Statistical Genetic Methods and Applications for Population Structure

Qian Sophia Zhang

A dissertation
submitted in partial fulfillment of the
requirements for the degree of

Doctor of Philosophy

University of Washington

2018

Reading Committee:

Bruce Spencer Weir, Chair

Sharon Ruth Browning

Kenneth Martin Rice

Timothy A. Thornton

Program Authorized to Offer Degree:
Biostatistics

University of Washington

Abstract

Statistical Genetic Methods and Applications for Population Structure

Qian Sophia Zhang

Chair of the Supervisory Committee:
Professor Bruce Spencer Weir
Biostatistics

In Chapter 1, we estimate what we call ancestry-specific allele frequencies using data about a sample of individuals from an admixed population. An admixed population is one descended from multiple ancestral populations. For a bi-allelic marker with genotypes measured in a sample of admixed, diploid individuals, and measured in each ancestry's reference panel consisting of marker genotypes measured for a sample of diploid individuals from that ancestral population, one can infer the ancestral origin of each allele copy of each admixed individual at the marker. Since a diploid individual has two allele copies at a marker, their ancestral origins form a pair, which we call an "ancestry pair."

In a genome-wide association study (GWAS) of an admixed population, such as African-Americans or Hispanic-Americans, for a marker with a trait-associated allele, knowing each ancestry's ancestry-specific allele frequency, the relative frequency of an allele amongst the admixed population's chromosomes that are of a certain ancestry at the marker, can inform the design of a replication GWAS study. We provide methods for obtaining maximum likelihood estimates (MLEs) of ancestry-specific allele frequencies from observations on marker genotypes and ancestry pairs, with phase of genotype relative to ancestry pair unknown, obtained on a sample of admixed individuals. The difficulty is that when an individual is heterozygous at a marker and has ancestry pair with differing ancestries at the marker, it is not generally known which allele is of which ancestry. The problem is formally equivalent to

that of phase being unknown for multi-locus genotypes when haplotype frequencies are to be estimated, and we invoke the Expectation-Maximization (EM) algorithm we have described previously (Zhang et al., 2016). Instead of frequencies of haplotypes of alleles at two loci, we estimate frequencies of allele-ancestry pairs. These joint frequencies then provide marker allele frequencies conditional on ancestry, i.e. ancestry-specific allele frequencies.

Our new two-ancestry approach involves conducting a preliminary check for the existence of boundary allele-ancestry pair frequency estimates, where one or more combinations of allele and ancestry have estimated frequencies of zero, and uses different analytical formulas to find an MLE depending on which counts of allele and ancestry combinations are zero. Analytical formulas are used to obtain an MLE, whether allele-ancestry pair frequency estimates are on the boundary or not. Our new three-ancestry approach involves, under certain conditions on the data, employing an analytical formula for an MLE, and when those conditions do not hold, a two-dimensional grid search that can find a boundary MLE. As current SNP-array and DNA sequence data are revealing many SNPs with low minor allele frequencies, these boundary conditions apply to many markers. We evaluate our two- and three- ancestry approaches on simulated data, and demonstrate their application to 1000 Genomes data for an African-American population, with African and European ancestries, and for a Puerto Rican population, with African, European and Native American ancestries.

In Chapter 2, we estimate individual inbreeding coefficients. Many population genetic activities, ranging from evolutionary studies to association mapping to forensic identification, rely on estimates of individual inbreeding coefficients and between-individual coancestry coefficients. We examined existing method of moments estimators (MMEs) and maximum likelihood estimators (MLEs) introduced as estimators of an individual j's total inbreeding coefficient F_j (i.e. Hall et al.'s simple method of moments estimator, GCTA^{I} MME, GCTA^{III} MME, and Hall et al.'s first maximum likelihood estimator) or as an estimator of within-population inbreeding coefficient β_j (i.e. Weir and Goudet's MME), as well as our novel likelihood estimator of β_j. For each MME, we considered the estimator that combines over loci using the ratio of means (RM) method (a weighted estimator), and the mean of ratios (MR) method (an unweighted estimator). We bootstrapped over loci to provide interval estimates for MMEs and MLEs. We evaluated all methods, except Hall et al.'s simple MME, on simulated data. We found that, despite previously published claims of unbiasedness for some evaluated estimators, none of these estimators were unbiased for F_j, and the Weir and Goudet MME RM had very low bias for β_j. We showed that evaluated estimators were better at estimating certain non-F_j parameters (which we refer to as "pedigree parameters" or "parametric pedigree values") than at estimating F_j, in that they had lower squared bias and lower MSE for these other parameters than for F_j, and tracked more closely with these other parameters than with F_j. In other words, estimators intended to estimate F_j have been mischaracterized as F_j estimators, and are actually better at estimating other parameters. Lastly, we recommend the Weir and Goudet MME RM for estimation of β_j.

In Chapter 3, we estimate individual pair coancestry coefficient $\theta_{jj'}$ for individuals j and j' and population-specific F_{ST}. We show that various estimators (simple MME, Weir and Goudet weighted $\hat{\beta}_{jj',w}$ and unweighted $\hat{\beta}_{jj',u}$ MMEs, GCTAw and GCTAu MMEs) are better estimators of parameters that we call "pedigree-determined parameters" than of $\theta_{jj'}$. We then review existing estimators of population-specific F_{ST} (Weir and Goudet MME, the Bayesian estimator of BayeScan software, the Bayesian estimator of Bayenv2.0 software). We introduce a novel, two-population Bayesian estimator of population-specific F_{ST} that allows populations' allele frequencies to be dependent random variables (so that negative population-specific F_{ST} estimates are possible), while modeling marginal allele frequencies as Beta-distributed. We compare methods on simulated data, demonstrate them on real data from the HealthABC cohort study, and make recommendations about when to use which estimator.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

ACKNOWLEDGMENTS

Tremendous thanks goes to my adviser Dr. Bruce S. Weir (BSW) for his generous time and attention advising me, including holding research meetings, giving me feedback on my writing and presentation slides, and responding to research emails quickly. I also thank him for providing me with vital research funding and a workspace.

I thank my reading committee Dr. Sharon Browning (SB), Dr. Ken Rice (KR), Dr. Tim Thornton (TT), and BSW for reading and providing feedback on multiple drafts of my dissertation. I thank Dr. Cinnamon Hillyard for serving as my Graduate School Representative. I thank my entire supervisory committee for discussing my research at committee meetings.

For their advising on papers not in this dissertation (Zhang et al., 2015, 2016; Kerr et al., 2017), which included research emails, conversations, and reading draft papers, I thank Dr. SB, Dr. Brian Browning, and Dr. Katie Kerr. Considerable credit goes to Dr. SB for her primary role advising me on two of my first author papers, providing me with research funding, and contributing to a welcome Browning Lab environment through lab meetings and the like. For working with me on a paper not in this dissertation (Thornton et al., 2012), discussing research outside of committee meetings, and checking on how I was doing in graduate school, I thank Dr. TT. For lab rotations, I thank Dr. SB, Dr. Gail Jarvik, and Dr. Pei Wang.

I thank everyone else who has discussed research (e.g. tools, general advice, particular projects) with me. Conversations helped me develop ideas or practice teaching. They include Dr. Mathias Drton, Dr. Scott Emerson, Dr. Mary Emond, population genetics seminar participants (including Dr. Joe Felsenstein who discussed island model transition equations with me), University of Washington (UW) Biostat Slack channel participants (including Lucy Gao (LG) and channel creator Jean Feng (JF)), Jay Garlapati, Jithin George, Dr. Jérôme Goudet, Kelsey Grinde, UW Biostat student seminar participants (including Dr. Asad Haris who mentioned copulas), Dr. Peter Hoff, Dr. Marshall Horwitz (MH), Dr. Deepti Jain (who asked me about ASAFE for research), Dr. Vladimir Minin, Anna Plantinga (AP), Pavel Panchekha (PP), Dr. Noah Simon, Arjun Sondhi, Xiaowen Tian, Drs. Elizabeth Blue, Elizabeth Thompson, and Ellen Wijsman through the statistical genetics journal club and when bumping into each other, Dr. Jonathan Wakefield, Dr. Sam Wang, David Whitney, Katie Wilson (KW), and Dr. Sen Zhao. Rebecca Ferrell practiced my general exam talk with me. Sanne Aalbers (SA), JF, Dr. Jane Lange, and Dr. Ethan Jewett practiced my final exam talk with me.

I thank UW Genetic Analysis Center members (e.g. Jai Broome, Dr. Stephanie Gogarten, Dr. Ben Heaver, and Dr. Xiuwen Zheng), Robert Moulton, and the people of the eScience Institute, for discussing computer programming with me or providing programming resources. I thank Dr. KR and Dr. Lurdes Inoue for letting me audit their classes, as well as Dr. Marco Carone for sharing his survival analysis course materials and Dr. Barbara McKnight for sharing her course materials on effective teaching. I thank Gitana Garofalo for helping me and other students navigate graduate school, and the UW Biostat fiscal team for making sure I got paid and encouraging me.

For fun and educational dinner meetings and retreats, I thank the Medical Scientist Training Program (MSTP) people (e.g. Marcie Buckner, Dr. MH, Dr. Mary-Claire King, Dr. Julie Overbaugh, and of course the other MSTP students). For checking on how I was doing during graduate school and practicing clinical skills with me to help me transition back to medical school, I thank my amazing college mentor Dr. Glen Tamura! Mentors like him are salt of the Earth.

I thank my family and friends for their moral support. For keeping me company at work, I thank Paul Fisk, Jon Fintzi, the Graduate Opportunities and Minority Achievement Program (GO-MAP) writing group, Dr. Jason C. Liang, Tedrick Leung, Richard Li, Dr. Shih-Yin Lin, the National Center for Faculty Development and Diversity dissertation writing support group (e.g. Sarah Nelson), PP, AP, Katherine Tan, and KW. I thank Dr. Carolyn W. Jackson for organizing the GO-MAP writing group. For other encouragement or being my sounding board, I thank Dr. Lisa Brown, Dr. Hui Feng (Joy) Chen, SA, Shengying Chen (who provided me with a large Teddy bear), LG, Dr. Alexandra Grier, Dr. Fiona Grimson, Dr. Yu Huang, Tung Le, Dr. Sherry Lee, Dr. Steven Levitte, Dr. Natalie Miller, Dr. Andrew Nato, PP, Anna Simpson, Dr. Tamar Sofer (who told me she found my research useful), Dr. Theo Sottero, Dr. Yvonne Tyler, Dr. Yuefeng Zhang, James Zhang, and Dr. Michael Zhang. Sorry if I've missed someone!

DEDICATION

I dedicate this dissertation to researchers and teachers - especially graduate student researchers and teachers - who seek to preserve and grow knowledge for the good of society.

Chapter 1

ANCESTRY-SPECIFIC ALLELE FREQUENCY ESTIMATION (ASAFE)

1.1 Abstract

An admixed population is one descended from multiple ancestral populations. For a bi-allelic marker with genotypes measured in a sample of admixed, diploid individuals, and measured in each ancestry's reference panel consisting of marker genotypes measured for a sample of diploid individuals from that ancestral population, one can infer the ancestral origin of each allele copy of each admixed individual at the marker. Since a diploid individual has two allele copies at a marker, their ancestral origins form a pair, which we call an "ancestry pair."

In a genome-wide association study (GWAS) of an admixed population, such as African-Americans or Hispanic-Americans, for a marker with a trait-associated allele, knowing each ancestry's ancestry-specific allele frequency, the relative frequency of an allele amongst the admixed population's chromosomes that are of a certain ancestry at the marker, can inform the design of a replication GWAS study. We provide methods for obtaining maximum likelihood estimates (MLEs) of ancestry-specific allele frequencies from observations on marker genotypes and ancestry pairs, with phase of genotype relative to ancestry pair unknown, obtained on a sample of admixed individuals. The difficulty is that when an individual is heterozygous at a marker and has ancestry pair with differing ancestries at the marker, it is not generally known which allele is of which ancestry. The problem is formally equivalent to that of phase being unknown for multi-locus genotypes when haplotype frequencies are to be estimated, and we invoke the Expectation-Maximization (EM) algorithm we have described previously (Zhang et al., 2016). Instead of frequencies of haplotypes of alleles at two loci, we estimate frequencies of allele-ancestry pairs. These joint frequencies then provide marker

allele frequencies conditional on ancestry, i.e. ancestry-specific allele frequencies.

Our new two-ancestry approach involves conducting a preliminary check for the existence of boundary allele-ancestry pair frequency estimates, where one or more combinations of allele and ancestry have estimated frequencies of zero, and uses different analytical formulas to find an MLE depending on which counts of allele and ancestry combinations are zero. Analytical formulas are used to obtain an MLE, whether allele-ancestry pair frequency estimates are on the boundary or not. Our new three-ancestry approach involves, under certain conditions on the data, employing an analytical formula for an MLE, and when those conditions do not hold, a two-dimensional grid search that can find a boundary MLE. As current SNP-array and DNA sequence data are revealing many SNPs with low minor allele frequencies, these boundary conditions apply to many markers.

We evaluate our two- and three- ancestry approaches on simulated data, and demonstrate their application to 1000 Genomes data for an African-American population, with African and European ancestries, and for a Puerto Rican population, with African, European and Native American ancestries.

1.2 Introduction

Human genome-wide association studies (GWASs) have focused on populations of European origin. GWAS of an admixed population, one that has descended from multiple ancestral populations, could identify novel SNP-trait associations, since mutations have different frequencies in different human lineages. In a GWAS of an admixed population, one collects genotypes and phenotypes on a sample of admixed individuals, and can infer local ancestries from admixed genotypes and each ancestral population's reference panel of genotypes. Admixed individuals' global ancestry proportions and local ancestry pairs at each marker are not fixed by design. They are not determined before obtaining data on admixed individuals, so are considered random. An example of such a GWAS is given by the Hispanic Community Health Study, in which Hispanic individuals' local ancestries are inferred after recruitment of individuals (Burkart et al., 2018; Browning et al., 2016).

After discovering that an allele is significantly associated with a trait in an initial GWAS of admixed individuals, it may be useful to know the allele's ancestry-specific frequencies, since these frequencies inform the choice of population to study in follow-up work. In order to maximize power to replicate the discovery signal, one would ideally study individuals primarily of the ancestral background with highest ancestry-specific allele frequency. For a sample of Hispanic individuals, an ancestry-specific allele frequency would be the relative frequency of the allele amongst chromosomes in the Hispanic population that have a particular ancestral origin (either African, European, or Native American) at the marker.

To obtain estimates of ancestry-specific allele frequencies, one might find the observed allele frequency in each reference panel. However, this approach is infeasible when a SNP for which one wants ancestry-specific allele frequency estimates is typed in the admixed sample and not in a reference panel, which occurs when the reference panels are genotyped on a SNP array and when the admixed individuals have sequence data. For example, the Native American reference panel used by Martin et al. (2017) consisted of SNP array data (Mao et al., 2007).

