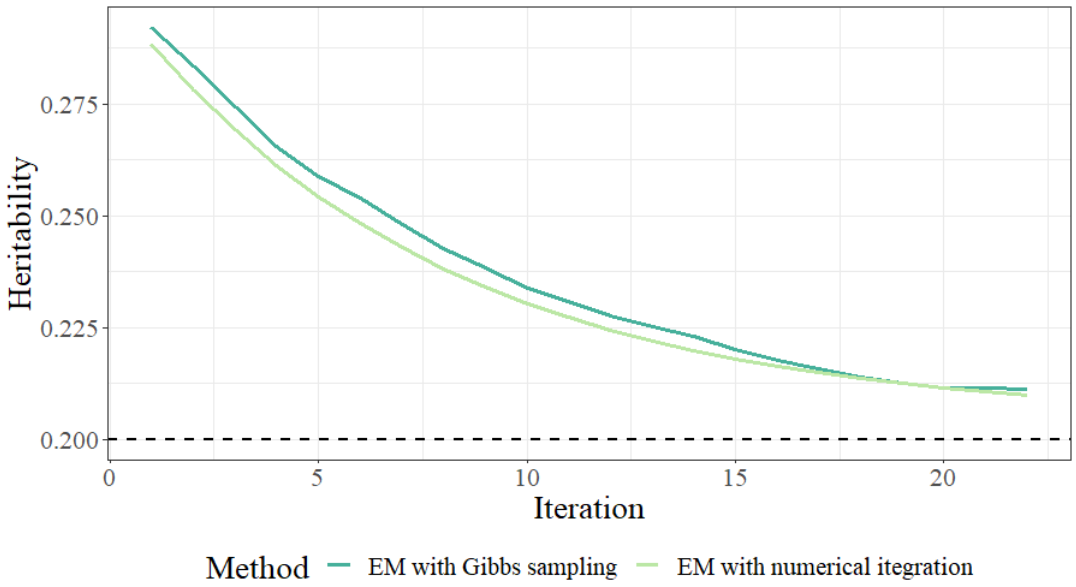

Fig. 6.2 The LTMH example dataset contains 500 families that were randomly generated with heritability of 0.2 and prevalence of 0.1. The number of Gibbs samples is increased as the estimate converges to the true value.

LTMH considers independent families with small family sizes, such as the human population. Hence the conditional mean and variance can be computed separately for each family using numerical integration in a reasonable amount of time. In contrast, the Gibbs sampler takes longer to reach the same results as numerical integration for such datasets with a large number of small families, but the computational time does not increase exponentially as the family size increases. Furthermore, LTMH uses a Newton-Raphson algorithm to maximize the log-likelihood at each maximization step, and it is not straightforward to incorporate the sampling weights in the algorithm.

6.3 The weighted MLE

Section 6.2 described an approach for fitting linear mixed models with correlated random effects. In this section, I will extend this approach to fit such models under two-phase sampling.

Let R be the sampling indicator with $R_i = 1$ if the i-th individual is sampled in phase II, $R_i = 0$ otherwise, and π be the matrix pairwise sampling probabilities. Let $\odot$ denote the symbol for element-wise matrix multiplication. Then we can re-weight the log-likelihood in

Eq.6.1 using sampling weights,

$$\hat{\ell}_n(\beta,h^2;y) = -\frac{1}{2}\left[\log|\Xi| + \text{tr}\left(\frac{R}{\pi}\odot\Xi^{-1}\mathbb{V}(y^*|y)\right)\right.$$
$$\left.+(\mathbb{E}[y^*|y]-X\beta)^T\left(\frac{R}{\pi}\odot\Xi^{-1}\right)(\mathbb{E}[y^*|y]-X\beta)\right]. \tag{6.8}$$

When data is available for all individuals, $R=1$ and $\pi=1$, hence Eq.6.8 is equivalent to Eq.6.1. Let O denote the phase II samples whose genotype and phenotype are observed, and M denotes the individuals with missing data. We can reorder terms in Eq.6.8 by

$$y=\begin{pmatrix}y_O\\y_M\end{pmatrix}, X=\begin{pmatrix}X_O\\X_M\end{pmatrix}, \mathbb{E}[y^*|y]=\begin{pmatrix}\mathbb{E}[y^*|y]_O\\\mathbb{E}[y^*|y]_M\end{pmatrix},$$

$$R=\begin{pmatrix}R_{OO}&R_{OM}\\R_{MO}&R_{MM}\end{pmatrix}, \pi=\begin{pmatrix}\pi_{OO}&\pi_{OM}\\\pi_{MO}&\pi_{MM}\end{pmatrix}, \Xi^{-1}=\begin{pmatrix}(\Xi^{-1})_{OO}&(\Xi^{-1})_{OM}\\(\Xi^{-1})_{MO}&(\Xi^{-1})_{MM}\end{pmatrix},$$
$$\mathbb{V}(y^*|y)=\begin{pmatrix}\mathbb{V}(y^*|y)_{OO}&\mathbb{V}(y^*|y)_{OM}\\\mathbb{V}(y^*|y)_{MO}&\mathbb{V}(y^*|y)_{MM}\end{pmatrix} \tag{6.9}$$

Since R_{OO} is a matrix of ones and R_{OM}, R_{MO}, R_{MM} are matrices of zeros, the log-likelihood in Eq.6.8 can be written as

$$\hat{\ell}_n(\beta,h^2;y) = -\frac{1}{2}\left[\log|\Xi| + \text{tr}\left(\frac{1}{\pi_{OO}}\odot\left(\Xi^{-1}\right)_{OO}\mathbb{V}(y^*|y)_{OO}\right)\right.$$
$$\left.+(\mathbb{E}[y^*|y]_O-X_O\beta)^T\left(\frac{1}{\pi_{OO}}\odot\left(\Xi^{-1}\right)_{OO}\right)(\mathbb{E}[y^*|y]_O-X_O\beta)\right] \tag{6.10}$$

However, it is generally impossible to calculate $\mathbb{E}[y^*|y]_O$ and $\mathbb{V}(y^*|y)_{OO}$ in two-phase design as they require X and y for the whole population in Eq.6.2 and Eq.6.4. Under outcome-dependent sampling, we may know y for the whole population, but X is only obtained for the sample. An alternative solution is to approximate $\mathbb{E}[y^*|y]_O$ and $\mathbb{V}(y^*|y)_{OO}$ using $\mathbb{E}[y^*_O|y_O]$ and $\mathbb{V}(y^*_O|y_O)$, and express the log-likelihood as

$$\hat{\ell}_n(\beta,h^2;y) = -\frac{1}{2}\left[\log|\Xi| + \text{tr}\left(\frac{1}{\pi_{OO}}\odot(\Xi^{-1})_{OO}\mathbb{V}(y^*_O|y_O)\right)\right.$$
$$\left.+(\mathbb{E}[y^*_O|y_O]-X_O\beta)^T\left(\frac{1}{\pi_{OO}}\odot(\Xi^{-1})_{OO}\right)(\mathbb{E}[y^*_O|y_O]-X_O\beta)\right]. \tag{6.11}$$

Then $\hat{h^2}$ can be obtained by maximizing the weighted log-likelihood in Eq.6.11 with the constraint $0 < h^2 < 1$, and the weighted MLE for β is given by

$$\hat{\beta} = \left[X_O^T \left(\frac{1}{\pi_{OO}} \odot (\Xi^{-1})_{OO} \right) X_O \right]^{-1} X_O^T \left(\frac{1}{\pi_{OO}} \odot (\Xi^{-1})_{OO} \right) \hat{\mathbb{E}}[y_O^* | y_O].$$

6.3.1 Simulation study

Since it is much harder to make inference on the generalized linear mixed model parameters than on the linear mixed model parameters, especially for small datasets, I use simulated kākāpō datasets with known true parameters to evaluate the performance of proposed weighted MLE method. The data are simulated based on the kākāpō pedigree which reveals a high level of inbreeding in the population, thus it should still be an excellent example to demonstrate the performance of the proposed approach for other species with complex population structures. I also simulated datasets with population structure similar to the human population in order to extend the results to larger populations with a simple population structure.

Simulated kākāpō phenotype with complex pedigree structure

For some values of β and σ^2, the vectors of continuous liability y^* and disease status y for 158 kākāpō in the first simulated dataset are generated by

$$\begin{aligned} y^* &= \beta_0 + X\beta_1 + u, y^* \sim \mathcal{N}(\beta_0 + X\beta_1, \sigma^2\Phi + I), \\ y &= \mathbb{1}_{y^* \geq 0}. \end{aligned}$$

where 1 denotes affected by disease and 0 denotes unaffected (see Figure 6.3). The second dataset is generated with the same parameters but based on a larger pedigree (as shown in Figure 6.4) that is simulated by connecting two subsets of the kākāpō pedigree and contains 292 individuals. This allows us to assess the performance of the method using a pedigree that is larger but retains the complexity of the actual kākāpō pedigree.

Consider an individual-based outcome-dependent design that oversamples the cases and undersamples the controls for both datasets. Figure 6.5 compares the generalized linear mixed model inference under the outcome-dependent sampling using the proposed weighted MLE approach with unweighted MLE. The unweighted MLE is computed by `lme4qtl` [173] that takes the Laplace approximation of the integrand of the likelihood function.

Fig. 6.3 The a subset of the true kākāpō pedigree, including 53 affected kākāpō (colored ones) and 105 unaffected kākāpō.

Fig. 6.4 Simulated kākāpō-like pedigree, including 112 affected kākāpō (colored ones) and 180 unaffected kākāpō.

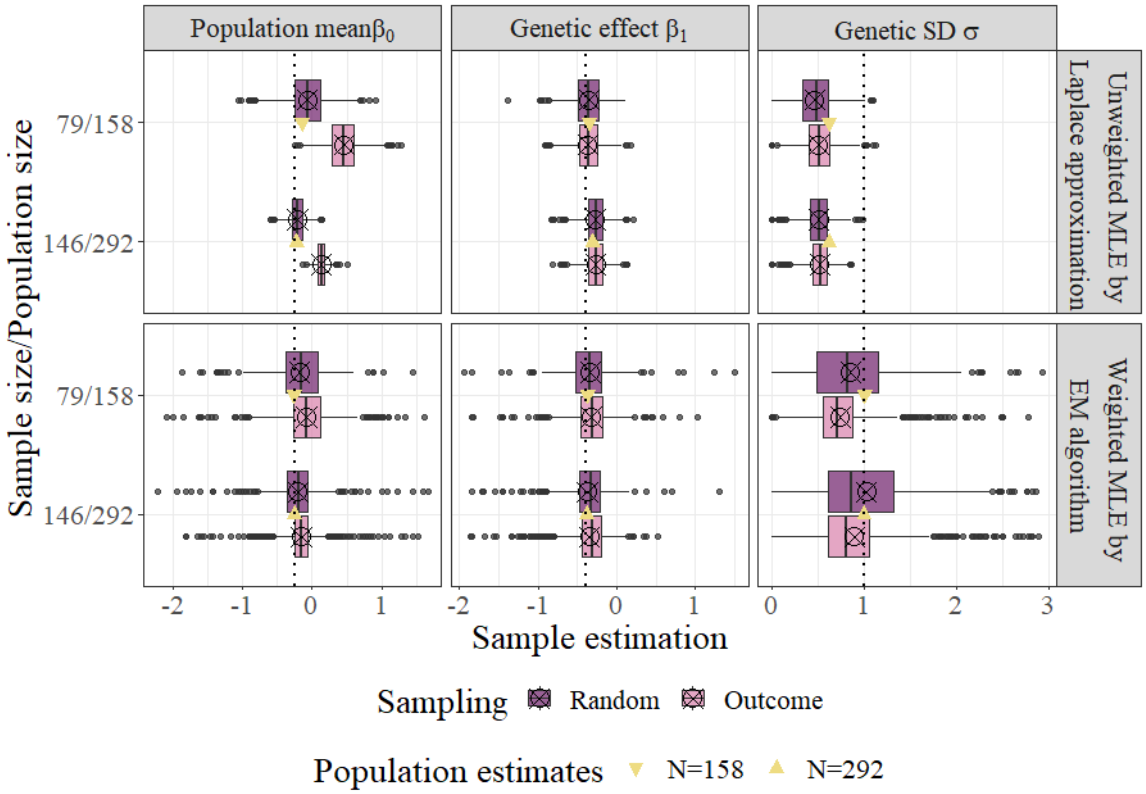

Fig. 6.5 Compare the generalized linear mixed model inference on the two simulated kākāpō datasets under individual-based outcome-dependent using the proposed weighted MLE approach and MLE approach. The model inference under random sampling serves as a baseline of sample estimation. The vertical dotted lines represent the true parameters of the simulated data.

The top row of Figure 6.5 shows the unweighted MLE by treating the samples as the whole population. Although it is less obvious than the overestimation of the population mean, the genetic SD is also overestimated by MLE under outcome dependent sampling compared to random sampling. However, the method that uses the Laplace approximation still underestimates the true genetic SD even under outcome-dependent sampling. This is not surprising as it was shown that the Laplace approximation underestimates the variance components, especially for binary data with small cluster sizes [17, 19].

While both Laplace approaximation and EM algorithm underestimate the population genetic SD using sample data, the sample estimation by EM algorithm in the bottom row of Figure 6.5 is much closer to the true parameters. For the weighted MLE, the proposed EM algorithm tends to underestimate the genetic SD under outcome-dependent sampling compared to random sampling, but the bias reduces as the data size increases. It is possible to simulate data based on a larger complex pedigree to check the hypothesis and also

reduce the variability of the weighted MLE, but it would take too long to compute of the conditional mean and variance in Eq.6.11 using Gibbs sampling as the faster approach numerical integration is impracticable for complex pedigree.

Large dataset with simple pedigree structure

The second set of examples contains 2251 and 10011 individuals from 500 and 2200 families with family sizes ranging from 3 to 6, where parents are considered to be genetically unrelated. To better compare the two methods, a smaller value is chosen for the standard deviation, so the estimator based on the Laplace approximation should be less biased [19]. Consider a family-based outcome-dependent sampling, the idea of the sampling strategy is that families with a higher proportion of affected members have higher sampling probabilities. For example, the distribution of the smaller dataset before and after the sampling is shown in Figure 6.6.

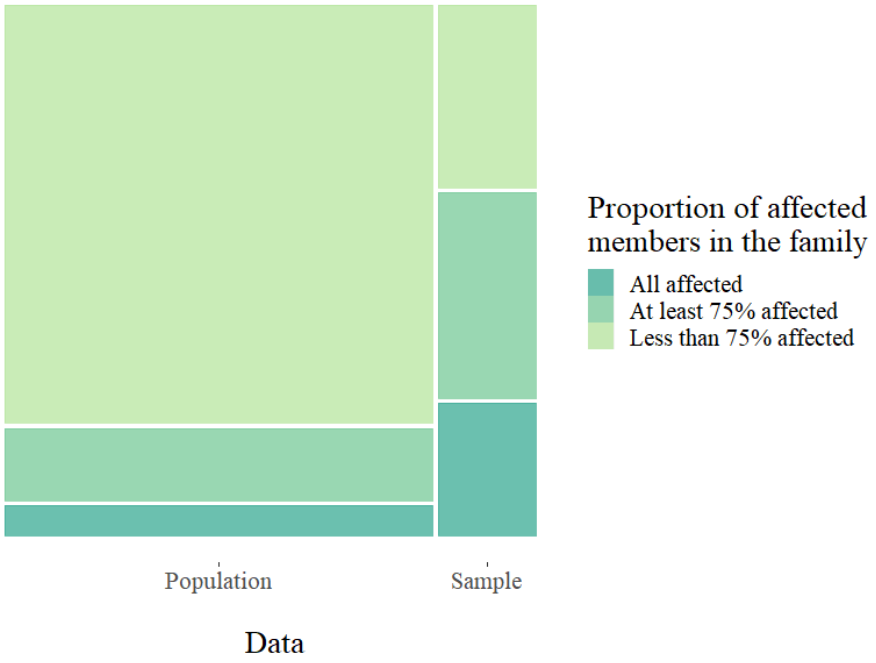

Fig. 6.6 Distribution of the simulated nuclear family data ($N = 2251$) before and after the family-based outcome-dependent sampling. The numbers are the counts of the families.

In contrast to Figure 6.5, it is more obvious in the top row of Figure 6.7 that the genetic SD under outcome-dependent sampling without adjustment is overestimated. Moreover, the sampling bias tends to increase as the data size increases, whereas the proposed EM

algorithm corrects the bias and the accuracy improves as the data size increases (see the bottom row of Figure 6.7).