The ADMIXTURE and STRUCTURE programs can use admixed genotypes to estimate ancestry-specific allele frequencies, but require genotypes from reference individuals of known ancestry in order to identify the frequencies with particular ancestries (Alexander et al., 2009; Pritchard et al., 2000). Furthermore, ADMIXTURE assumes linkage equilibrium amongst markers, and STRUCTURE's linkage model assumes linkage equilibrium amongst markers descended from the same ancestral population (Falush et al., 2003). In contrast, we utilize estimates of local ancestry, such as those from the RFMix program (Maples et al., 2013). Estimates of local ancestry from such programs utilize haplotype frequencies, and thus make use of linkage disequilibrium. At markers genotyped in the admixed sample and all the reference panels, one could possibly directly obtain ancestry-specific allele frequencies from admixed individuals' genotypes and admixed individuals' corresponding local ancestry estimates that are also phased with respect to genotypes. However, analysts may not save or report the phase of local ancestry estimates with respect to genotypes in the admixed sample. For example, local ancestry estimates are reported in Phase 1 of the Thousand Genomes Project for African and Puerto Rican individuals, but their phase with respect to admixed individuals' genotypes is not given (Kenny, 2017).

Zhang et al. (2016) derived an EM algorithm (ASAFE) to obtain maximum likelihood estimates (MLEs) of ancestry-specific allele frequencies at a bi-allelic marker in a three-way admixed diploid population, given admixed local ancestry pairs and genotypes at the marker, with each admixed individual's genotype allele order relative to ancestry order unknown. One might obtain admixed individuals' local ancestry estimates that are phased with respect to individuals' genotypes and then directly count alleles of each ancestry to obtain ancestry-specific allele frequency estimates. However, an advantage of ASAFE over this alternative approach is that ASAFE can be used when local ancestry estimates are not phased with respect to genotypes. For example, admixed local ancestry pairs and admixed genotypes in the 1000 Genomes Phase 1 data used here are not phased with respect to each other, and some local ancestry calling programs such as HAPMIX (Price et al., 2009) do not produce local ancestry estimates that are phased with respect to genotypes. A second advantage is ASAFE is applicable to SNPs genotyped in the admixed sample that were not genotyped in a reference panel, since one can impute an individual's ancestry pair at such a SNP using ancestry pairs estimated at nearby SNPs.

We extend the work of the original ASAFE paper (Zhang et al., 2016) in a number of ways. First, we present a likelihood framework for any finite number of ancestries K, with $K = 2$ and $K = 3$ ancestries given as examples. Secondly, it follows from score equations that a stationary point of the log likelihood function has corresponding marginal frequencies that are observed sample frequencies, and we make use of these sample frequencies to constrain the set of points in the domain to consider when searching for joint allele and ancestry probability maximum likelihood estimates used to obtain ancestry-specific allele frequency estimates. (We use the terms "probability" and "frequency" interchangeably.) These constraints reduce the number of free parameters that the log likelihood function depends on, from $(2K-1)$ parameters down to $(K-1)$ parameters, and inspires our two-and three-ancestry approaches. In the two-ancestry case, because the log likelihood is a function of a single parameter, we propose an analytical approach for obtaining MLE's. In the three-ancestry case, because the log likelihood is a function of two parameters, we propose an analytical approach for obtaining MLE's when certain conditions on the observed counts of genotype and ancestry pair combinations are satisfied. When conditions are not satisfied, we propose a two-dimensional grid search. Critically, these approaches allow the calculation of exact boundary MLEs, which the ASAFE EM implementation cannot achieve. Thirdly, we prove conditions on the observed counts that guarantee that a stationary point of the log likelihood function is an MLE. These conditions can be checked automatically with computer code. Lastly, we give a condition (which can also be checked automatically) that is sufficient, but not necessary, for an MLE to lie in the interior of the parameter space. The ASAFE EM algorithm implementation increments an estimate $\hat{p}_{ga}$ of the joint probability p_{ga} of allele g with ancestry a (defined in Section 1.3.1) by $\epsilon = 10^{-8}$ above boundary value 0 and decrements a $\hat{p}_{ga}$ estimate by ϵ below boundary value 1, in order to avoid divide by 0 error in the algorithm. One might then be concerned that the ASAFE EM implementation does not return an MLE $\hat{\mathbf{p}}$ that takes a boundary value. When the condition sufficient for the MLE to lie in the interior of the domain of the likelihood function holds, one knows that the ASAFE EM implementation cannot have missed a boundary MLE, because an MLE is not on the boundary.

In what follows, we give the likelihood framework for any finite number K of ancestries. We then detail "Extended ASAFE," a two-ancestry analytical approach for obtaining MLEs and their standard error estimators and a three-ancestry mixed analytical and grid search approach for obtaining MLEs. We then describe applying the two-ancestry and three-ancestry approaches to simulated data, applying the two-ancestry approach to 1000 Genomes African data, applying the three-ancestry approach to 1000 Genomes Puerto Rican data, comparing our three-ancestry approach to the ASAFE EM algorithm on 1000 Genomes Puerto Rican data, and using ancestry-specific allele frequency estimates to estimate population-specific F_{ST}. We then describe results of all these endeavors and conclude with a discussion.

1.3 Methods

1.3.1 Methods: Likelihood Framework for K Ancestries.

1.3.1.1 Notation

A set of n individuals is typed at a series of SNPs and have ancestry pairs at each SNP estimated e.g. by RFMix (Maples et al., 2013). The difficulty is that if the individual is heterozygous at a SNP and has two different ancestries at the SNP, it is not known which allele is of which ancestry. This is illustrated in Figure 1.1.

We refer to the allele-ancestry pair ga for allele type g of ancestry type a as a ga-allele and write its probability as p_{ga}. Assuming that ga-alleles come together independently, individuals resulting from the union of ga and $g'a'$ alleles, in that order, have probability $p_{ga,g'a'} = p_{ga}p_{g'a'}$. The observed genotypes have unknown phase with respect to ancestry pairs so individuals observed to have alleles g, g' and ancestries a, a' have probabilities

$$p_{gg',aa'} = \begin{cases} 2[p_{ga}p_{g'a'} + p_{ga'}p_{g'a}] & g \neq g' \quad \text{and} \quad a \neq a' \\ 2p_{ga}p_{g'a'} & g = g' \quad \text{and} \quad \mathrm{a} \neq \mathrm{a}' \text{ ,or } \mathrm{g} \neq \mathrm{g}' \quad \text{and} \quad \mathrm{a} = \mathrm{a}' \\ p_{ga}p_{g'a'} & g = g' \quad \text{and} \quad a = a' \end{cases} \tag{1.1}$$

Note that $p_{ga,g'a'}$ is the probability that allele g is of ancestry a (ga) and allele g' is of ancestry a' ($g'a'$), which differs from $p_{gg',aa'}$, the probability of having genotype gg' and ancestry pair aa', where we don't know which allele (g or g') is of which ancestry (a or a').

Genotype/ancestry probabilities can be summed over either category to provide genotype probabilities and ancestry-pair probabilities:

$$\text{Genotypes} \qquad p_{gg',..} = \sum_{a=1}^{K} p_{gg',aa} + \frac{1}{2}\sum_{a=1}^{K}\sum_{a'=1,a\neq a'}^{K} p_{gg',aa'}$$

$$\text{Ancestry pairs} \qquad p_{..,aa'} = \sum_{g=0}^{1} p_{gg,aa'} + \frac{1}{2}\sum_{g=0}^{1}\sum_{g'=0,g\neq g'}^{1} p_{gg',aa'} \,.$$

These, in turn, can be summed to provide allele probabilities $p_{g.}$'s and ancestry probabilities $p_{.a}$'s:

$$\text{Alleles} \qquad p_{g.} = p_{gg,..} + \frac{1}{2} \sum_{gg', g \neq g'} p_{gg',..}$$
$$\text{Ancestries} \qquad p_{.a} = p_{..,aa} + \frac{1}{2} \sum_{aa', a \neq a'} p_{..,aa'}.$$

Allele probabilities $p_{g.}$'s and ancestry probabilities $p_{.a}$'s also follow by summing ga-allele probabilities p_{ga}'s of ga-alleles for SNP allele $g; g = 0, 1$ and ancestry $a; a = 1, 2, \ldots K$ over either subscript:

$$\text{Alleles} \qquad p_{g.} = \sum_{a=1}^{K} p_{ga} \text{ and} \tag{1.2}$$
$$\text{Ancestries} \qquad p_{.a} = \sum_{g=0}^{1} p_{ga}. \tag{1.3}$$

Because counts of genotype/ancestry pair combinations are multinomially distributed, the marginal genotype probabilities $p_{gg',..}$'s and the marginal ancestry pair probabilities $p_{..,aa'}$'s have observed proportions as maximum likelihood estimates (MLEs). Similarly, the marginal allele probabilities $p_{g.}$'s and the marginal ancestry probabilities $p_{.a}$'s have their observed proportions as MLEs under the assumption of Hardy-Weinberg equilibrium (Weir, 1996).

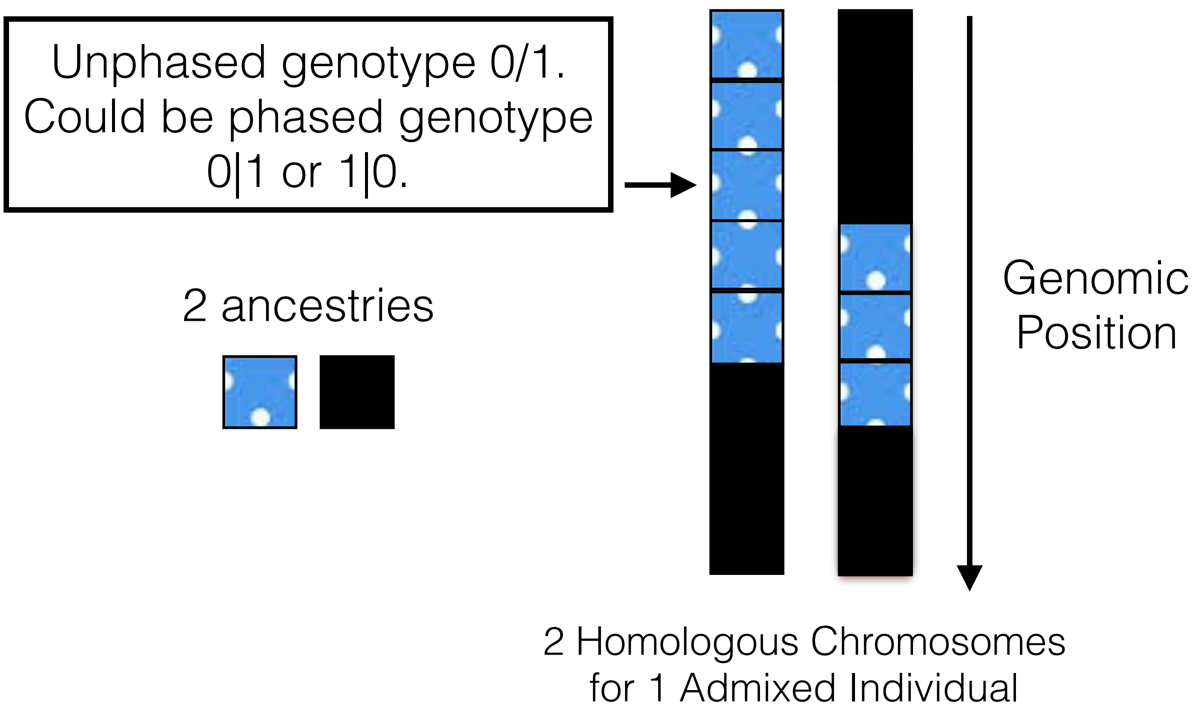

Figure 1.1: Available Data: At a SNP, we know admixed individuals' genotypes and ancestry pairs, but not the genotype order (i.e. phase) relative to the ancestry pair. The horizontal arrow points to a marker. This admixed individual has unphased genotype 0/1 and ancestry pair dotted blue | solid black at the marker. An unphased genotype 0/1 is consistent with a phased genotype of 0|1, in which case the 0 allele is of dotted blue ancestry and the 1 allele of solid black ancestry, and a phased genotype of 1|0, in which case the 1 allele is of dotted blue ancestry and the 0 allele of solid black ancestry.

1.3.1.2 Estimation

If the number of individuals with a genotype/ancestry pair combination with genotype phase relative to ancestry pair unknown is written $n_{gg',aa'}$, the likelihood function is

$$L(\{p_{ga}\}) = C \prod_{gg'} \prod_{aa'} \left[p_{gg',aa'}(\mathbf{p}) \right]^{n_{gg',aa'}},$$

where $\mathbf{p}$ is a vector of p_{ga}'s, which is the same as the collection $\{p_{ga}\}$ of p_{ga}'s, and C is an arbitrary constant. The domain of the likelihood is the probability simplex $D = \{\mathbf{p} \in [0,1]^{(2)(|K|)}$ s.t. $\sum_{g\in\{0,1\}} \sum_{a\in S_a} p_{ga} = 1\}$, where $S_a = \{1, ..., K\}$ is the set of ancestries of cardinality K, and Equation 1.1 gives $p_{gg',aa'}(\mathbf{p})$ as a function of p_{ga}'s, assuming Hardy-Weinberg equilibrium (HWE). The log likelihood function is

$$\ln(L(\{p_{ga}\})) = \ln(C) + \sum_{gg'} \sum_{aa'} n_{gg',aa'} \ln(p_{gg',aa'}(\mathbf{p})).$$

If a count $n_{gg',aa'} = 0$, then we re-define the likelihood function to exclude the factor $\left[p_{gg',aa'}(\mathbf{p}) \right]^{n_{gg',aa'}}$, in order to keep the log likelihood function defined at a $\mathbf{p}$ with a component $p_{ga} = 0$. If we did not re-define the likelihood function, and for example, $n_{00,AA}$ were 0, then when $p_{0A} = 0$, the term $(n_{00,AA})ln(p_{00,AA}(\mathbf{p}))$ would evaluate to $(0)ln(0)$, which is undefined. Excluding terms with 0 counts from the log likelihood function has been done previously by Casella and Berger (2002) in Example 7.2.7 on page 318.

It will be convenient to define $X_{ga} := 2n_{gg,aa} + \sum_{a'=1,a'\neq a}^{K} n_{gg,aa'} + \sum_{g'=0,g'\neq g}^{1} n_{gg',aa}$, the observed number of gametes that carry allele g of ancestry a. By making use of Equation 1.1 that gives $p_{gg',aa'}(\mathbf{p})$ as a function of p_{ga}'s, we can rewrite the log likelihood function in terms of X_{ga}'s:

$$\begin{aligned} l(\mathbf{p}) := \ln(L(\{p_{ga}\})) &= \ln(C) + \sum_{g=0}^{1} \sum_{a=1}^{K} X_{ga} \ln(p_{ga}) \\ &+ \sum_{gg' \in S_{gg'}, g\neq g'} \sum_{aa' \in S_{aa'}, a\neq a'} n_{gg',aa'} \ln(2p_{ga}p_{g'a'} + 2p_{ga'}p_{g'a}), \end{aligned} \tag{1.4}$$

where the set $S_{gg'}$ of genotypes is $\{0/0, 0/1, 1/1\}$ and the set of ancestry pairs $S_{aa'}$ consists of all combinations of alleles a and a'. Writing the log likelihood function in this form helps us see that when an $X_{ga} > 0$, an MLE $\hat{\mathbf{p}}$ cannot have component $\hat{p}_{ga} = 0$, which we can prove by contradiction. Take $X_{ga} > 0$ and MLE $\hat{\mathbf{p}}$. Assume component $\hat{p}_{ga}$ of MLE $\hat{\mathbf{p}}$ is 0. Then the term $X_{ga}\ln(\hat{p}_{ga}) = -\infty$, so that $ln(L(\hat{\mathbf{p}})) = -\infty$, contradicting $\hat{\mathbf{p}}$ being an MLE, so $\hat{p}_{ga}$ cannot be 0. Since an X_{ga} being strictly positive implies that MLE $\hat{\mathbf{p}}$ cannot have component $\hat{p}_{ga} = 0$, we see that when every $X_{ga} > 0$, an MLE $\hat{\mathbf{p}}$ must be in the interior of the parameter space D. The contrapositive statement says that if an MLE is on the boundary of the parameter space, then at least one $X_{ga} = 0$. Considering every combination of the X_{ga}s being zero would consider all cases where an MLE $\hat{\mathbf{p}}$ could lie on the boundary.

We define Lagrangian function $F(\lambda, \mathbf{p}) := ln(L(\mathbf{p})) + \lambda(\sum_{g=0}^{1}\sum_{a=1}^{K} p_{ga} - 1)$ and set its partial derivatives with respect to each p_{ga} and with respect to λ to 0. One might be concerned that for some g and a, $\partial F/\partial p_{ga}(\mathbf{p})$ is undefined at $\mathbf{p}$ with $p_{ga} = 0$, due to the term X_{ga}/p_{ga}. However, when $X_{ga} > 0$, we do not consider points $\mathbf{p}$ with $p_{ga} = 0$ when we maximize Lagrangian function $F(\lambda, \mathbf{p})$, because an MLE $\hat{\mathbf{p}}$ cannot have $\hat{p}_{ga} = 0$, and when $X_{ga} = 0$, recall we excluded factors $\left[p_{gg',aa'}(\mathbf{p})\right]^{n_{gg',aa'}}$ with 0 $n_{gg',aa'}$ counts from $\ln(L(\mathbf{p}))$, hence excluding $X_{ga}\ln(p_{ga})$ from $\ln(L(\mathbf{p}))$, so X_{ga}/p_{ga} is not a term in $\partial F/\partial p_{ga}(\mathbf{p})$.

Setting partial derivatives of the Lagrangian function to 0 yields the following equation for each g and a:

$$\begin{aligned} p^*_{ga} &= \tilde{p}_{gg,aa} + \frac{1}{2}\left(\sum_{a'=1,a'\neq a}^{K} \tilde{p}_{gg,aa'} + \sum_{g'=0,g'\neq g}^{1} \tilde{p}_{gg',aa}\right) \\ &+ \sum_{g'\neq g}\sum_{a'\neq a} \frac{\tilde{p}_{gg',aa'} p^*_{ga} p^*_{g'a'}}{2(p^*_{ga}p^*_{g'a'} + p^*_{ga'}p^*_{g'a})}, \end{aligned} \tag{1.5}$$

where $\tilde{p}_{gg'aa'}$ is defined as $n_{gg',aa'}/n$, and a starred p^*_{ga} indicates a p_{ga} component of $\mathbf{p}^*$, a stationary point of $\ln(L(\mathbf{p}))$.

From Equation 1.5, we see that a stationary point $\mathbf{p}^*$ has the following property: Marginal probability estimates $p^*_{.a}$ and $p^*_{g.}$ are observed sample frequencies $\tilde{p}_{.a}$ and $\tilde{p}_{g.}$: $p^*_{.a} = \sum_{g=0}^{1} p^*_{ga} =$

$\tilde{p}_{.a}$ and $p^*_{g.} = \sum_{a=1}^{K} p^*_{ga} = \tilde{p}_{g.}$. Since an MLE $\hat{\mathbf{p}}$ is one of the points $\mathbf{p}^*$ that solve the system of equations obtained by considering Equation 1.5 for every g and a, an MLE $\hat{\mathbf{p}}$ must also have this property. Hence, when searching for an MLE $\hat{\mathbf{p}}$, we need only consider points $\mathbf{p}$ that have the property that $\sum_{g=0}^{1} p_{ga} = \tilde{p}_{.a}$ and $\sum_{a=1}^{K} p_{ga} = \tilde{p}_{g.}$. Restricting our search for an MLE $\hat{\mathbf{p}}$ to points $\mathbf{p}$ with this property reduces the number of free parameters in our log likelihood function $\ln(L(\mathbf{p}))$ from $(2K-1)$ p_{ga} parameters down to $(2-1)*(K-1) = (K-1)$ p_{ga} parameters. This reduction in the dimensionality of the problem inspires our two- and three-ancestry estimation approaches. When we restrict to points with this property, the domain of the log likelihood function $\ln(L(\mathbf{p}))$ has the constraint that for every g and a, $p_{ga} \in [0, min(\tilde{p}_{.a}, \tilde{p}_{g.})]$, a fact we will use in our two and three-ancestry estimation approaches.

Equation 1.5 can also be obtained from the EM algorithm, considering incomplete observed categories to be genotype gg' and ancestry pair aa' combinations gg', aa''s with genotype order relative to ancestry pair unknown, and considering complete unobserved categories to be genotype gg' and ancestry pair aa' combinations $ga/g'a'$ with genotype order relative to ancestry pair known. If the types gg', aa' could be distinguished by which allele was of which ancestry, then maximum likelihood estimates for p_{ga} would simply be the sample proportions of those ga-combinations. In the absence of that, the EM algorithm starts by estimating (E) the counts $n_{ga/g'a'}$'s of hidden classes $\{ga/g'a'$ such that $g \neq g', a \neq a'\}$ of individuals carrying allele g with ancestry a and allele g' with ancestry a'. These estimated counts are based on the Hardy-Weinberg assumption that ga-alleles come together in Hardy-Weinberg proportions:

$$n_{ga/g'a'} = \frac{n_{gg',aa'} p_{ga} p_{g'a'}}{p_{ga} p_{g'a'} + p_{ga',g'a}}.$$

The maximization (M) step then proceeds by counting the number of ga combinations in the data and dividing by the number $2n$ of alleles in the data. Equation 1.5 can be used iteratively. Initial guesses for the p^*_{ga} are used on the right-hand side of the equation and the next iterates are given by the left-hand side. These left-hand side values are subsequently used in the right-hand side, and so on, until successive iterates are sufficiently close (within a user-specified ϵ of each other) that convergence can be declared. Alternatively, one could declare convegence when successive iterates evaluate to log likelihood values that are within ϵ of each other. The iterative EM algorithm will give a stationary point for the likelihood with respect to each p_{ga}. Different starting values may lead to different stationary points, and maximum likelihood estimates $\{\hat{p}_{ga}\}$ are identified by comparing likelihoods evaluated at stationary points.

To be more concise, we re-write Equation 1.5 with $X_{ga}/(2n)$, where recall X_{ga} was defined earlier as $2n_{gg,aa} + \sum_{a'=1,a'\neq a}^{K} n_{gg,aa'} + \sum_{g'=0,g'\neq g}^{1} n_{gg',aa}$, replacing the first three terms on the right-hand side:

$$p_{ga}^{*} \quad = \quad X_{ga}/(2n) + \sum_{g'\neq g}\sum_{a'\neq a} \frac{\tilde{p}_{gg',aa'} p_{ga}^{*} p_{g'a'}^{*}}{2(p_{ga}^{*} p_{g'a'}^{*} + p_{ga'}^{*} p_{g'a}^{*})}. \tag{1.6}$$

Following Weir and Cockerham (1979), we examine the system of equations obtained by considering Equation 1.6 for all g and a to detect possible boundary solutions. We see immediately two situations that allow simplification. If $X_{ga} = 0$, then

$$p_{ga}^{*} \quad = \quad \sum_{g'\neq g}\sum_{a'\neq a} \frac{\tilde{p}_{gg',aa'} p_{ga}^{*} p_{g'a'}^{*}}{2(p_{ga}^{*} p_{g'a'}^{*} + p_{ga'}^{*} p_{g'a}^{*})}, \tag{1.7}$$

so that $p_{ga}* = 0$ or

$$1 \quad = \quad \sum_{g'\neq g}\sum_{a'\neq a} \frac{\tilde{p}_{gg',aa'} p_{g'a'}^{*}}{2(p_{ga}^{*} p_{g'a'}^{*} + p_{ga'}^{*} p_{g'a}^{*})}. \tag{1.8}$$

Similarly, if $n_{gg',aa'} = 0$ for all $g' \neq g, a \neq a'$, then

$$p_{ga}^{*} \quad = \quad X_{ga}/(2n). \tag{1.9}$$

The parameters of interest are conditional probabilities $\{p(g=1|a), a \in \{1, ..., K\}\}$ for a bi-allelic SNP with alleles $g = 0$ and 1 and an admixed population descended from K ancestral populations 1, ..., K. We previously called these probabilities ancestry-specific allele frequencies (Zhang et al., 2016).

1.3.2 Methods: Two Ancestries.

In the case of two ancestries, the problem of estimating ancestry-specific allele frequencies is formally equivalent to the problem of estimating linkage disequilibrium for pairs of biallelic markers, when there are double heterozygotes of unknown phase, considered by Weir and Cockerham (1979). We extend their discussion of boundary conditions and provide estimators of standard errors for ancestry-specific allele frequency estimators $\hat{p}(1|A)$ and $\hat{p}(1|E)$.

For notational convenience, we label the two ancestries as A and E (e.g. African and European for African-American data) and keep 0 and 1 for the two alleles at a SNP. Then there are four allele/ancestry combinations, 0A, 0E, 1A, and 1E. Constraints imposed by marginal probabilities on p_{ga} probabilities are $\sum_g p_{ga} = p_{.a}$ and $\sum_a p_{ga} = p_{g.}$ for all g and a (e.g. $p_{0A} + p_{1A} = p_{.A}$). Only one p_{ga} probability needs to be estimated, as the other three probabilities' estimators follow from constraints imposed by sample marginal frequencies, illustrated in Table 1.1.

Equation 1.6 provides

$$p^*_{0A} \quad = \quad \frac{X_{0A}}{2n} + \frac{\tilde{p}_{01,AE} p^*_{0A} p^*_{1E}}{2(p^*_{0A} p^*_{1E} + p^*_{0E} p^*_{1A})}. \tag{1.10}$$

Critically, we can express p^*_{0E} as $\tilde{p}_{0.} - p^*_{0A}$, p^*_{1A} as $\tilde{p}_{.A} - p^*_{0A}$, and p^*_{1E} as $1 - \tilde{p}_{0.} - \tilde{p}_{.A} + p^*_{0A}$, so that Equation 1.10 is a polynomial equation of a single variable, p^*_{0A}:

$$p^*_{0A} \quad = \quad \frac{X_{0A}}{2n} + \frac{\tilde{p}_{01,AE} p^*_{0A}(1 - \tilde{p}_{0.} - \tilde{p}_{.A} + p^*_{0A})}{2\Big[(p^*_{0A})(1 - \tilde{p}_{0.} - \tilde{p}_{.A} + p^*_{0A}) + (\tilde{p}_{0.} - p^*_{0A})(\tilde{p}_{.A} - p^*_{0A})\Big]}. \tag{1.11}$$

Table 1.1: Two-ancestry ga-allele probabilities corresponding to a stationary point $\mathbf{p}^*$, for allele $g \in \{0, 1\}$ and ancestry $a \in \{A, E\}$. The frequency p^*_{0A} is constrained by sample marginal frequencies $\tilde{p}_{0.}$ and $\tilde{p}_{.A}$, such that $max(0, \tilde{p}_{0.} + \tilde{p}_{.A} - 1) \leq p^*_{0A} \leq min(\tilde{p}_{0.}, \tilde{p}_{.A})$, where the bounds on p^*_{0A} arise by considering that $p^*_{ga} \in [0, 1]$ for all $ga \in \{0A, 0E, 1A, 1E\}$ and writing p^*_{0E}, p^*_{1A}, and p^*_{1E} in terms of sample marginal frequencies and p^*_{0A}.

	Ancestry		
Allele	A	E	Total
0	p^*_{0A}	$p^*_{0E} = \tilde{p}_{0.} - p^*_{0A}$	$\tilde{p}_{0.}$
1	$p^*_{1A} = \tilde{p}_{.A} - p^*_{0A}$	$p^*_{1E} = 1 - \tilde{p}_{0.} - \tilde{p}_{.A} + p^*_{0A}$	$\tilde{p}_{1.}$
Total	$\tilde{p}_{.A}$	$\tilde{p}_{.E}$	1

We follow Weir and Cockerham (1979) by breaking cases according to the positivity of counts. Weir and Cockerham (1979) considered cases where $n_{01,AE} = 0$, or only one or two X_{ga}'s are 0. We additionally consider cases where only three or all four X_{ga}'s are 0. In each case, we use analytical formulas to obtain the p_{0A} components p^*_{0A}'s of stationary points $\mathbf{p}^*$'s. If there are multiple stationary points, we evaluate the log likelihood function at each point, and take a point with the highest log likelihood value to be an MLE.

When $n_{01,AE} = 0$, direct counting provides $\hat{p}_{0A} = X_{0A}/(2n) = \tilde{p}_{00,AA} + (\tilde{p}_{00,AE} + \tilde{p}_{01,AA})/2$ as well as $\hat{p}_{0E} = X_{0E}/(2n), \hat{p}_{1A} = X_{1A}/(2n), \hat{p}_{1E} = X_{1E}/(2n)$. Incidentially, in Section 1.6.2 of the Supplement, we prove that when the $n_{gg',aa'}$ counts for categories with heterozygous genotype and differing ancestries in the ancestry pair are zero, the negative log likelihood function -$l(\mathbf{p})$ is a convex function of $\mathbf{p}$, so a stationary point of $l(\mathbf{p})$ must be an MLE.