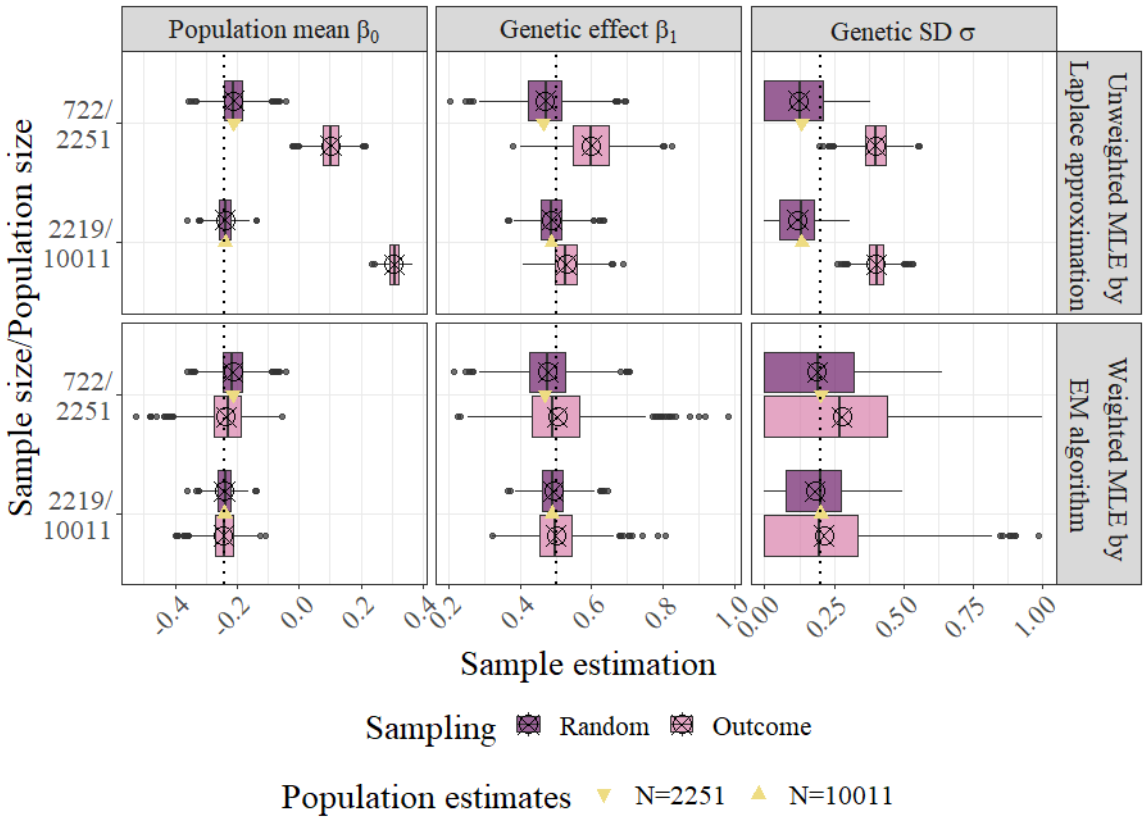

Fig. 6.7 Compare the generalized linear mixed model inference on the two simulated nuclear family datasets under family-based outcome-dependent using the proposed weighted MLE approach and MLE approach. The model inference under random sampling serves as a baseline of sample estimation. The vertical dotted lines represent the true parameters of the simulated data.

An advantage of the family-based design is that, for the simulated nuclear family data where the covariance matrix Ξ is a block diagonal matrix, $\mathbb{E}[y^*|y]_O = \mathbb{E}[y^*_O|y_O]$ and $\mathbb{V}(y^*|y)_{OO} = \mathbb{V}(y^*_O|y_O)$, hence the log-likelihood in Eq.6.11 is the same as the log-likelihood in Eq.6.10. On the other hand, this is not true for individual-based sampling design, which explains the bias of the sample weighted MLE in the bottom row of Figure 6.5 beyond data size.

To see whether varying model parameters affects the conclusions on the proposed weighted approach for generalized linear mixed model, I simulated four nuclear family datasets with different parameter values ($\beta_1 = 0.5$ or $\beta_1 = 1$, $\sigma = 0.2$ or $\sigma = 0.5$) and the results are shown in Figure 6.8. Dataset A and B investigate the effect of varying β_1 only,

Dataset A and C investigate the effect of varying σ only, and Dataset A and D investigate the effect of varying both β_1 and σ. Figure 6.8 confirms that varying model parameters has no effect on the sampling bias correction of the weighted approach, whereas the unweighted estimation of all parameters are always biased, and the bias in σ seems to increase as β_1 and σ increase without weights adjustment.

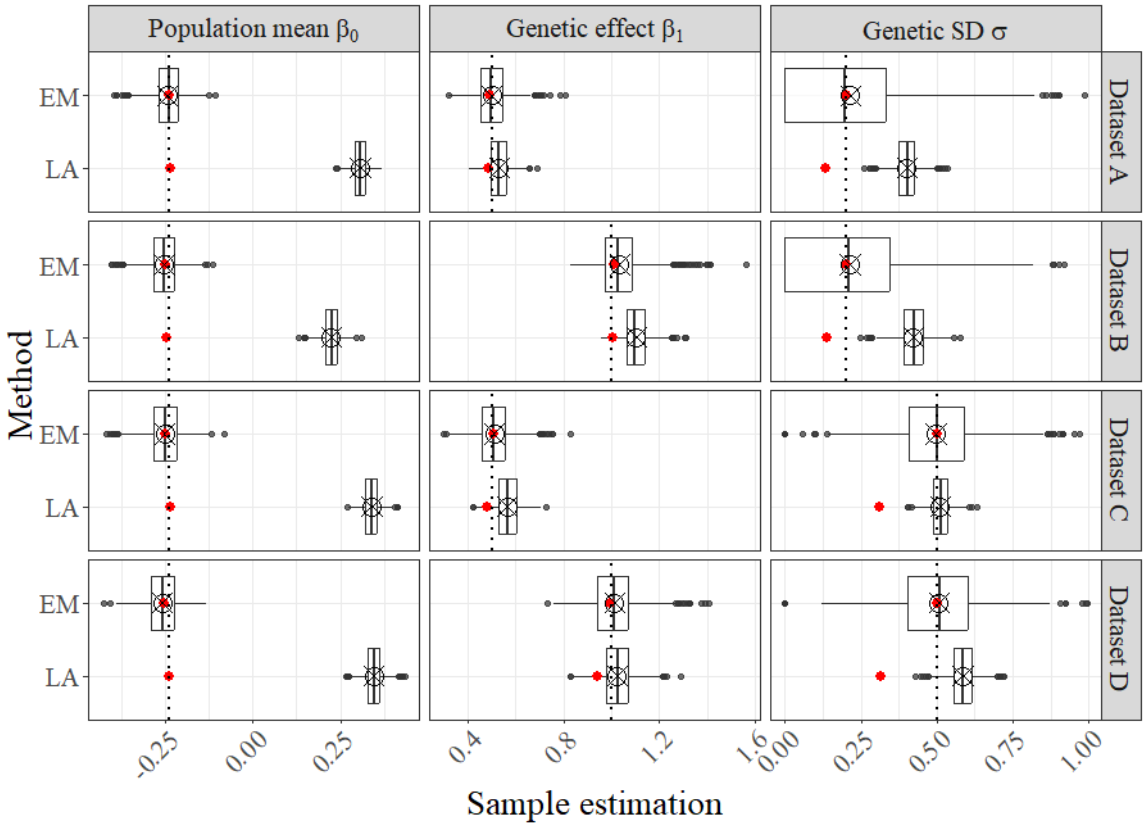

Fig. 6.8 The effect of varying model parameters (β_1,σ). The four simulated nuclear family datasets A, B, C and D ($N_A = N_B = N_C = N_D = 10011$) are generated with $\beta_1 = 0.5$ or $\beta_1 = 1$, $\sigma = 0.2$ or $\sigma = 0.5$. The samples from datasets A, B, C and D are selected under family-based outcome-dependent sampling ($n_A \approx 2200$, $n_B \approx 2256$, $n_C \approx 2208$, $n_D \approx 2237$). The vertical dotted lines represent the true parameters of the simulated data and the red dots represent the population estimates of the simulated data.

6.4 Summary

Recall the weighted log-likelihood in Eq 5.2. In general, the log-likelihood of linear mixed models cannot be written as a pairwise sum. Hence, under informative sampling, the sampling weights appear in the weighted estimating equation in a non-linear form and the weighted log-likelihood is not design-unbiased. One approach to solve this is to consider

a composite log-likelihood, which is a sum of individual components of the log-likelihood. While the composite likelihood is an estimation of the population likelihood, Chapter 5 considered a special case of linear mixed models where the non-log-determinant part of the population log-likelihood can be written as a sum over all pairs of individuals with pairwise sampling indicator and weights. It was expected that this approach would be more efficient than an estimation of the population likelihood.

This idea cannot be extended to most generalized linear mixed models except the probit model where the probability (i.e., liability of a binary trait) follows a normal distribution like the linear mixed model in Chapter 5. Unlike quantitative traits, the liability of a binary trait is unobserved. Therefore, this chapter has been focusing on constructing a weighted estimation of the log-likelihood of a generalized linear mixed model in a linear form, and obtaining weighted estimation of model parameters using the Monte Carlo EM algorithm.

In contrast to the weighted log-likelihood of linear mixed models in Eq 5.2, the weighted log-likelihood of generalized linear mixed models requires calculations of expectation and variance of the latent variable conditional on the complete data either by Gibbs sampling or numerical integration. This is achievable with sample data only when the $(\Xi^{-1})_{OO} = (\Xi_{OO})^{-1}$, i.e., the covariance matrix Ξ is a block diagonal matrix and the design samples the whole blocks rather than individuals. When this is not the case (e.g. under individual-based outcome-dependent sampling), the log-likelihood in Eq.6.11 tends to underestimate the variance component, but the bias tends to reduce as the data size increases. In conclusion, the weighted estimation can be extended from the linear mixed models to the generalized linear mixed models only under family-based sampling designs. Apart from the second case study in section 6.3.1, another possible two-phase design is the proband case-control design for the human population.

Chapter 7

Future work

In this thesis, I explored two kinds of approaches for handling incomplete data in two-phase sampling designs: obtaining complete data through genotype imputation and model inference using incomplete data. This chapter summarizes the work that has been done in this thesis with respect to the two approaches, including the contribution to the field, connection with the current literature, limitations and potential extensions and applications.

Chapter 3 investigated a few factors that are likely to influence the performance of genotype imputation for the endangered and inbred kākāpō species, such as the type of low-density genotype data, reference subject selection, and whether or not relatedness is taken into account. It was found that the type of low-density genotype data is the major factor that affects both imputation accuracy and the number of imputed genotypes. In comparison to reference SNPs, SNPs called from low-depth GBS data had a higher error rate and a larger proportion of missing genotypes, and therefore lead to poor performance in genotype imputation when it is served as the low-density genotype data.

Bilton [13] developed an approach that is particularly suitable for low-depth GBS data, that extends the existing models in genetic analyses by incorporating a binomial-type sampling model for the conditional probability of read count for reference/alternative allele given latent genotype. In contrast to using called genotypes from low-depth GBS data in genetic analyses, the advantage of Bilton's approach is utilizing all available information and incorporating the uncertainty of the low-depth GBS data directly into the model for genetic analysis. Bilton [13] extended the models to GBS data in genetic linkage maps construction and estimation of genetic relatedness estimation. This can also be done for genotype imputation with low-depth GBS data by developing a new model for genotype imputation that incorporates the GBS error process using the read count information.

In Chapter 5, I proposed a weighted maximum likelihood approach for fitting linear mixed models by taking advantage of the fact the kākāpō population relatedness structure is known,

making it possible to incorporate the population covariance matrix rather than the sample covariance matrix into the model. Since the population relatedness structure is often known either exactly or approximately for endangered species, the proposed approach provides a general solution for fitting linear mixed models under two-phase sampling designs in complex pedigrees in conservation genetics. In other words, this allows obtaining population parameter estimations in linear mixed models using only sample data but without sampling bias, which can greatly reduce the genotyping/phenotyping cost in conservation studies.

The performance of the proposed weighted maximum likelihood method was also evaluated using a simulated dataset containing typical human pedigrees and with a ten times larger population size compared to the kākāpō data. The two case studies demonstrated that the sampling bias could be corrected by re-weighting the samples regardless of relatedness structure. However, in contrast to the larger simulated dataset, the weighted sample estimation is quite variable for the kākāpō data given there are only 104 individuals with known egg lengths. It is expected that the proposed method will work better if all kākāpō are phenotyped or for less endangered species.

In Chapter 6, I extended the idea of re-weighting the observations with unequal sampling probabilities to analyze binary traits by assuming they have a continuous normally-distributed liability that measures the susceptibility to the traits. More specifically, the weighted maximum likelihood approach for fitting generalized linear mixed models under complex sampling design can be carried out using a Monte-Carlo EM algorithm.

However, the proposed method has some limitations when fitting a generalized linear mixed model. Unlike the linear mixed models, the sample weighted log-likelihood for generalized linear mixed models requires calculations of expectation and variance of the latent variable conditional on the complete data either by Gibbs sampling or numerical integration. As discussed in Chapter 6, this is only possible when the population covariance matrix is a block diagonal matrix and the design samples blocks rather than individuals. Otherwise, the variance component will be underestimated, particularly when the sample size is small.

The proposed weighted maximum likelihood approach for fitting linear mixed models under two-phase designs is available as an R package called `WLMM` on GitHub (https://github.com/zoeluo15/WLMM). For generalized linear mixed models, the R code for weighted maximum likelihood estimation via the Monte Carlo EM algorithm is also available in the same GitHub repository.

Other possible approaches for fitting mixed models under two-phased designs include the full likelihood approach which is infeasible to implement in general as demonstrated in Chapter 5 and pseudolikelihood approaches. Two relatively closely related methods are the

sample weighted pseudolikelihood proposed by Rabe-Hesketh and Skrondal [112] and the pairwise likelihood proposed by Rao et al. [114] and Yi et al. [165] (see Section 4.3.3 for more details). To the best of my knowledge, the sample weighted pseudolikelihood proposed by Rabe-Hesketh and Skrondal [112] is the only method for mixed model inference under complex sampling design with available software (see `gllamm` in `Stata` [138]). However, Rabe-Hesketh and Skrondal [112] assumes the model cluster is the same as the sampling unit, which is impossible for kākāpō due to a high level of inbreeding, and therefore their pseudolikelihood approach cannot be applied to the kākāpō data. Huang [68] showed that this assumption could be relaxed for the pairwise likelihood proposed by Rao et al. [114] and Yi et al. [165], but it is expected that there will be an efficiency loss as a result of considering only pairs.

It is then natural to ask whether the proposed approach that utilizes the sample covariance matrix will lead to a gain of efficiency compared to the pairwise likelihood. Unfortunately, a similar loss of efficiency as pairwise likelihood was found for the proposed sample weighted log-likelihood for reasons that are not fully understood. Therefore, another direction of future work is to improve the efficiency by calibration that takes advantage of knowing the genotype information for some individuals. Breslow et al. [20, 21] use calibrated weights to improve the asymptotic efficiency for the target parameter and describe a general strategy for constructing an auxiliary variable that is linearly correlated with the estimator of the target parameter. In brief, their strategy involves: (1) developing an imputation model for the partially observed phase II variable from the fully observed phase I variables; (2) using this model used to predict the values of the phase II variable for all phase I individuals; (3) estimating the influence functions from the outcome model using the complete data; (4) using the influence functions as auxiliary variables in calibration.

For unrelated individuals, it is relatively straightforward to predict the missing genotypes from observed genotypes. On the other hand, building an imputation model for related individuals can be much more complicated because the imputed genotypes need to be consistent with the inheritance pattern. Nevertheless, calibration does not require an error model to achieve high imputation accuracy, and it is helpful as long as the auxiliary variables are correlated with the variable of interest.

As a summary, this thesis explored the two classes of approaches to handling incomplete data in two-phasing sampling designs under different situations. For both class of approaches, there are more works can be done in the future. To achieve high accuracy in genotype imputation using low-depth GBS data, a new model for genotype imputation that incorporates the GBS error process needed to be developed. For the maximum likelihood estimator, there is a loss in efficiency using either the weighted likelihood or the pairwise pseudolikelihood,

hence a new likelihood estimator that combines the information used by both methods is needed.

References

[1] Abecasis, G. R., Cherny, S. S., Cookson, W. O., and Cardon, L. R. (2001). Merlin—rapid analysis of dense genetic maps using sparse gene flow trees. *Nature Genetics*, 30(1):97–101.

[2] Akdemir, D. and Isidro-Sánchez, J. (2019). Design of training populations for selective phenotyping in genomic prediction. *Scientific Reports*, 9(1):1–15.

[3] Akdemir, D., Sanchez, J. I., and Jannink, J.-L. (2015). Optimization of genomic selection training populations with a genetic algorithm. *Genetics Selection Evolution*, 47(1):1–10.

[4] Almasy, L., Dyer, T. D., Peralta, J. M., Jun, G., Wood, A. R., Fuchsberger, C., Almeida, M. A., Kent, J. W., Fowler, S., Blackwell, T. W., et al. (2014). Data for Genetic Analysis Workshop 18: human whole genome sequence, blood pressure, and simulated phenotypes in extended pedigrees. In *BMC proceedings*, volume 8, page S2. BioMed Central.

[5] Aluru, S. (2005). *Handbook of Computational Molecular Biology*. Chapman and Hall/CRC.

[6] Antwis, R. E., Edwards, K. L., Unwin, B., Walker, S. L., and Shultz, S. (2019). Rare gut microbiota associated with breeding success, hormone metabolites and ovarian cycle phase in the critically endangered eastern black rhino. *Microbiome*, 7(1):1–12.

[7] Atwell, S. et al. (2010). Genome-wide association study of 107 phenotypes in Arabidopsis thaliana inbred lines. *Nature*, 465(7298):627–631.

[8] Aulchenko, Y. S., De Koning, D.-J., and Haley, C. (2007). Genomewide rapid association using mixed model and regression: a fast and simple method for genomewide pedigree-based quantitative trait loci association analysis. *Genetics*, 177(1):577–585.

[9] Axelsson, E., Webster, M. T., Smith, N. G., Burt, D. W., and Ellegren, H. (2005). Comparison of the chicken and turkey genomes reveals a higher rate of nucleotide divergence on microchromosomes than macrochromosomes. *Genome Research*, 15(1):120–125.

[10] Backström, N., Forstmeier, W., Schielzeth, H., Mellenius, H., Nam, K., Bolund, E., Webster, M. T., Öst, T., Schneider, M., Kempenaers, B., et al. (2010). The recombination landscape of the zebra finch *Taeniopygia guttata* genome. *Genome Research*, 20(4):485–495.

[11] Bates, D., Mächler, M., Bolker, B., and Walker, S. (2014). Fitting linear mixed-effects models using lme4. *ArXiv:1406.5823*. doi:https://doi.org/10.48550/arXiv.1406.5823.