There are four cases where only a single X_{ga} is 0 (i.e. only $X_{0A} = 0$, only $X_{0E} = 0$, only $X_{1A} = 0$, or only $X_{1E} = 0$). Such cases are discussed by Weir and Cockerham (1979). For example, if there are no individuals seen that are known to have allele 0 with ancestry A (i.e. only $X_{0A} = 0$ while other X_{ga}'s are positive), then from Equation 1.11, either $\hat{p}_{0A} = 0$ or

$$1 = \frac{\tilde{p}_{01,AE}(1 - \tilde{p}_{0.} - \tilde{p}_{.A} + p^*_{0A})}{2[p^*_{0A}(1 - \tilde{p}_{0.} - \tilde{p}_{.A} + p^*_{0A}) + (\tilde{p}_{0.} - p^*_{0A})(\tilde{p}_{.A} - p^*_{0A})]}, \tag{1.12}$$

which, after some rearrangement, is a quadratic equation in p^*_{0A}, analogous to the quadratic that Weir and Cockerham (1979) gave on page 109 to the right of "$X_{11} = 0$":

$$(p^*_{0A})^2 + (W - Y)(p^*_{0A}) + \Big[Z + Y(\hat{p}_{0.} + \hat{p}_{.A} - 1)\Big] = 0,$$

where

$$W = (1 - 2\hat{p}_{0.} - 2\hat{p}_{.A})/2$$
$$Y = n_{0/1,A/E}/(4n)$$
$$Z = \hat{p}_{0.}\hat{p}_{.A}/2,$$

whose roots can be obtained with the quadratic formula $p^*_{0A} = (-b + / - \sqrt{b^2 - 4ac})/(2a)$, where $a = 1$, $b = W - Y$, and $c = \Big[Z + Y(\hat{p}_{0.} + \hat{p}_{.A} - 1)\Big]$. (Note that the denominator in the right-hand side of Equation 1.12 cannot be 0, because we only consider points $\mathbf{p}^*$ with all component p^*_{ga}'s other than p^*_{0A} positive; since all X_{ga}'s besides X_{0A} are positive by assumption, an MLE $\hat{\mathbf{p}}$ cannot have any of those component $\hat{p}_{ga}$'s be 0, so we do not consider points $\mathbf{p}^*$ with any of those component p^*_{ga}'s being 0.) If this quadratic equation has no real roots, then $\hat{p}_{0A} = 0$ is the MLE. Otherwise, the real roots need to be checked for being in the range $[\max(\tilde{p}_{0.} + \tilde{p}_{.A} - 1,\, 0),\, \min(\hat{p}_{0.},\, \hat{p}_{.A})]$, and then the likelihood function's values at $p^*_{0A} = 0$ and each within-bounds real root are compared to determine an MLE.

There are four situations where only two X_{ga}'s sharing one subscript are zero, and two situations where only two X_{ga}'s sharing no subscripts are zero. Such cases are discussed by Weir and Cockerham (1979). First, if $X_{0A} = X_{0E} = 0$ for example, there are three stationary values: $p^*_{0A} = 0, \tilde{p}_{0.}$ and $\tilde{p}_{.A} - (1 - \tilde{p}_{0.})/2$. Weir and Cockerham (1979) gave these expressions on page 110 to the right of "$X_{11} = X_{12} = 0$." The three likelihoods need to be compared to determine an MLE. Second, if $X_{0A} = X_{1E} = 0$ for example, $\tilde{p}_{0.} + \tilde{p}_{.A} = 1$ and $\hat{p}_{0A} = 0$ is the MLE.

There are four situations of only three X_{ga}'s being zero. For example, $X_{0A} = X_{1E} = X_{1A} = 0$ so there are only two observed types in the data: $01, AE$ and $00, EE$, and $\tilde{p}_{0.} + \tilde{p}_{.A} = 1$. The stationary points are $p^*_{0A} = 0$, $\tilde{p}_{.A}$, and $\tilde{p}_{0.}/2$, and their log likelihood values are compared to find an MLE.

In the degenerate case of all four X_{ga}'s being zero, there is no unique estimate for p_{0A}.

Finally, there is the case where no count is 0, in which case Equation 1.11, expressed in terms of only p^*_{0A}, is a cubic equation in p^*_{0A}: $a(p^*_{0A})^3 + b(p^*_{0A})^2 + c(p^*_{0A}) + d = 0$, where we define $a := 4n$, $b := (2n)\Big[1 - 2(\hat{p}_{0.} + \hat{p}_{.A})\Big]$ - $\Big[2X_{0A} + n_{01,AE}\Big]$, $c := (2n)(\hat{p}_{.A})(\hat{p}_{0.})$ - $(X_{0A})(1 - 2\hat{p}_{0.} - 2\hat{p}_{.A})$ - $n_{01,AE}(1 - \hat{p}_0 - \hat{p}_{.A})$, and $d := -(X_{0A})(\hat{p}_{.A})(\hat{p}_{0.})$. We implement an analytical method for finding its roots (Gaunt et al., 2007). The number of unique roots depends on the positivity of a discriminant.

We provide an information-based approach for obtaining a standard error (SE) estimator for point estimator $T = \hat{p}(1|A)$ or $\hat{p}(1|E)$. The SE estimator is based on the observed Fisher information (Efron and Hinkley, 1978), and relies on standard asymptotic theory that assumes the true parameter lies in the interior of the parameter space. See Section 1.6.3 for derivations of information-based SE formulas.

We implemented this two-ancestry approach on simulated data, as described in Section 1.3.4 with results described in Section 1.4.2. We implemented this two-ancestry approach on real data, as described in Section 1.3.7 with results described in Section 1.4.4.

1.3.3 Methods: Three Ancestries.

Taking advantage of constraints $\sum_{a=1}^{K} p^*_{ga} = \tilde{p}_{g.}$ and $\sum_{g=0}^{1} p^*_{ga} = \tilde{p}_{.a}$ imposed on p^*_{ga} stationary point components by marginal frequency estimates and derived from score equations (given by considering Equation 1.5 for all g and a), and illustrated in Table 1.2, we can express the log likelihood function evaluated at a stationary point $l(\mathbf{p}^* = [p^*_{0A}, p^*_{0E}, p^*_{0N}, p^*_{1A}, p^*_{1E}, p^*_{1N}])$ in terms of only two free p^*_{ga} components, which we take to be p^*_{0A} and p^*_{0E}.

Table 1.2: Three-ancestry ga-allele probabilities corresponding to a stationary point $\mathbf{p}^*$, for allele $g \in \{0,1\}$ and ancestry $a \in \{A, E, N\}$. The frequencies p^*_{0A} and p^*_{0E} are constrained by sample marginal frequencies, such that $max(0, \tilde{p}_{0.}+\tilde{p}_{.A}+\tilde{p}_{.E}-1-p^*_{0E}) \leq p^*_{0A} \leq min(\tilde{p}_{.A}, \tilde{p}_{0.}-p^*_{0E})$ and $max(0, \tilde{p}_{0.}+\tilde{p}_{.A}+\tilde{p}_{.E}-1-p^*_{0A}) \leq p^*_{0E} \leq min(\tilde{p}_{.E}, \tilde{p}_{0.}-p^*_{0A})$.

	Ancestry			
Allele	A	E	N	Total
0	p^*_{0A}	p^*_{0E}	$p^*_{0N} = \tilde{p}_{0.} - p^*_{0A} - p^*_{0E}$	$\tilde{p}_{0.}$
1	$p^*_{1A} = \tilde{p}_{.A} - p^*_{0A}$	$p^*_{1E} = \tilde{p}_{.E} - p^*_{0E}$	$p^*_{1N} = 1 - \tilde{p}_{0.} - \tilde{p}_{.A} - \tilde{p}_{.E} + p^*_{0A} + p^*_{0E}$	$\tilde{p}_{1.}$
Total	$\tilde{p}_{.A}$	$\tilde{p}_{.E}$	$\tilde{p}_{.N} = 1 - \tilde{p}_{.A} - \tilde{p}_{.E}$	1

When $n_{01,AE} = n_{01,NA} = n_{01,NE} = 0$ (Case 1), we obtain an MLE $\hat{\mathbf{p}}$ analytically as $\hat{p}_{ga} = X_{ga}/(2n)$ for all g and a. Otherwise (Case 2), we perform a two-dimensional grid search of the log likelihood function $l(p^*_{0A}, p^*_{0E})$. Unlike with the two-ancestry approach, we do not analytically consider all boundary solutions, before considering grid search. Rather, the grid search can uncover boundary solutions.

1.3.3.1 Case 1: $n_{01,AE} = n_{01,NA} = n_{01,NE} = 0$.

Under this condition, the log likelihood function reduces to that of a multinomial distribution with category counts X_{ga}'s that sum to $2n$ and category probabilities p_{ga}'s. For such a distribution, each MLE $\hat{\mathbf{p}}$ component $\hat{p}_{ga}$ is $X_{ga}/(2n)$.

We can alternatively derive $\hat{p}_{ga} = X_{ga}/(2n)$ by solving the system of equations obtained by setting partial derivatives of the log likelihood function plus Langrangian term $\lambda(\sum_{g=0}^{1}\sum_{a=1}^{K} p_{ga} - 1)$ to 0 to find that $\hat{p}_{0A} = X_{0A}/(2n)$ and $\hat{p}_{0E} = X_{0E}/(2n)$, and then solve for $\hat{p}_{0N} = \hat{p}_{0.} - \hat{p}_{0A} - \hat{p}_{0E}$, $\hat{p}_{1A} = \hat{p}_{.A} - \hat{p}_{0A}$, $\hat{p}_{1E} = \hat{p}_{.E} - \hat{p}_{0E}$, and $\hat{p}_{1N} = 1 - \hat{p}_{0.} - \hat{p}_{.A} - \hat{p}_{.E} + \hat{p}_{0A} + \hat{p}_{0E}$, where recall marginal frequency MLEs are observed sample frequencies ($\hat{p}_{g.} = \tilde{p}_{g.}$ and $\hat{p}_{.a} = \tilde{p}_{.a}$ for all g and a).

The result derived in Section 1.6.2 of the Supplement applies to this case: Because all counts with heterozygous genotype and differing ancestries in the ancestry pair are zero, the negative log likelihood function $-l(\mathbf{p})$ is a convex function, so a stationary point of $l(\mathbf{p})$ must be an MLE.

1.3.3.2 Case 2: If not Case 1, perform a 2-D grid search of the log likelihood function

The grid search is as follows:

1. Consider a grid of points $[p_{0A}, p_{0E}]$, where p_{0A} ranges from 0 to $min(\hat{p}_{0.}, \hat{p}_{.A})$ inclusive, by a user-specified increment of ϵ (default value 0.01), and p_{0E} ranges from 0 to $min(\hat{p}_{0.}, \hat{p}_{.E})$ inclusive, by an increment of ϵ. Inclusivity is important to search for a boundary MLE. The log likelihood function evaluates to $-\infty$ at grid points outside its domain. It evaluates to a finite number, only for grid points $[p_{0A}, p_{0E}]$ that lie within constraints derived by expressing p_{0N}, p_{1A}, p_{1E}, and p_{1N} in terms of p_{0A}, p_{0E} and sample marginal frequencies $\{\tilde{p}_{g.}\}$ and $\{\tilde{p}_{.a}\}$ so that $\sum_g p_{ga} = \tilde{p}_{.a}$ and $\sum_a p_{ga} = \tilde{p}_{g.}$, and considering $0 \leq p_{ga} \leq 1$:

$$max(0, \tilde{p}_{0.} + \tilde{p}_{.A} + \tilde{p}_{.E} - 1 - p_{0E}) \leq p_{0A} \leq min(\tilde{p}_{0.} - p_{0E}, \tilde{p}_{.A})$$
$$max(0, \tilde{p}_{0.} + \tilde{p}_{.A} + \tilde{p}_{.E} - 1 - p_{0A}) \leq p_{0E} \leq min(\tilde{p}_{0.} - p_{0A}, \tilde{p}_{.E})$$

2. Determine which of the within-constraints grid point(s) yield(s) the largest log likelihood value.

3. If there is a unique $[p_{0A}, p_{0E}]$ maximizer from Step 2, report the maximizer.

4. If there are multiple maximizers of the log likelihood function from Step 3, do a second, denser grid search on a subgrid of $[p_{0A}, p_{0E}] \in [0, 1] \times [0, 1]$, bordering the collection of maximizers by the user-specified ϵ grid increment used in the first grid search. More specifically, the subgrid points $[p_{0A}, p_{0E}]$ are constrained as follows:

$$p_{0A} \geq max(min(\{p_{0A} \text{ coordinate values of maximizers}$$
$$\text{from the first } \epsilon \text{ grid search }\}) - \epsilon, 0)$$
$$p_{0A} \leq min(max(\{p_{0A} \text{ coordinate values of maximizers}$$
$$\text{from the first } \epsilon \text{ grid search }\}) + \epsilon, 1)$$
$$p_{0E} \geq max(min(\{p_{0E} \text{ coordinate values of maximizers}$$
$$\text{from the first } \epsilon \text{ grid search }\}) - \epsilon, 0)$$
$$p_{0E} \leq min(max(\{p_{0E} \text{ coordinate values of maximizers}$$
$$\text{from the first } \epsilon \text{ grid search }\}) + \epsilon, 1)$$

This second grid search is denser in that we use an increment that is 10 times finer than the original increment, i.e. using an increment of $(\epsilon/10)$.

5. Round maximizer(s) from the subgrid search to the nearest ϵ and report the unique, rounded values.

1.3.4 Methods: Applying the Two-Ancestry Approach to Simulated Data

Under each of six simulation settings, we simulated ga-alleles for $1,000$ admixed individuals at $1,000$ SNPs. Each simulation setting was defined by a value for $\mathbf{p} = [p_{0A}, p_{0E}, p_{1A}, p_{1E}]$, and corresponding values for conditional probabilities $p(1|A)$ and $p(1|E)$, as given in Table 1.3. These settings were chosen to be realistic in that they were designed to encourage the applicability of our two-ancestry method cases (e.g. only $n_{01,AE} = 0$, only one zero X_{ga}, or only two zero X_{ga}s) that applied to real data. We drew $2,000$ ga-alleles from $\{0A, 0E, 1A, 1E\}$ independently with probabilities $[p_{0A}, p_{0E}, p_{1A}, p_{1E}]$. We considered each pair of consecutively drawn ga-alleles as belonging to one diploid individual. Two ga-alleles then determined the individual's genotype $gg' \in \{0/0, 0/1, 1/1\}$ and ancestry pair $aa' \in \{A/A, A/E, E/E\}$. We then counted the number of individuals with each genotype and ancestry pair combination, determining numbers $\{n_{gg',aa'}\}$ of individuals, which constituted the data for one SNP. If a simulated SNP's data did not have both alleles 0 and 1 and both ancestries A and E present, or if a simulated SNP's data resulted in multiple MLEs, we did not keep that data and re-simulated the SNP's data.

Simulation Setting	p_{0A}	p_{0E}	p_{1A}	p_{1E}	$p(1\|$ A)	$p(1\|$ E)
1	0.00	0.01	0.01	0.98	1.00	0.99
2	0.00	0.40	0.40	0.20	1.00	0.33
3	0.00	0.01	0.50	0.49	1.00	0.98
4	0.00	0.01	0.99	0.00	1.00	0.00
5	0.70	0.25	0.01	0.04	0.01	0.14
6	0.60	0.25	0.10	0.05	0.14	0.17

Table 1.3: Extended ASAFE: Two-Ancestry Simulation Settings.