[12] Benscoter, A. M., Reece, J. S., Noss, R. F., Brandt, L. A., Mazzotti, F. J., Romañach, S. S., and Watling, J. I. (2013). Threatened and endangered subspecies with vulnerable ecological traits also have high susceptibility to sea level rise and habitat fragmentation. *PLOS One*, 8(8):e70647.

[13] Bilton, T. P. (2020). *Developing statistical methods for genetic analysis of genotypes from genotyping-by-sequencing data*. PhD thesis, University of Otago.

[14] Borecki, I. B. and Province, M. A. (2008). Linkage and association: basic concepts. *Advances in Genetics*, 60:51–74.

[15] Breslow, N. and Cain, K. (1988). Logistic regression for two-stage case-control data. *Biometrika*, 75(1):11–20.

[16] Breslow, N. E. (1996). Statistics in epidemiology: The case-control study. *Journal of the American Statistical Association*, 91(433):14–28.

[17] Breslow, N. E. and Clayton, D. G. (1993). Approximate inference in generalized linear mixed models. *Journal of the American Statistical Association*, 88(421):9–25.

[18] Breslow, N. E. and Holubkov, R. (1997). Maximum likelihood estimation of logistic regression parameters under two-phase, outcome-dependent sampling. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 59(2):447–461.

[19] Breslow, N. E. and Lin, X. (1995). Bias correction in generalised linear mixed models with a single component of dispersion. *Biometrika*, 82(1):81–91.

[20] Breslow, N. E., Lumley, T., Ballantyne, C. M., Chambless, L. E., and Kulich, M. (2009a). Improved Horvitz–Thompson estimation of model parameters from two-phase stratified samples: applications in epidemiology. *Statistics in Biosciences*, 1(1):32–49.

[21] Breslow, N. E., Lumley, T., Ballantyne, C. M., Chambless, L. E., and Kulich, M. (2009b). Using the whole cohort in the analysis of case-cohort data. *American Journal of Epidemiology*, 169(11):1398–1405.

[22] Browning, B. L. and Browning, S. R. (2009). A unified approach to genotype imputation and haplotype-phase inference for large data sets of trios and unrelated individuals. *The American Journal of Human Genetics*, 84(2):210–223.

[23] Browning, B. L. and Browning, S. R. (2016). Genotype imputation with millions of reference samples. *The American Journal of Human Genetics*, 98(1):116–126.

[24] Browning, S. R. and Browning, B. L. (2011). Haplotype phasing: existing methods and new developments. *Nature Reviews Genetics*, 12(10):703–714.

[25] Buckler, E. S., Yu, J., Pressoir, G., Briggs, W. H., Vroh Bi, I., Yamasaki, M., Doebley, J. F., McMullen, M. D., Gaut, B. S., Nielsen, D. M., Holland, J. B., and Kresovich, S. (2006). A unified mixed-model method for association mapping that accounts for multiple levels of relatedness. *Nature Genetics*, 38(2):203–208.

[26] Buniello, A., MacArthur, J. A. L., Cerezo, M., Harris, L. W., Hayhurst, J., Malangone, C., McMahon, A., Morales, J., Mountjoy, E., Sollis, E., et al. (2019). The NHGRI-EBI GWAS Catalog of published genome-wide association studies, targeted arrays and summary statistics 2019. *Nucleic Acids Research*, 47(D1):D1005–D1012.

[27] Bürkner, P.-C., Gabry, J., and Vehtari, A. (2021). Efficient leave-one-out cross-validation for Bayesian non-factorized normal and Student-t models. *Computational Statistics*, 36(2):1243–1261.

[28] Burt, D. (2002). Origin and evolution of avian microchromosomes. *Cytogenetic and Genome Research*, 96(1-4):97–112.

[29] Carmi, S., Palamara, P. F., Vacic, V., Lencz, T., Darvasi, A., and Pe'er, I. (2013). The variance of identity-by-descent sharing in the Wright–Fisher model. *Genetics*, 193(3):911–928.

[30] Carroll, R. J. and Wand, M. P. (1991). Semiparametric estimation in logistic measurement error models. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 53(3):573–585.

[31] Chan, J. S. and Kuk, A. Y. (1997). Maximum likelihood estimation for probit-linear mixed models with correlated random effects. *Biometrics*, 53(1):86–97.

[32] Chatterjee, N., Chen, Y.-H., and Breslow, N. E. (2003). A pseudoscore estimator for regression problems with two-phase sampling. *Journal of the American Statistical Association*, 98(461):158–168.

[33] Cheung, C. Y., Blue, E. M., and Wijsman, E. M. (2014). A statistical framework to guide sequencing choices in pedigrees. *The American Journal of Human Genetics*, 94(2):257–267.

[34] Cheung, C. Y., Thompson, E. A., and Wijsman, E. M. (2013). GIGI: an approach to effective imputation of dense genotypes on large pedigrees. *The American Journal of Human Genetics*, 92(4):504–516.

[35] Choi, S.-H., Liu, C., Dupuis, J., Logue, M. W., and Jun, G. (2011). Using linkage analysis of large pedigrees to guide association analyses. In *BMC proceedings*, volume 5, pages 1–4. Springer.

[36] Christian Fuchsberger (2010). ExomePicks - Genome Analysis Wiki. https://genome.sph.umich.edu/wiki/ExomePicks. [Online; accessed 14-August-2019].

[37] Cordell, H. J. and Clayton, D. G. (2002). A unified stepwise regression procedure for evaluating the relative effects of polymorphisms within a gene using case/control or family data: Application to hla in type 1 diabetes. *American Journal of Human Genetics*, 70(1):124–141.

[38] Danecek, P., Auton, A., Abecasis, G., Albers, C. A., Banks, E., DePristo, M. A., Handsaker, R. E., Lunter, G., Marth, G. T., Sherry, S. T., McVean, G., and Durbin, R. (2011). The variant call format and vcftools. *Bioinformatics*, 27(15):2156–2158.

[39] David, O., Le Rouzic, A., and Dillmann, C. (2022). Optimization of sampling designs for pedigrees and association studies. *Biometrics*, 78(3):1056–1066.

[40] de Valpine, P., Turek, D., Paciorek, C. J., Anderson-Bergman, C., Lang, D. T., and Bodik, R. (2017). Programming with models: writing statistical algorithms for general model structures with NIMBLE. *Journal of Computational and Graphical Statistics*, 26(2):403–413.

[41] Dempster, A. P., Laird, N. M., and Rubin, D. B. (1977). Maximum likelihood from incomplete data via the em algorithm. *Journal of the Royal Statistical Society. Series B, Methodological*, 39(1):1–38.

[42] Deville, J.-C. and Särndal, C.-E. (1992). Calibration estimators in survey sampling. *Journal of the American Statistical Association*, 87(418):376–382.

[43] Dodds, K. G., McEwan, J. C., Brauning, R., Anderson, R. M., van Stijn, T. C., Kristjánsson, T., and Clarke, S. M. (2015). Construction of relatedness matrices using genotyping-by-sequencing data. *BMC Genomics*, 16(1):1047–1047.

[44] Dussex, N., Van Der Valk, T., Morales, H. E., Wheat, C. W., Díez-del Molino, D., Von Seth, J., Foster, Y., Kutschera, V. E., Guschanski, K., Rhie, A., et al. (2021). Population genomics of the critically endangered kākāpō. *Cell Genomics*, 1(1):100002.

[45] Edge, P., Bafna, V., and Bansal, V. (2017). HapCUT2: robust and accurate haplotype assembly for diverse sequencing technologies. *Genome Research*, 27(5):801–812.

[46] Ellegren, H. (2013). The evolutionary genomics of birds. *Annual Review of Ecology, Evolution, and Systematics*, 44(1):239–259.

[47] Elshire, R. J., Glaubitz, J. C., Sun, Q., Poland, J. A., Kawamoto, K., Buckler, E. S., and Mitchell, S. E. (2011). A robust, simple genotyping-by-sequencing (gbs) approach for high diversity species. *PLOS One*, 6(5):1–10.

[48] Elston, R. C. and Stewart, J. (1971). A general model for the genetic analysis of pedigree data. *Human Heredity*, 21(6):523–542.

[49] Espinoza, T., Burke, C. L., Carpenter-Bundhoo, L., Marshall, S., Roberts, D., and Kennard, M. J. (2020). Fine-scale acoustic telemetry in a riverine environment: movement and habitat use of the endangered Mary River cod *Maccullochella mariensis*. *Endangered Species Research*, 42:125–131.

[50] Falconer, D. S. (1965). The inheritance of liability to certain diseases, estimated from the incidence among relatives. *Annals of Human Genetics*, 29(1):51–76.

[51] Fisher, R. A. (1919). XV.—The correlation between relatives on the supposition of Mendelian inheritance. *Earth and Environmental Science Transactions of the Royal Society of Edinburgh*, 52(2):399–433.

[52] Flanders, W. D. and Greenland, S. (1991). Analytic methods for two-stage case-control studies and other stratified designs. *Statistics in Medicine*, 10(5):739–747.

[53] Foster, Y., Dutoit, L., Grosser, S., Dussex, N., Foster, B. J., Dodds, K. G., Brauning, R., Van Stijn, T., Robertson, F., McEwan, J. C., et al. (2021). Genomic signatures of inbreeding in a critically endangered parrot, the kākāpō. *G3*, 11(11):jkab307.

[54] Fragoso, C. A., Heffelfinger, C., Zhao, H., and Dellaporta, S. L. (2016). Imputing genotypes in biallelic populations from low-coverage sequence data. *Genetics (Austin)*, 202(2):487–495.

[55] Fuchsberger, C., Flannick, J., Teslovich, T. M., et al. (2016). The genetic architecture of type 2 diabetes. *Nature*, 536(7614):41–47.

[56] Gudbjartsson, D. F., Helgason, H., Gudjonsson, S. A., Zink, F., Oddson, A., Gylfason, A., Besenbacher, S., Magnusson, G., Halldorsson, B. V., Hjartarson, E., et al. (2015). Large-scale whole-genome sequencing of the Icelandic population. *Nature Genetics*, 47(5):435–444.

[57] Guhlin, J. (2020). Private Communication.

[58] Guhlin, J., Lec, M. F. L., Wold, J., Koot, E., Winter, D., Biggs, P., Galla, S. J., Urban, L., Foster, Y., Cox, M. P., Digby, A., Uddstrom, L., Eason, D., Vercoe, D., Davis, T., Howard, J. T., Jarvis, E., Robertson, F. E., Robertson, B. C., Gemmell, N., Steeves, T. E., Santure, A. W., and Dearden, P. K. (2022). Species-wide genomics of kākāpo provides transformational tools to accelerate recovery. *bioRxiv*.

[59] Guilloud-Bataille, M., Bouzigon, E., Annesi-Maesano, I., Bousquet, J., Charpin, D., Gormand, F., Hochez, J., Just, J., Lemainque, A., Le Moual, N., et al. (2008). Evidence for linkage of a new region (11p14) to eczema and allergic diseases. *Human Genetics*, 122(6):605–614.

[60] Haataja, R., Karjalainen, M. K., Luukkonen, A., Teramo, K., Puttonen, H., Ojaniemi, M., Varilo, T., Chaudhari, B. P., Plunkett, J., Murray, J. C., et al. (2011). Mapping a new spontaneous preterm birth susceptibility gene, IGF1R, using linkage, haplotype sharing, and association analysis. *PLOS Genetics*, 7(2):e1001293.

[61] Han, S., Yang, B.-Z., Kranzler, H. R., Oslin, D., Anton, R., Farrer, L. A., and Gelernter, J. (2012). Linkage analysis followed by association show NRG1 associated with cannabis dependence in African Americans. *Biological Psychiatry*, 72(8):637–644.

[62] Heckerman, D., Gurdasani, D., Kadie, C., Pomilla, C., Carstensen, T., Martin, H., Ekoru, K., Nsubuga, R. N., Ssenyomo, G., Kamali, A., et al. (2016). Linear mixed model for heritability estimation that explicitly addresses environmental variation. *Proceedings of the National Academy of Sciences*, 113(27):7377–7382.

[63] Hickey, J. M., Crossa, J., Babu, R., and de los Campos, G. (2012). Factors affecting the accuracy of genotype imputation in populations from several maize breeding programs. *Crop Science*, 52(2):654–663.

[64] Hing, S., Jones, K. L., Rafferty, C., Thompson, R. A., Narayan, E. J., and Godfrey, S. S. (2017). Wildlife in the line of fire: evaluating the stress physiology of a critically endangered Australian marsupial after bushfire. *Australian Journal of Zoology*, 64(6):385–389.

[65] Horvitz, D. G. and Thompson, D. J. (1952). A generalization of sampling without replacement from a finite universe. *Journal of the American Statistical Association*, 47(260):663–685.

[66] Howie, B. N., Donnelly, P., and Marchini, J. (2009). A flexible and accurate genotype imputation method for the next generation of genome-wide association studies. *PLOS Genetics*, 5(6):e1000529–e1000529.

[67] Huang, L., Li, Y., Singleton, A. B., Hardy, J. A., Abecasis, G., Rosenberg, N. A., and Scheet, P. (2009). Genotype-imputation accuracy across worldwide human populations. *The American Journal of Human Genetics*, 84(2):235–250.

[68] Huang, X. (2019). *Mixed Models for Complex Survey Data*. PhD thesis, The University of Auckland.

[69] Hutz, J. E., Manning, W. A., Province, M. A., and McLeod, H. L. (2011). Genomewide analysis of inherited variation associated with phosphorylation of PI3K/AKT/mTOR signaling proteins. *PLOS One*, 6(9):e24873.

[70] International HapMap 3 Consortium and others (2010). Integrating common and rare genetic variation in diverse human populations. *Nature*, 467(7311):52–58.

[71] International HapMap Consortium and others (2003). The international HapMap project. *Nature*, 426(6968):789–796.

[72] International HapMap Consortium and others (2007). A second generation human haplotype map of over 3.1 million SNPs. *Nature*, 449(7164):851–861.

[73] Kalbfleisch, J. and Lawless, J. (1988). Likelihood analysis of multi-state models for disease incidence and mortality. *Statistics in Medicine*, 7(1-2):149–160.

[74] Kang, H. M., Sul, J. H., Service, S. K., Freimer, N. B., Zaitlen, N. A., Eskin, E., Sabatti, C., and Kong, S.-y. (2010). Variance component model to account for sample structure in genome-wide association studies. *Nature Genetics*, 42(4):348–354.

[75] Kang, H. M., Zaitlen, N. A., Wade, C. M., Kirby, A., Heckerman, D., Daly, M. J., and Eskin, E. (2008). Efficient control of population structure in model organism association mapping. *Genetics*, 178(3):1709–1723.

[76] Kim, W., Kwak, S. H., and Won, S. (2019). Heritability estimation of dichotomous phenotypes using a liability threshold model on ascertained family-based samples. *Genetic Epidemiology*, 43(7):761–775.

[77] Klein, R. J., Zeiss, C., Chew, E. Y., Tsai, J.-Y., Sackler, R. S., Haynes, C., Henning, A. K., SanGiovanni, J. P., Mane, S. M., Mayne, S. T., et al. (2005). Complement factor H polymorphism in age-related macular degeneration. *Science*, 308(5720):385–389.

[78] Knapp, M., Clarke, A. C., Horsburgh, K. A., and Matisoo-Smith, E. A. (2012). Setting the stage – Building and working in an ancient DNA laboratory. *Annals of Anatomy*, 194(1):3–6.

[79] Koffler, S., de Matos Peixoto Kleinert, A., and Jaffé, R. (2017). Quantitative conservation genetics of wild and managed bees. *Conservation Genetics*, 18(3):689–700.

[80] Kruglyak, L., Daly, M. J., Reeve-Daly, M. P., and Lander, E. S. (1996). Parametric and nonparametric linkage analysis : A unified multipoint approach. *American Journal of Human Genetics*, 58(6):1347–1363.

[81] Lander, E. S. and Green, P. (1987). Construction of multilocus genetic linkage maps in humans. *Proceedings of the National Academy of Sciences*, 84(8):2363–2367.

[82] Lee, J. H., Cheng, R., Honig, L. S., Feitosa, M., Kammerer, C. M., Kang, M. S., Schupf, N., Lin, S. J., Sanders, J. L., Bae, H., et al. (2014). Genome wide association and linkage analyses identified three loci—4q25, 17q23. 2, and 10q11. 21—associated with variation in leukocyte telomere length: the long life family study. *Frontiers in Genetics*, 4:310.

[83] Lee, S. H., Wray, N. R., Goddard, M. E., and Visscher, P. M. (2011). Estimating missing heritability for disease from genome-wide association studies. *The American Journal of Human Genetics*, 88(3):294–305.

[84] Li, B. and Leal, S. M. (2008). Methods for detecting associations with rare variants for common diseases: application to analysis of sequence data. *The American Journal of Human Genetics*, 83(3):311–321.

[85] Li, N. and Stephens, M. (2003). Modeling linkage disequilibrium and identifying recombination hotspots using single-nucleotide polymorphism data. *Genetics*, 165(4):2213–2233.

[86] Li, Y., Willer, C. J., Ding, J., Scheet, P., and Abecasis, G. R. (2010). MaCH: using sequence and genotype data to estimate haplotypes and unobserved genotypes. *Genetic Epidemiology*, 34(8):816–834.

[87] Lin, S. and Zhao, H. (2010). *Handbook on Analyzing Human Genetic Data Computational Approaches and Software*. Springer.

[88] Little, R. J. (1995). Modeling the drop-out mechanism in repeated-measures studies. *Journal of the American Statistical Association*, 90(431):1112–1121.

[89] Loh, P.-R., Tucker, G., Bulik-Sullivan, B. K., Vilhjálmsson, B. J., Finucane, H. K., Salem, R. M., Chasman, D. I., Ridker, P. M., Neale, B. M., Berger, B., Patterson, N., and Price, A. L. (2015). Efficient Bayesian mixed-model analysis increases association power in large cohorts. *Nature Genetics*, 47(3):284–290.