For each SNP, we applied the two-ancestry approach to obtain ancestry-specific allele frequency estimates $\hat{p}(1|A)$ and $\hat{p}(1|E)$. For each simulation setting, we considered the true ancestry-specific allele frequencies to be the frequencies $[p(1 \mid \text{Ancestry } A), p(1 \mid \text{Ancestry } E)]$ corresponding to p_{ga} values used to generate ga-alleles. We calculated error as the estimated frequency $\hat{p}(1|a)$ minus the true frequency $p(1|a)$, for each ancestry $a \in S_a = \{A, E\}$. For each ancestry, we calculated the mean of $1,000$ SNPs' errors and the standard deviation of these errors. Results are discussed in Section 1.4.2.

1.3.5 Methods: Assessing Two-Ancestry SE Estimators on Simulated Data

We assessed observed Fisher information-based two-ancestry estimators of standard errors $SD(\hat{P}(1|A))$ and $SD(\hat{P}(1|E))$ of Section 1.6.3 with simulated data from Section 1.3.4.

For each of six simulation settings given in Table 1.3 and each MLE $\hat{p} \in \{\hat{p}_{0A}, \hat{p}_{0E}, \hat{p}_{1A}, \hat{p}_{1E}, \hat{p}(1|A), \hat{p}(1|E)\}$, we made a histogram comparing $1,000$ standard error estimates to the empirical standard error, the sample standard deviation of $1,000$ MLE point estimates. We denote standard error estimators as $\widehat{SD}$. For example, $\widehat{SD}(\hat{p}_{0A})$ is the estimated standard error of MLE $\hat{p}_{0A}$.

We did not expect that our standard error estimators would perform well in simulation settings 1-4 because the true parameter $\mathbf{p} = [p_{0A}, p_{0E}, p_{1A}, p_{1E}]$ was on the boundary of the parameter space, while our estimators were derived assuming the parameter was in the interior. Nevertheless, we assessed our standard error estimators in these settings to see if results would confirm our expectation. It was possible that our expectation might not be confirmed, since Wellner wrote on page 22 of a situation in which, despite a parameter taking a boundary value (contrary to the assumptions of standard theory), standard theoretical results still applied.

1.3.6 Methods: Applying the Three-Ancestry Approach to Simulated Data

We simulated ancestry pairs for 250 admixed individuals at 1,000 SNPs as follows. For every individual i's allele copy j at a SNP, the allele copy's ancestry $a_{ij} \in \{1, ..., K\}$ was drawn from a distribution with equal mass on each ancestry: $a_{ij} \sim_{iid} Multinomial(n = 1, [p_1, ..., p_K] = [1/K, ..., 1/K])$. Given a drawn ancestry a_{ij}, the allele copy's allelic type $g_{ij} \in \{0, 1\}$ was drawn from a Bernoulli distribution with ancestry-specific allele frequency $p(g_{ij} = 1|a_{ij})$ of being allelic type 1: $(g_{ij}|a_{ij}) \sim Bernoulli(p(g_{ij} = 1|a_{ij}))$.

For each SNP, we applied the three-ancestry approach with default $\epsilon = 0.01$ to obtain ancestry-specific allele frequency estimates $\hat{p}(1|A)$, $\hat{p}(1|E)$, and $\hat{p}(1|N)$. We considered the true ancestry-specific allele frequencies to be frequencies [$p(1$ | Ancestry $A)$, $p(1$ | Ancestry $E)$, $p(1$ | Ancestry $N)$] used to generate alleles from ancestries. We calculated error as the estimated frequency $\hat{p}(1|a)$ minus the true frequency $p(1|a)$, for each ancestry $a \in S_a = \{A, E, N\}$. For each ancestry, we calculated the mean of 1,000 SNPs' errors and the standard deviation of these errors. Results are discussed in Section 1.4.3.

1.3.7 Methods: Applying the Two-Ancestry Approach to 1000 Genomes ASW Data

African Americans are an admixed population, descended from Africans and Europeans. We took genotypes and ancestry calls for 61 African American (ASW) individuals in the 1000 Genomes Phase 1 data.

We obtained Chr 22 marker genotypes from the 1000 Genomes FTP site, with URL given in Section 1.6.1. We obtained ancestry calls for ASW individuals from the 1000 Genomes FTP site, with URL given in Section 1.6.1.

We filtered the vcf file containing genotypes as follows: We removed rows that had a reference allele or alternative allele that was not A, T, C, or G. We then removed rows with the same bp position, since such rows could correspond to an incorrectly mapped marker, and removed rows with the same rsID, since such rows could correspond to a SNP with more than two alleles. The remaining markers were bi-allelic SNPs less likely to have mapping errors. Of these SNPs, we considered only the ones that also had ancestry pairs called, and that had both alleles 0 and 1 as well as ancestries A and E present in at least one copy in the sample. We applied the two-ancestry approach to estimate African-specific and European-specific allele frequencies, $p(1|A)$ and $p(1|E)$, for these remaining 141,318 bi-allelic SNPs. In Section 1.4.4, we will describe this application using multiple SNPs as illustrative examples.

1.3.8 Methods: Applying the Three-Ancestry Approach to 1000 Genomes PUR Data

Puerto Ricans are an admixed population, descended from Africans, Europeans, and Native Americans. We considered genotypes and ancestry calls for 55 Puerto Rican (PUR) individuals in the 1000 Genomes Phase 1 data.

We obtained Chr 22 SNP genotypes from the 1000 Genomes FTP site, with URL given in Section 1.6.1. The genotype file contained 494,328 SNPs for 1,099 individuals, including 55 PUR individuals. We obtained ancestry calls for 55 PUR individuals from the 1000 Genomes FTP site, with URL given in Section 1.6.1. We determined each PUR individual's ancestry pair at the 494,328 SNPs in the vcf file.

Each individual had missing ancestry pairs for certain SNPs. After excluding SNPs that were missing an ancestry pair in any individual, 94,731 SNPs with non-missing ancestry pairs remained. To reduce computation time, we randomly sampled 1,000 of these 94,731 SNPs. We then applied the three-ancestry method with default grid interval $\epsilon = 0.01$ to non-missing genotypes and ancestry pairs for these 1,000 SNPs, and estimated African-specific, European-specific, and Native American-specific allele frequencies, $p(1|A)$, $p(1|E)$, and $p(1|N)$. In Section 1.4.5, we will describe this application using multiple SNPs as illustrative examples.

1.3.9 Methods: Comparing the Three-Ancestry Approach to the ASAFE EM Algorithm with 1000 Genomes PUR Data

Recall from Section 1.3.3 that our novel three-ancestry approach involves an analytical solution (Case 1) for the unique MLE $\hat{\mathbf{p}}$ when Assumption 1 about counts $\{n_{gg',aa'}\}$ is satisfied, and a grid search (Case 2) with a user-specified grid interval (default 0.01) otherwise. Assumption 1 said that no individuals have genotype and ancestry pair combination with heterozygous genotype and differing ancestries, i.e. every observed category gg', aa' with heterozygous genotype gg' and ancestry pair aa' with differing ancestries has count $n_{gg',aa'} = 0$. In Section 1.6.2 of the Appendix, we proved that this assumption is sufficient for the negative log likelihood function to be convex, so that a stationary point is an MLE.

The ASAFE EM algorithm implementation applies to only three ancestries, yields a stationary point $\mathbf{p}^*$, starts at a user-specified point (default $\mathbf{p} = [1/6, 1/6, 1/6, 1/6, 1/6, 1/6]$), and stops iterating when two estimates $\mathbf{p}_k$ and $\mathbf{p}_{k+1}$ are within a user-specified ϵ (default 10^{-8}) of each other, or when a user-specified number of iterations (default 1000) has passed, whichever comes first. The algorithm avoids divide by zero error by incrementing (or decrementing) each k-th iterate's $p_{k,ga}$ from the M step away from a boundary value of 0 (or 1) by a user-specified ϵ (default 10^{-8}).

How would our novel three-ancestry approach's estimate differ from or be similar to the ASAFE EM algorithm implementation's estimate, and under what conditions would we expect these differences or similarities? When the log likelihood function is multimodal, the ASAFE EM algorithm's estimate can differ from our novel three-ancestry approach's estimate, because an EM algorithm with an unlucky starting value can yield a stationary point that is not an MLE, while our three-ancestry approach's grid search with a sufficiently small grid interval should yield an MLE. When the log likelihood function is unimodal, a stationary point is an MLE, so we would expect the ASAFE EM algorithm's estimate to be close to our novel three-ancestry approach's estimate, regardless of starting value. When an MLE is on the boundary of the parameter space, i.e. has a 0 element, the ASAFE EM algorithm's estimate can differ from our novel three-ancestry approach's estimate in that the EM estimate will not involve exact zeroes, while our novel three-ancestry approach's estimate can have exact zeroes. Case 1 (analytical formula) of our novel three-ancestry approach can return exact zeroes and Case 2 (grid search) of our three-ancestry approach can if $\hat{p}_{0A}$ and $\hat{p}_{0E}$ take boundary values, 0 and $min(\hat{p}_{0.}, \hat{p}_{.A})$ for $\hat{p}_{0A}$, and 0 and $min(\hat{p}_{0.}, \hat{p}_{.E})$ for $\hat{p}_{0E}$. Case 2 of our novel three-ancestry approach could possibly not return an exact boundary estimate $\hat{\mathbf{p}} = [\hat{p}_{0A}, \hat{p}_{0E}, \hat{p}_{0N}, \hat{p}_{1A}, \hat{p}_{1E}, \hat{p}_{1N}]$ if $\hat{p}_{0N} = 0$ or $\hat{p}_{1N} = 0$, while $\hat{p}_{0A}$ and $\hat{p}_{0E}$ did not take boundary values (0 and $min(\hat{p}_{0.}, \hat{p}_{.A})$ for $\hat{p}_{0A}$, 0 and $min(\hat{p}_{0.}, \hat{p}_{.E})$ for $\hat{p}_{0E}$) and did not correspond to a non-boundary $[p_{0A},\, p_{0E}]$ grid point.

We applied the ASAFE EM algorithm implementation with default arguments to the same 1,000 SNPs that we applied our novel three-ancestry approach to in Section 1.3.8. For each $\hat{p}_{ga}, g \in \{0, 1\}, a \in \{A, E, N\}$ and $\hat{p}(1|a), a \in \{A, E, N\}$, we plotted $1,000$ three-ancestry approach estimates against $1,000$ ASAFE EM estimates.

We considered searching amongst these $1,000$ SNPs for a SNP with a multimodal log likelihood function, a SNP with a unimodal log likelihood function, and a SNP with a boundary MLE, to check for expected differences and similarities. To look for a SNP with a multimodal log likelihood function, we looked for a SNP with a novel three-ancestry approach estimate far from its ASAFE EM algorithm estimate. Our idea was to take such a SNP's $n_{gg',aa'}$ counts and observe a 3-D plot of the log likelihood function $l(p_{0A}, p_{0E})$ over a grid with $p_{0A} \in [0, min(\hat{p}_{0.}, \hat{p}_{.A})]$ and $p_{0E} \in [0, min(\hat{p}_{0}, \hat{p}_{.E})]$, looking for multiple modes. Considering

$p_{0N} = \hat{p}_{0.} - p_{0A} - p_{0E}$, $p_{1A} = \hat{p}_{.A} - p_{0A}$, $p_{1E} = \hat{p}_{.E} - p_{0E}$, and $p_{1N} = 1 - \hat{p}_{0.} - \hat{p}_{.A} - \hat{p}_{.E} + p_{0A} + p_{0E}$, these $[p_{0A}, p_{0E}]$ grid points would correspond to potential stationary points of the full log likelihood function $l(p_{0A}, p_{0E}, p_{0E}, p_{1A}, p_{1E}, p_{1N})$, those points satisfying $\sum_g p_{ga} = \hat{p}_{.a}$ and $\sum_a p_{ga} = \hat{p}_{g.}$. To look for a SNP with a unimodal log likelihood function, we examined these $1,000$ SNPs for any that met Assumption 1, which we proved was sufficient for a unimodal log likelihood function. To look for a SNP with a boundary MLE, we searched amongst the $1,000$ SNPs that met Assumption 1 for a SNP with a novel three-ancestry approach $\hat{\mathbf{p}}$ estimate with an exact zero element.

1.3.10 Methods: Using Allele Frequency Estimates for Population Structure

As described in Sections 1.3.8 and 1.4.5, at 1,000 SNPs randomly selected from 94,731 Chr 22 SNPs with non-missing genotypes and ancestry pairs in 55 PUR individuals, we estimated African, European, and Native American-specific allele frequencies, i.e. $p(1|A)$, $p(1|E)$, and $p(1|N)$. We used these frequency estimates to calculate method of moments estimators (MMEs) for population-specific F_{ST}'s (also known as β_i's) for African, European, and Native American ancestries, using formulas from Buckleton et al. (2016) and Weir and Goudet (2017), which we now describe.

Buckleton et al. (2016) defined the between-population i and j, locus-l-specific match statistic as

$$\tilde{M}_{ijl} = \sum_u \tilde{p}_{ilu} \tilde{p}_{jlu} = (\tilde{p}_{il})(\tilde{p}_{jl}) + (1 - \tilde{p}_{il})(1 - \tilde{p}_{jl}),$$

where the sample proportion of allele-u at locus-l in a sample from population-i is $\tilde{p}_{ilu}$ and in a sample from population j is $\tilde{p}_{jlu}$. The latter equality holds if we consider a bi-allelic SNP, and let $\tilde{p}_{il}$ be the population-i, locus-l sample proportion of one allele and $\tilde{p}_{jl}$ be the population-j, locus-l sample proportion of that allele.

The between-population, locus-l-specific average match statistic is

$$\tilde{M}_{Bl} = \frac{\sum_{i=1}^{r_l}\sum_{j=1,j\neq i}^{r_l}\tilde{M}_{ijl}}{(r_l)(r_l-1)},$$

where locus l is measured in r_l populations, and i and j index populations. In words, the match statistic $\tilde{M}_{Bl}$ is the mean of $\tilde{M}_{ijl}$'s, over pairs of distinct populations i and j. Section A.1. of Appendix A of Buckleton et al. (2016) gives the between-population, locus-l match statistic as $\tilde{M}_{Bl} = \left[\sum_{i=1}^{r_l}\sum_{j=1}^{r_l,j\neq i}\tilde{M}_{ijl}\right]/\left[(r_l)(1-r_l)\right]$, but the $(1-r_l)$ is a typo, and should be (r_l-1), so that the denominator is $(r_l^2-r_l)$.