[90] Lumley, T. and Scott, A. (2017). Fitting regression models to survey data. *Statistical Science*, 32(2):265–278.

[91] Lumley, T., Shaw, P. A., and Dai, J. Y. (2011). Connections between survey calibration estimators and semiparametric models for incomplete data. *International Statistical Review*, 79(2):200–220.

[92] MacDonald, M. E., Novelletto, A., Lin, C., Tagle, D., Barnes, G., Bates, G., Taylor, S., Allitto, B., Altherr, M., Myers, R., et al. (1992). The Huntington's disease candidate region exhibits many different haplotypes. *Nature Genetics*, 1(2):99–103.

[93] Magee, L. (1998). Improving survey-weighted least squares regression. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 60(1):115–126.

[94] Mahon, P. B., Payne, J. L., MacKinnon, D. F., Mondimore, F. M., Goes, F. S., Schweizer, B., Jancic, D., Consortium, N. G. I. B. D., Consortium, B., Coryell, W. H., et al. (2009). Genome-wide linkage and follow-up association study of postpartum mood symptoms. *American Journal of Psychiatry*, 166(11):1229–1237.

[95] Manolio, T. A., Brooks, L. D., and Collins, F. S. (2008). A HapMap harvest of insights into the genetics of common disease. *The Journal of Clinical Investigation*, 118(5):1590–1605.

[96] Marchini, J. and Howie, B. (2010). Genotype imputation for genome-wide association studies. *Nature Reviews Genetics*, 11(7):499–511.

[97] Martin, M., Patterson, M., Garg, S., Fischer, S. O., Pisanti, N., Klau, G. W., Schöenhuth, A., and Marschall, T. (2016). WhatsHap: fast and accurate read-based phasing. *bioRxiv:085050*. doi:https://doi.org/10.1101/085050.

[98] McCulloch, C. E. (1994). Maximum likelihood variance components estimation for binary data. *Journal of the American Statistical Association*, 89(425):330–335.

[99] Minster, R. L., Hawley, N. L., Su, C.-T., Sun, G., Kershaw, E. E., Cheng, H., Buhule, O. D., Lin, J., Tuitele, J., Naseri, T., et al. (2016). A thrifty variant in CREBRF strongly influences body mass index in Samoans. *Nature Genetics*, 48(9):1049–1054.

[100] Neuhaus, J. M., Scott, A. J., and Wild, C. J. (2002). The analysis of retrospective family studies. *Biometrika*, 89(1):23–37.

[101] Neuhaus, J. M., Scott, A. J., and Wild, C. J. (2006). Family-specific approaches to the analysis of case–control family data. *Biometrics*, 62(2):488–494.

[102] Newey, W. K. (1991). Uniform convergence in probability and stochastic equicontinuity. *Econometrica: Journal of the Econometric Society*, 59(4):1161–1167.

[103] Neyman, J. (1938). Contribution to the theory of sampling human populations. *Journal of the American Statistical Association*, 33(201):101–116.

[104] Ott, J., Kamatani, Y., and Lathrop, M. (2011). Family-based designs for genome-wide association studies. *Nature Reviews Genetics*, 12(7):465–474.

[105] Ozaki, K., Ohnishi, Y., Iida, A., Sekine, A., Yamada, R., Tsunoda, T., Sato, H., Sato, H., Hori, M., Nakamura, Y., et al. (2002). Functional SNPs in the lymphotoxin-α gene that are associated with susceptibility to myocardial infarction. *Nature Genetics*, 32(4):650–654.

[106] Pepe, M. S. and Fleming, T. R. (1991). A nonparametric method for dealing with mismeasured covariate data. *Journal of the American Statistical Association*, 86(413):108–113.

[107] Pfeffermann, D., Skinner, C. J., Holmes, D. J., Goldstein, H., and Rasbash, J. (1998). Weighting for unequal selection probabilities in multilevel models. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 60(1):23–40.

[108] Pfeffermann, D. and Sverchkov, M. (1999). Parametric and semi-parametric estimation of regression models fitted to survey data. *Sankhyā: The Indian Journal of Statistics, Series B*, 61(1):166–186.

[109] Powell, M. J. (2009). The BOBYQA algorithm for bound constrained optimization without derivatives. *Cambridge NA Report NA2009/06, University of Cambridge, Cambridge*, 26.

[110] Prentice, R. L. and Pyke, R. (1979). Logistic disease incidence models and case-control studies. *Biometrika*, 66(3):403–411.

[111] R Core Team (2022). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria.

[112] Rabe-Hesketh, S. and Skrondal, A. (2006). Multilevel modelling of complex survey data. *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, 169(4):805–827.

[113] Ramnarine, S., Zhang, J., Chen, L.-S., Culverhouse, R., Duan, W., Hancock, D. B., Hartz, S. M., Johnson, E. O., Olfson, E., Schwantes-An, T.-H., et al. (2015). When does choice of accuracy measure alter imputation accuracy assessments? *PLOS One*, 10(10):e0137601.

[114] Rao, J., Verret, F., and Hidiroglou, M. A. (2013). A weighted composite likelihood approach to inference for two-level models from survey data. *Survey Methodology*, 39(2):263–282.

[115] Reilly, M. and Pepe, M. S. (1995). A mean score method for missing and auxiliary covariate data in regression models. *Biometrika*, 82(2):299–314.

[116] Rhie, A., McCarthy, S. A., Fedrigo, O., Damas, J., Formenti, G., Koren, S., Uliano-Silva, M., Chow, W., Fungtammasan, A., Kim, J., et al. (2021). Towards complete and error-free genome assemblies of all vertebrate species. *Nature*, 592(7856):737–746.

[117] Rivera, C. and Lumley, T. (2016). Using the whole cohort in the analysis of counter-matched samples. *Biometrics*, 72(2):382–391.

[118] Rivera-Rodriguez, C., Spiegelman, D., and Haneuse, S. (2019). On the analysis of two-phase designs in cluster-correlated data settings. *Statistics in Medicine*, 38(23):4611–4624.

[119] Robins, J. M., Hernan, M. A., and Brumback, B. (2000). Marginal structural models and causal inference in epidemiology. *Epidemiology*, 11(5):550–560.

[120] Robins, J. M., Rotnitzky, A., and Zhao, L. P. (1994). Estimation of regression coefficients when some regressors are not always observed. *Journal of the American Statistical Association*, 89(427):846–866.

[121] Rubin, D. B. (1976). Inference and missing data. *Biometrika*, 63(3):581–592.

[122] Saad, M. and Wijsman, E. M. (2014). Power of family-based association designs to detect rare variants in large pedigrees using imputed genotypes. *Genetic Epidemiology*, 38(1):1–9.

[123] Sambrook, J., Fritsch, E., and Maniatis, T. (1989). *Molecular Cloning: A Laboratory Manual, Second Edition*. Cold Spring Harbor Laboratory.

[124] Scheet, P. and Stephens, M. (2006). A fast and flexible statistical model for large-scale population genotype data: applications to inferring missing genotypes and haplotypic phase. *The American Journal of Human Genetics*, 78(4):629–644.

[125] Schildcrout, J. S., Garbett, S. P., and Heagerty, P. J. (2013). Outcome vector dependent sampling with longitudinal continuous response data: stratified sampling based on summary statistics. *Biometrics*, 69(2):405–416.

[126] Schill, W., Jöckel, K., Drescher, K., and Timm, J. (1993). Logistic analysis in case-control studies under validation sampling. *Biometrika*, 80(2):339–352.

[127] Schmidt, K. J., Soluk, D. A., Maestas, S. E. M., and Britten, H. B. (2021). Persistence and accumulation of environmental DNA from an endangered dragonfly. *Scientific Reports*, 11(1):1–8.

[128] Scott, A. J. and Wild, C. J. (1991). Fitting logistic regression models in stratified case-control studies. *Biometrics*, 47(2):497–510.

[129] Scott, A. J. and Wild, C. J. (1997). Fitting regression models to case-control data by maximum likelihood. *Biometrika*, 84(1):57–71.

[130] Servin, B. and Stephens, M. (2007). Imputation-based analysis of association studies: candidate regions and quantitative traits. *PLOS Genetics*, 3(7):e114.

[131] Shi, S., Yuan, N., Yang, M., Du, Z., Wang, J., Sheng, X., Wu, J., and Xiao, J. (2018). Comprehensive assessment of genotype imputation performance. *Human Heredity*, 83(3):107–116.

[132] Skinner, C. and Mason, B. (2012). Weighting in the regression analysis of survey data with a cross-national application. *Canadian Journal of Statistics*, 40(4):697–711.

[133] Smith, J., Bruley, C., Paton, I., Dunn, I., Jones, C., Windsor, D., Morrice, D., Law, A., Masabanda, J., Sazanov, A., et al. (2000). Differences in gene density on chicken macrochromosomes and microchromosomes. *Animal Genetics*, 31(2):96–103.