The between-population match statistic is

$$\tilde{M}_B = \sum_{l=1}^{L}\frac{\tilde{M}_{Bl}}{L}.$$

In words, the match statistic $\tilde{M}_B$ is the mean over L loci of between-population, locus-l-specific match statistics $\tilde{M}_{Bl}$'s.

The population-i, locus-l-specific match statistic is

$$\tilde{M}_{il} \approx \sum_u \tilde{p}_{ilu}^2 = (\tilde{p}_{il}^2) + (1-\tilde{p}_{il})^2,$$

where u indexes alleles at locus l, and the latter equality holds for a bi-allelic SNP, with $\tilde{p}_{il}$ being the sample proportion of one of the alleles in population i at locus l.

The population-i-specific match statistic is

$$\tilde{M}_i = \frac{\sum_{l=1}^{L}\tilde{M}_{il}}{L}.$$

In words, the match statistic $\tilde{M}_i$ is the mean over L loci of population-i, locus-l-specific match statistic $\tilde{M}_{il}$'s.

The methods of moments estimator (MME) of population-i-specific F_{ST} β_i is

$$\hat{\beta}_i := \frac{\tilde{M}_i - \tilde{M}_B}{1-\tilde{M}_B} = \frac{\sum_{l=1}^{L}(\tilde{M}_{il}-\tilde{M}_{Bl})}{\sum_{l=1}^{L}(1-\tilde{M}_{Bl})}.$$

We estimated the variance of $\hat{\beta}_i$ by bootstrapping over loci, as discussed in Weir (1996) and done previously by Weir and Goudet (2017). The parameter β_i reflects variation among evolutionary replicates of sampled populations, and this variation is captured by variation among loci. Further discussion of the rationale for bootstrapping over loci is given in Section 1.6.4 of the supplement.

More specifically, the following procedure was used to calculate bootstrap confidence intervals for each $\hat{\beta}_i$ MME:

Step (0) We considered three-ancestry estimates $[\hat{p}(1|A), \hat{p}(1|E), \hat{p}(1|N)]$ for each of 1,000 loci.

Step (1) We sampled $1,000$ of these 3-vector estimates with replacement, obtaining a bootstrap sample of 1,000 3-vector estimates.

Step (2) With this bootstrap sample, we calculated $\hat{\beta}_{African}$, $\hat{\beta}_{European}$, and $\hat{\beta}_{NativeAmerican}$ MMEs, which we call bootstrap MMEs.

We performed Steps (0)-(2) $B = 7,000$ times. We then calculated a 95% bootstrap percentile interval for each β_i by taking the 2.5% quantile and the 97.5% quantile of 7,000 $\hat{\beta}_i$ bootstrap MMEs and rounding the quantiles to the nearest 0.01 (Wasserman, 2006; Wakefield, 2013). Each quantile was calculated using the quantile() function in R with argument type = 1 for the inverse of the empirical cdf.

Repeating this bootstrap procedure a second time, we obtained confidence intervals that were the same as the intervals obtained from the first bootstrap procedure, indicating that B was large enough to get a representative sample from the empirical distribution of $\hat{\beta}_i$ MMEs. The value $B = 7,000$ was reached by starting with $B = 1,000$, seeing if two bootstrap confidence intervals had corresponding endpoints that were the same to the nearest 0.01, and increasing B until the corresponding endpoints were the same.

Results are given in Section 1.4.7.

1.4 Results

1.4.1 Results: Two-Ancestry Approach Applied to Simulated Data

Table 1.4 shows error summary statistics of applying the two-ancestry approach to simulated data, as discussed in Section 1.3.4. The largest (across $[p(1|A), p(1|E)]$ allele frequency combinations, and ancestry-specific allele frequencies $p(1|A)$ and $p(1|E)$) absolute value of mean of errors was 0.0081. The largest standard deviation of errors was 0.0195.

Table 1.4: Mean of Errors (i.e. Bias of Estimator) and SD of Errors (i.e. SD of Estimator), where Error = Estimated Ancestry-Specific Allele Frequency - True Ancestry-Specific Allele Frequency, for 1,000 SNPs per Combination of Ancestral Population Frequencies $p(1|A)$ and $p(1|E)$. The frequencies in the table are the ancestral population frequencies used to generate the data, and were considered the true ancestry-specific allele frequencies used to calculate errors. The mean of errors in estimating ancestry-a specific allele frequency is Mean Error $p(1|a), a \in \{A, E\}$. The standard deviation of errors in estimating ancestry-a specific allele frequency is SD Error $p(1|a), a \in \{A, E\}$. A pictoral display of estimates and true parameter values is given in Figure 1.2.

p(1 | A)	p(1 | E)	Mean Error p(1 | A)	SD Error p(1 | A)	Mean Error p(1 | E)	SD Error p(1 | E)
1.00	0.99	-8.07×10^{-3}	1.87×10^{-2}	1.62×10^{-4}	2.24×10^{-3}
1.00	0.33	0	0	5.90×10^{-4}	1.35×10^{-2}
1.00	0.98	0	0	-9.40×10^{-5}	4.34×10^{-3}
1.00	0.00	0	0	0	0
0.01	0.14	-4.91×10^{-5}	3.72×10^{-3}	5.74×10^{-4}	1.44×10^{-2}
0.14	0.17	-1.59×10^{-4}	1.08×10^{-2}	4.18×10^{-4}	1.95×10^{-2}

Table 1.5 counts two-ancestry approach cases used per simulation setting.

Figure 1.2 gives violin plots of ancestry-specific allele frequency $[p(1|A), p(1|E)]$ estimates (black) with ancestry-specific allele frequencies used to generate alleles given ancestry draws (red). Estimates have a mean close to the ancestry-specific allele frequency used to generate

Table 1.5: Number of 1,000 SNPs per Combination of Ancestral Population Frequencies $p(1|A)$ and $p(1|E)$ for which Each Two-Ancestry Approach Case was Used. Ancestral population frequencies used to generate data are p(1 | A) and p(1 | E), and define a simulation setting. The fourth to the last rows correspond to two-ancestry approach cases. An entry corresponding to a case and a simulation setting gives the number of $1,000$ SNPs simulated under that setting for which that case was used to obtain an MLE $\hat{\mathbf{p}}$. Cases that were not used in any simulation setting are excluded from the table.

Simulation Setting	1	2	3	4	5	6
p(1 | A)	1.00	1.00	1.00	1.00	0.01	0.14
p(1 | E)	0.99	0.33	0.98	0.00	0.14	0.17
$n_{01,AE} = 0$	819	0	1	0	0	0
$X_{0A} = X_{0E} = X_{1E} = 0$	0	0	0	820	0	0
$X_{0A} = X_{1E} = 0$	0	0	0	180	0	0
$X_{0A} = 0$	181	1000	999	0	0	0
Cubic Discriminant Positive	0	0	0	0	0	1000
Cubic Discriminant Negative	0	0	0	0	1000	0

alleles given ancestry draws.

Table 1.6 illustrates that of 6,000 simulated SNPs (1,000 SNPs for each of six simulation settings), 2,000 SNPs met the condition that for all g and a, $X_{ga} > 0$. Of these 2,000 SNPs that met the condition, all had the MLE $\hat{\mathbf{p}}$ in the interior of the parameter space. Hence, our estimates are consistent with our earlier claim, proved in Section 1.3.1.2, that having every $X_{ga} > 0$ is sufficient for an MLE $\hat{\mathbf{p}}$ to be in the interior of the domain. Of 6,000 simulated SNPs, 2,181 SNPs had MLE $\hat{\mathbf{p}}$ in the interior of the domain. Of these 2,181 SNPs with an interior MLE $\hat{\mathbf{p}}$, 2,000 SNPs met the condition while 181 did not, indicating that the condition is not necessary for an MLE $\hat{\mathbf{p}}$ to be in the interior of the domain, i.e. it is not true that if the MLE is in the interior, then the condition is met.

Table 1.6: Co-Occurrence of Meeting the Condition that Every $X_{ga} > 0$, and the Two-Ancestry Approach Yielding an Interior MLE $\hat{\mathbf{p}}$, in Simulated Data. The lower row shows that whenever the condition is met, the estimate $\hat{\mathbf{p}}$ is in the interior, consistent with our claim in Section 1.3.1 that having every $X_{ga} > 0$ is sufficient for the MLE to be in the interior (i.e. the statement "condition holds $\implies$ interior estimate" is true). The right column indicates that it is possible for the estimate to be in the interior, without meeting the condition, (i.e. the converse statement, "interior estimate $\implies$ condition holds" is not true).

	$\hat{\mathbf{p}}$ Not In Interior	$\hat{\mathbf{p}}$ In Interior
Condition Not Met	3819	181
Condition Met	0	2000

A multi-modal likelihood with some peaks of equal height has multiple MLEs. In our simulation study, 50 SNPs in the fourth simulation setting had simulated data that resulted in multiple MLEs. Such SNPs' data were discarded and 50 other SNPs without multiple MLEs were simulated in their stead.

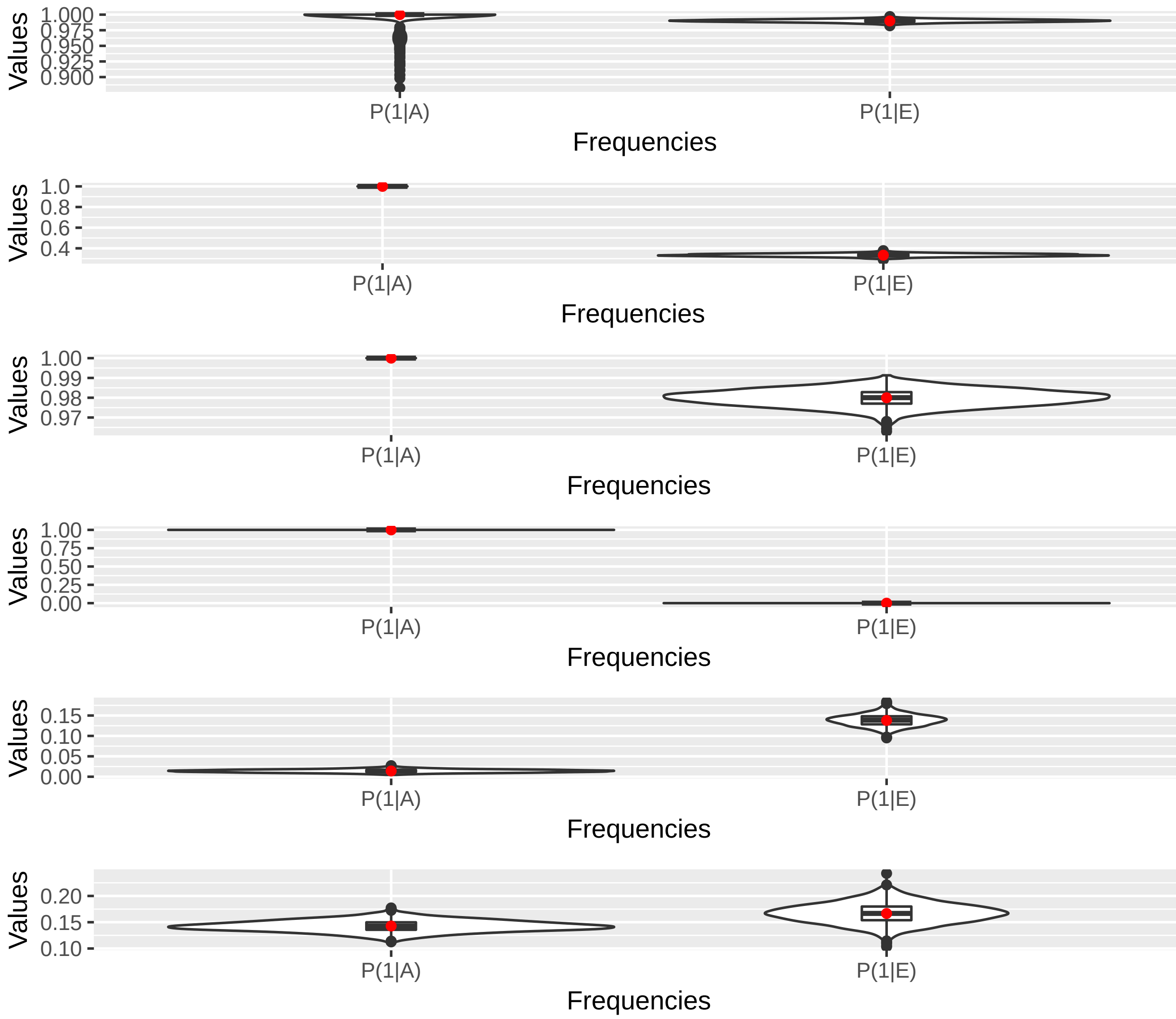

Figure 1.2: Violin plots of ancestry-specific allele frequency estimates (black) obtained by the two-ancestry approach, with true ancestry-specific allele frequency values used to generate data (red dot). Six simulation settings had allele frequency combinations $[p(1|A), p(1|E)]$, going from top to bottom: $[1.00, 0.99]$, $[1.00, 0.33]$, $[1.00, 0.98]$, $[1.00, 0.00]$, $[0.01, 0.14]$, and $[0.14, 0.17]$. For every simulation setting and ancestry-specific allele frequency $p(1|A)$ or $p(1|E)$, estimates (black) appear centered around the true parameter value (red dot), consistent with low mean error (i.e. bias) seen in Table 1.4.

1.4.2 Results: Assessing Two-Ancestry SE Estimators on Simulated Data

We had originally conducted the two-ancestry simulation study of Section 1.3.4 with $n = 100$ individuals. We found that 280 SNPs in simulation setting 3 had negative variance estimates, 2 SNPs in simulation setting 4 had NA variance estimates, and 23 SNPs in simulation setting 5 had negative variance estimates. All these SNPs had boundary MLEs. There were SNPs with boundary MLEs, and no NA or negative variance estimates. The two SNPs with NA variance estimates had $n_{11,AA} = 99$, $n_{00,EE} = 1$, and $X_{0A} = X_{0E} = X_{1A} = 0$. Zero counts resulted in 0 values for second derivatives $a = \partial^2 l/\partial^2 p_{0A}$, $b = \partial^2 l/\partial p_{0A}\partial p_{0E}$, $c = \partial^2 l/\partial^2 p_{0E}$, and $e = \partial^2 l/\partial p_{0E}\partial p_{1A}$, so that $1/(-adf + ae^2 + b^2 f - 2bce + c^2 d)$ in the inverse information matrix evaluated to NA, due to divide by 0 error. We increased the sample size to $n = 1,000$ individuals, and found no negative or NA SE estimates with that sample size.

Figures 1.3, 1.4, 1.5, 1.6, 1.7, and 1.8 display results of the assessment of two-ancestry SE estimators on simulated data with $n = 1,000$ individuals and $L = 1,000$ SNPs under six simulation settings described in Section 1.3.4. For each simulation setting, the Figures give black histograms of $1,000$ SNPs' standard error (SE) estimates with a red line for the sample standard deviation of $1,000$ SNPs' MLE point estimates. We see that in simulation settings 5 and 6 where the true parameter $\mathbf{p} = [p_{0A}, p_{0E}, p_{1A}, p_{1E}]$ is in the interior of the parameter space, black histograms are closer to centered at the red line value than in simulation settings 1-4, where the true parameter $\mathbf{p}$ is on the boundary of the parameter space.

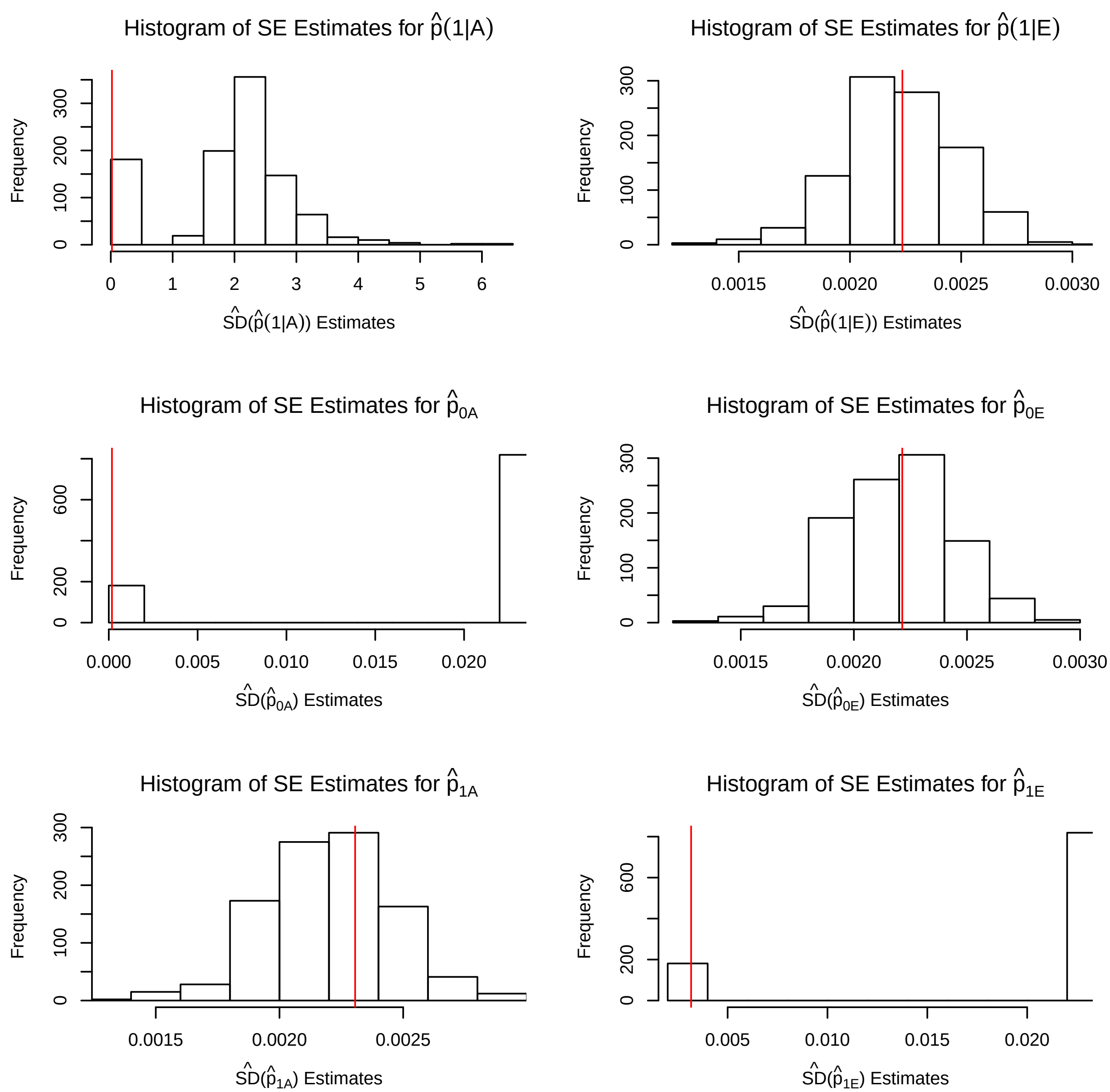

Figure 1.3: Simulation Setting 1: Histogram of $1,000$ SE Estimates and the Sample SD of $1,000$ MLEs Obtained on Simulated Two-Ancestry Data. Red lines: Sample SDs. Sample SDs for $\hat{p}_{0A}$, $\hat{p}_{0E}$, $\hat{p}_{1A}$, $\hat{p}_{1E}$, $\hat{p}(1|A)$, and $\hat{p}(1|E)$ were respectively 1.78×10^{-4}, 2.21×10^{-3}, 2.31×10^{-3}, 3.17×10^{-3}, 1.87×10^{-2}, and 2.24×10^{-3}.

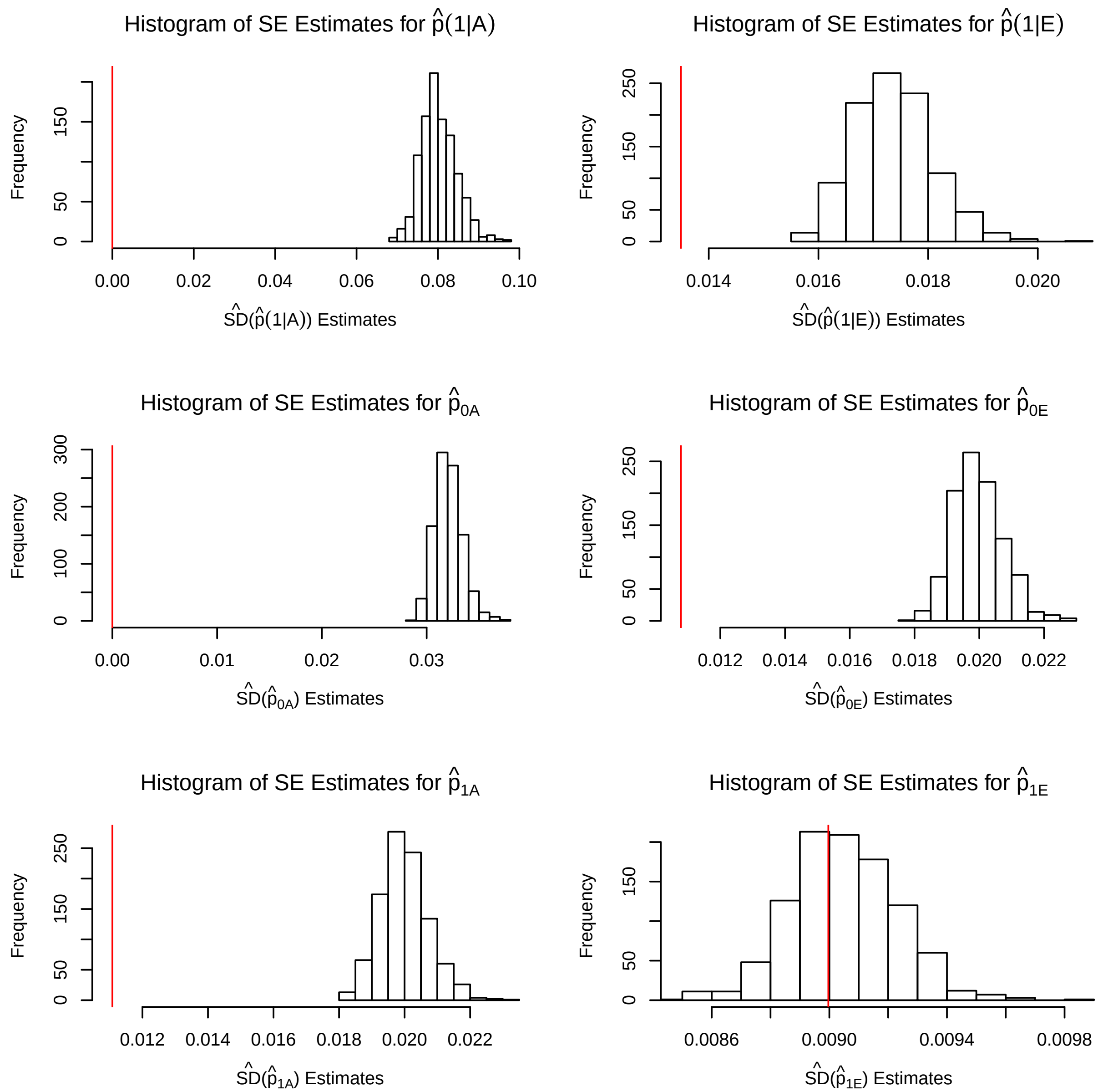

Figure 1.4: Simulation Setting 2: Histogram of 1,000 SE Estimates and the Sample SD of 1,000 MLEs Obtained on Simulated Two-Ancestry Data. Red lines: Sample SDs. Sample SDs for $\hat{p}_{0A}$, $\hat{p}_{0E}$, $\hat{p}_{1A}$, $\hat{p}_{1E}$, $\hat{p}(1|A)$, and $\hat{p}(1|E)$ were respectively 0, 1.08×10^{-2}, 1.11×10^{-2}, 9.00×10^{-3}, 0, and 1.35×10^{-2}.

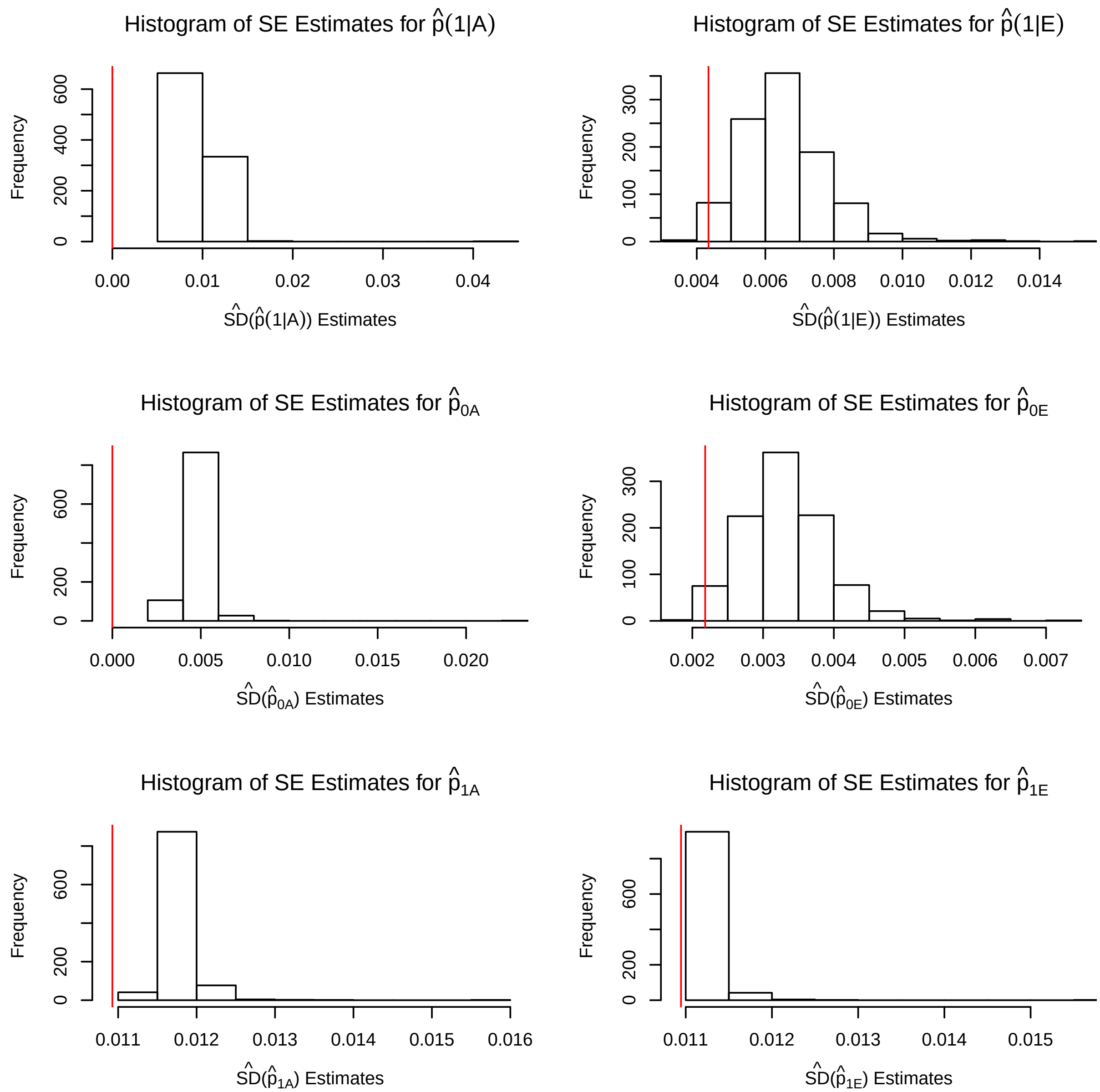

Figure 1.5: Simulation Setting 3: Histogram of 1,000 SE Estimates and the Sample SD of 1,000 MLEs Obtained on Simulated Two-Ancestry Data. Red lines: Sample SDs. Sample SDs for $\hat{p}_{0A}$, $\hat{p}_{0E}$, $\hat{p}_{1A}$, $\hat{p}_{1E}$, $\hat{p}(1|A)$, and $\hat{p}(1|E)$ were respectively 0, 2.18×10^{-3}, 1.09×10^{-2}, 1.09×10^{-2}, 0, and 4.34×10^{-3}.