[134] Sobel, E. and Lange, K. (1996). Descent graphs in pedigree analysis: applications to haplotyping, location scores, and marker-sharing statistics. *American Journal of Human Genetics*, 58(6):1323–1337.

[135] Sohail, M., Maier, R. M., Ganna, A., Bloemendal, A., Martin, A. R., Turchin, M. C., Chiang, C. W., Hirschhorn, J., Daly, M. J., Patterson, N., et al. (2019). Polygenic adaptation on height is overestimated due to uncorrected stratification in genome-wide association studies. *Elife*, 8.

[136] Speliotes, E. K., Willer, C. J., Berndt, S. I., et al. (2010). Association analyses of 249,796 individuals reveal eighteen new loci associated with body mass index. *Nature Genetics*, 42(11):937–948.

[137] Staples, J., Nickerson, D. A., and Below, J. E. (2013). Utilizing graph theory to select the largest set of unrelated individuals for genetic analysis. *Genetic Epidemiology*, 37(2):136–141.

[138] StataCorp (2017). Stata statistical software: Release 15. College Station, TX: StataCorp LLC.

[139] Stephens, M. and Balding, D. J. (2009). Bayesian statistical methods for genetic association studies. *Nature Reviews. Genetics*, 10(10):681–690.

[140] Støer, N. C. and Samuelsen, S. O. (2012). Comparison of estimators in nested case–control studies with multiple outcomes. *Lifetime Data Analysis*, 18(3):261–283.

[141] Svishcheva, G., Axenovich, T., Belonogova, N., Duijn, C., and Aulchenko, Y. (2012). Rapid variance components-based method for whole-genome association analysis. *Nature Genetics*, 44(10):1166–1170.

[142] Thompson, E. (2011). The structure of genetic linkage data: from LIPED to 1M SNPs. *Human Heredity*, 71(2):86–96.

[143] Thompson, E. and Neel, J. (1996). Private polymorphisms: how many? how old? how useful for genetic taxonomies? *Molecular Phylogenetics and Evolution*, 5(1):220–231.

[144] Thompson, E. A. (2013). Identity by descent: variation in meiosis, across genomes, and in populations. *Genetics*, 194(2):301–326.

[145] Toma, C., Shaw, A. D., Allcock, R. J., Heath, A., Pierce, K. D., Mitchell, P. B., Schofield, P. R., and Fullerton, J. M. (2018). An examination of multiple classes of rare variants in extended families with bipolar disorder. *Translational Psychiatry*, 8(1):65.

[146] Ullah, E., Kunji, K., Wijsman, E. M., and Saad, M. (2019a). GIGI2: A fast approach for parallel genotype imputation in large pedigrees. *BioRxiv:533687*. doi:https://doi.org/10.1101/533687.

[147] Ullah, E., Mall, R., Abbas, M. M., Kunji, K., Nato, A. Q., Bensmail, H., Wijsman, E. M., and Saad, M. (2019b). Comparison and assessment of family-and population-based genotype imputation methods in large pedigrees. *Genome Research*, 29(1):125–134.

[148] van Binsbergen, R., Bink, M. C., Calus, M. P., van Eeuwijk, F. A., Hayes, B. J., Hulsegge, I., and Veerkamp, R. F. (2014). Accuracy of imputation to whole-genome sequence data in Holstein Friesian cattle. *Genetics Selection Evolution*, 46(1):1–13.

[149] Van Gestel, S., Houwing-Duistermaat, J. J., Adolfsson, R., van Duijn, C. M., and Van Broeckhoven, C. (2000). Power of selective genotyping in genetic association analyses of quantitative traits. *Behavior Genetics*, 30(2):141–146.

[150] Visscher, P. M., Hill, W. G., and Wray, N. R. (2008). Heritability in the genomics era—concepts and misconceptions. *Nature Reviews Genetics*, 9(4):255–266.

[151] Wang, M., Jakobsdottir, J., Smith, A. V., and McPeek, M. S. (2016). G-STRATEGY: optimal selection of individuals for sequencing in genetic association studies. *Genetic Epidemiology*, 40(6):446–460.

[152] Weaver, M. A. and Zhou, H. (2005). An estimated likelihood method for continuous outcome regression models with outcome-dependent sampling. *Journal of the American Statistical Association*, 100(470):459–469.

[153] Weil, C. (1970). Selection of the valid number of sampling units and a consideration of their combination in toxicological studies involving reproduction, teratogenesis or carcinogenesis. *Food and Cosmetics Toxicology*, 8(2):177–182.

[154] Weir, B. S. and Goudet, J. (2017). A unified characterization of population structure and relatedness. *Genetics*, 206(4):2085–2103.

[155] White, J. E. (1982). A two stage design for the study of the relationship between a rare exposure and a rare disease. *American Journal of Epidemiology*, 115(1):119–128.

[156] Whittemore, A. S. (1995). Logistic regression of family data from case-control studies. *Biometrika*, 82(1):57–67.

[157] Wijsman, E. M. (2012). The role of large pedigrees in an era of high-throughput sequencing. *Human Genetics*, 131(10):1555–1563.

[158] Wijsman, E. M., Rothstein, J. H., and Thompson, E. A. (2006). Multipoint linkage analysis with many multiallelic or dense diallelic markers: Markov chain–Monte Carlo provides practical approaches for genome scans on general pedigrees. *American Journal of Human Genetics*, 79(5):846–858.

[159] Wilcox, M. A., Pugh, E. W., Zhang, H., Zhong, X., Levinson, D. F., Kennedy, G. C., and Wijsman, E. M. (2005). Comparison of single-nucleotide polymorphisms and microsatellite markers for linkage analysis in the COGA and simulated data sets for Genetic Analysis Workshop 14: Presentation Groups 1, 2, and 3. *Genetic Epidemiology*, 29(S1):S7–S28.

[160] Wilhelm, S. and G, M. B. (2022). *tmvtnorm: Truncated Multivariate Normal and Student t Distribution*. R package version 1.5.

[161] Xing, C. and Xing, G. (2009). Power of selective genotyping in genome-wide association studies of quantitative traits. In *BMC proceedings*, volume 3, pages 1–5. BioMed Central.

[162] Xu, W. and Zhou, H. (2012). Mixed effect regression analysis for a cluster-based two-stage outcome-auxiliary-dependent sampling design with a continuous outcome. *Biostatistics*, 13(4):650–664.

[163] Yang, J., Benyamin, B., McEvoy, B. P., Gordon, S., Henders, A. K., Nyholt, D. R., Madden, P. A., Heath, A. C., Martin, N. G., Montgomery, G. W., et al. (2010). Common SNPs explain a large proportion of the heritability for human height. *Nature Genetics*, 42(7):565–569.

[164] Yang, J., Zaitlen, N. A., Goddard, M. E., Visscher, P. M., and Price, A. L. (2014). Advantages and pitfalls in the application of mixed-model association methods. *Nature Genetics*, 46(2):100–106.

[165] Yi, G. Y., Rao, J., and Li, H. (2016). A weighted composite likelihood approach for analysis of survey data under two-level models. *Statistica Sinica*, 26(2):569–587.

[166] Zaitlen, N. and Kraft, P. (2012). Heritability in the genome-wide association era. *Human Genetics*, 131(10):1655–1664.

[167] Zhang, Z., Ersoz, E., Lai, C.-Q., Todhunter, R. J., Tiwari, H. K., Gore, M. A., Bradbury, P. J., Yu, J., Arnett, D. K., Ordovas, J. M., and Buckler, E. S. (2010). Mixed linear model approach adapted for genome-wide association studies. *Nature Genetics*, 42(4):355–360.

[168] Zhao, L., Hsu, L., Holte, S., Chen, Y., Quiaoit, F., and Prentice, R. (1998). Maximum likelihood estimation of case-control family data. *Biometrika*, 85:299–315.

[169] Zhao, L. and Lipsitz, S. (1992). Designs and analysis of two-stage studies. *Statistics in Medicine*, 11(6):769–782.

[170] Zheng, H.-F., Tobias, J. H., Duncan, E., Evans, D. M., Eriksson, J., Paternoster, L., Yerges-Armstrong, L. M., Lehtimäki, T., Bergström, U., Kähönen, M., et al. (2012). WNT16 influences bone mineral density, cortical bone thickness, bone strength, and osteoporotic fracture risk. *PLOS Genetics*, 8(7):e1002745.

[171] Zhou, H., Weaver, M. A., Qin, J., Longnecker, M., and Wang, M. (2002). A semi-parametric empirical likelihood method for data from an outcome-dependent sampling scheme with a continuous outcome. *Biometrics*, 58(2):413–421.

[172] Zhou, X., Carbonetto, P., and Stephens, M. (2013). Polygenic modeling with bayesian sparse linear mixed models. *PLOS Genetics*, 9(2):e1003264.

[173] Ziyatdinov, A., Vázquez-Santiago, M., Brunel, H., Martinez-Perez, A., Aschard, H., and Soria, J. M. (2018). lme4qtl: linear mixed models with flexible covariance structure for genetic studies of related individuals. *BMC Bioinformatics*, 19(1):1–5.