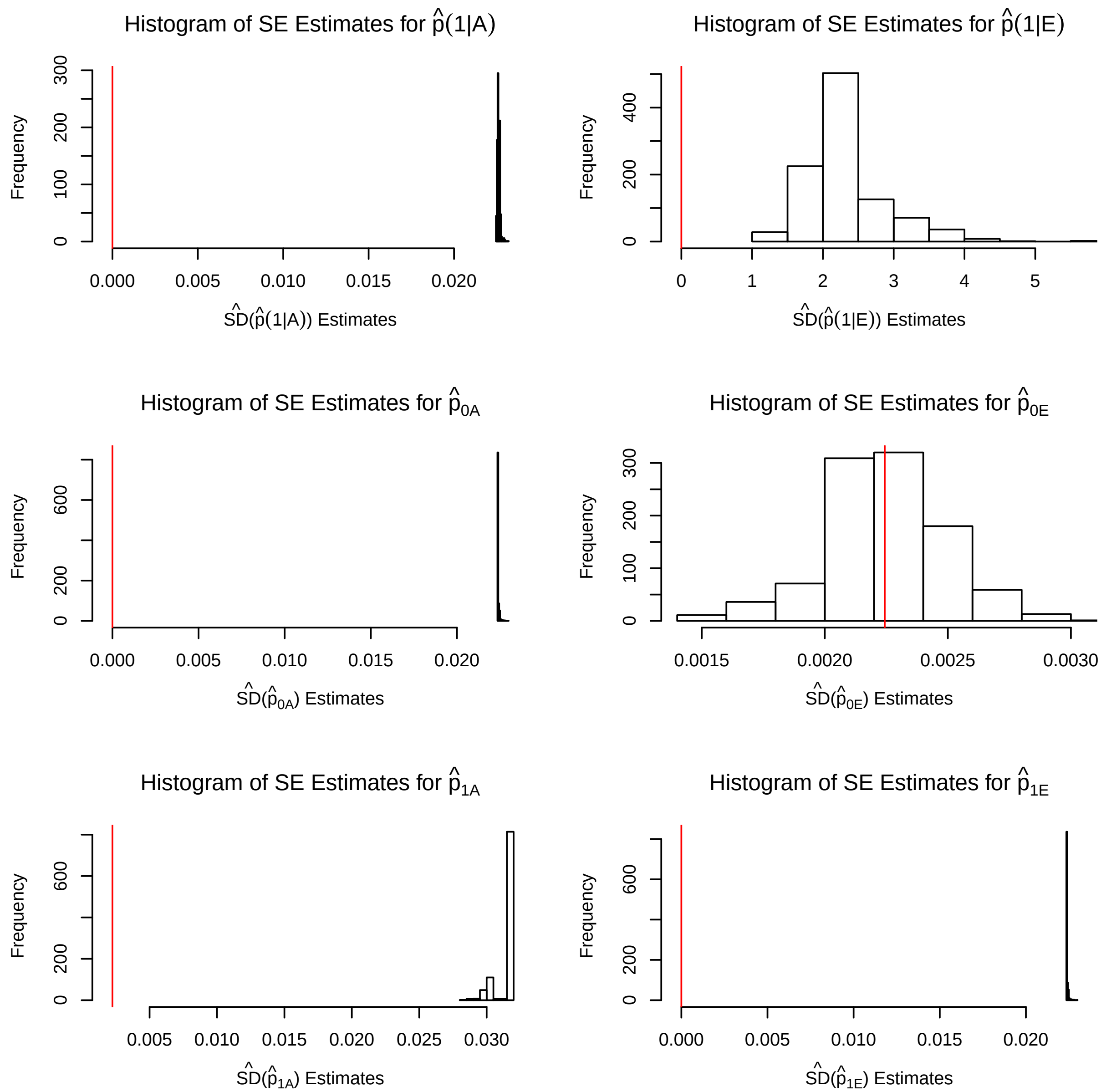

Figure 1.6: Simulation Setting 4: Histogram of $1,000$ SE Estimates and the Sample SD of $1,000$ MLEs Obtained on Simulated Two-Ancestry Data. Red lines: Sample SDs. Sample SDs for $\hat{p}_{0A}$, $\hat{p}_{0E}$, $\hat{p}_{1A}$, $\hat{p}_{1E}$, $\hat{p}(1|A)$, and $\hat{p}(1|E)$ were respectively 0, 2.24×10^{-3}, 2.24×10^{-3}, 0, 0, and 0.

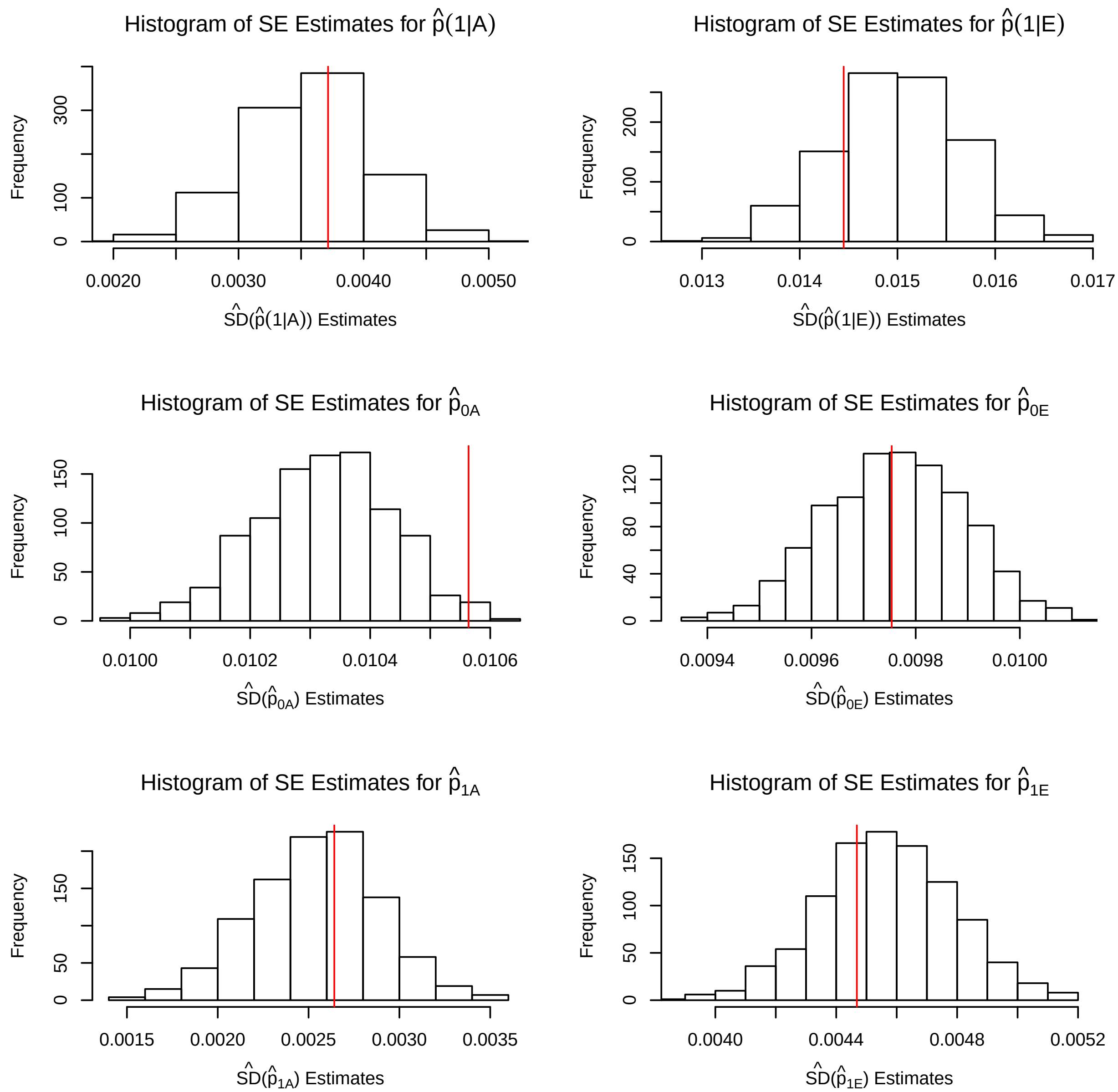

Figure 1.7: Simulation Setting 5: Histogram of $1,000$ SE Estimates and the Sample SD of $1,000$ MLEs Obtained on Simulated Two-Ancestry Data. Red lines: Sample SDs. Sample SDs for $\hat{p}_{0A}$, $\hat{p}_{0E}$, $\hat{p}_{1A}$, $\hat{p}_{1E}$, $\hat{p}(1|A)$, and $\hat{p}(1|E)$ were respectively 1.06×10^{-2}, 9.75×10^{-3}, 2.64×10^{-3}, 4.47×10^{-3}, 3.72×10^{-3}, and 1.44×10^{-2}.

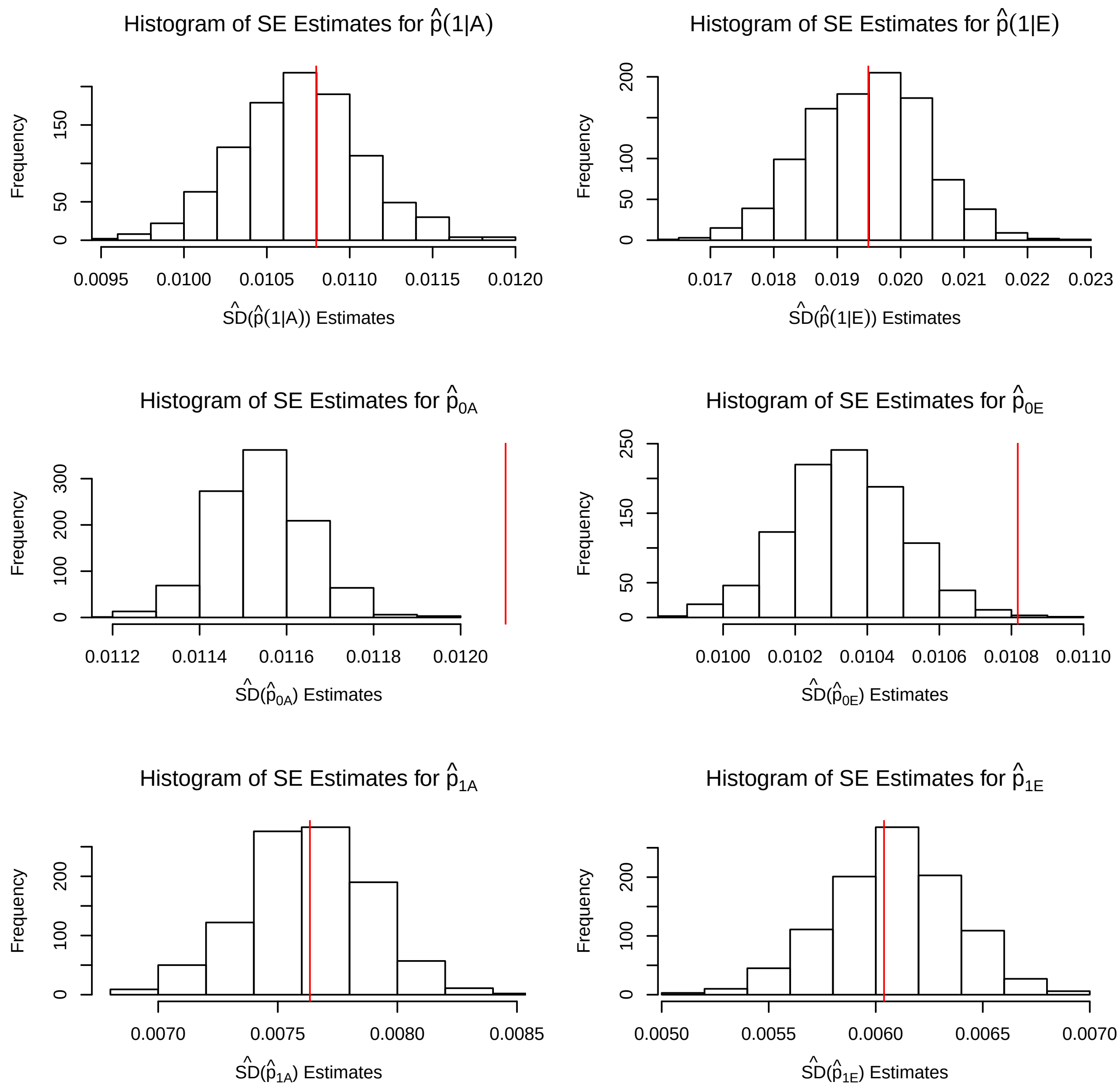

Figure 1.8: Simulation Setting 6: Histogram of $1,000$ SE Estimates and the Sample SD of $1,000$ MLEs Obtained on Simulated Two-Ancestry Data. Red lines: Sample SDs. Sample SDs for $\hat{p}_{0A}$, $\hat{p}_{0E}$, $\hat{p}_{1A}$, $\hat{p}_{1E}$, $\hat{p}(1|A)$, and $\hat{p}(1|E)$ were respectively 1.21×10^{-2}, 1.08×10^{-2}, 7.63×10^{-3}, 6.04×10^{-3}, 1.08×10^{-2}, and 1.95×10^{-2}

1.4.3 Results: Three-Ancestry Approach Applied to Simulated Data

Table 1.7 shows the results of applying the three-ancestry approach to simulated data, as discussed in Section 1.3.6. Comparing all allele frequency settings, the largest absolute value of mean of errors was 0.204, which occurred for $p(1|N)$ in Simulation Setting 1 where true frequencies were $p(1|A) = 0.0$, $p(1|E) = 0.5$, and $p(1|N) = 0.9$. Excluding 0.204, the largest absolute value of mean errors was 0.0125. The largest standard deviation of errors was 0.0527.

Due to uncertain phase of heterozygous genotype 0/1 relative to mixed ancestry pair (e.g. A/E), when the true $p(1|a), a \in \{A, E, N\}$ is 0, it is possible for a SNP's estimate $\hat{p}(1|a)$ to be non-zero, yielding non-zero mean error. For example, if allele 1 were simulated on the E background, and allele 0 were simulated on the A background, since the available data are observed counts $n_{gg',aa'}$'s of genotype gg' and ancestry pair aa' combinations that have no information regarding phase of genotype relative to ancestry pair, and a genotype 0/1 with an ancestry pair A/E is consistent with ga-alleles $1A$ and $0E$ as well as ga-alleles $0A$ and $1E$, it is reasonable for the estimate $\hat{p}(1, A)$ to be non-zero, so that the ancestry-specific estimate $\hat{p}(1|A)$ is non-zero, so that the mean error for $p(1|A)$ is non-zero in a simulation setting where the true $p(1|A)$ is 0.

Figure 1.9 gives violin plots of ancestry-specific allele frequency $[p(1|A), p(1|E), p(1|N)]$ estimates (black) with ancestry-specific allele frequencies (blue) used to generate alleles given ancestry draws. Estimates have a mean close to the ancestry-specific allele frequency used to generate alleles given ancestry draws. As a check that the method and our implementation of it give reasonable results, the $[p(1|A), p(1|E), p(1|N)] =$ [0.0, 0.5, 0.9], [0.5, 0.9, 0.0], and [0.9, 0.0, 0.5] allele frequency settings, which are permutations of the same collection of ancestry-specific allele frequency values, result in permuted error plots.

Table 1.8 illustrates that the estimates we obtain are consistent with our earlier claim, proved in Section 1.3.1.2, that having every $X_{ga} > 0$ is sufficient for an MLE $\hat{\mathbf{p}}$ to be in the interior of the domain. The table further illustrates that having every $X_{ga} > 0$ is not

Table 1.7: Mean of Errors (i.e. Bias of Estimator) and SD of Errors (i.e. SD of Estimator), where Error = Estimated Ancestry-Specific Allele Frequency - True Ancestry-Specific Allele Frequency, for 1,000 SNPs per Combination of Ancestral Population Frequencies $p(1|A)$, $p(1|E)$, and $p(1|N)$. The frequencies in the table are the ancestral population frequencies used to generate the data, and were considered the true ancestry-specific allele frequencies used to calculate errors. The mean of errors in estimating ancestry-a specific allele frequency is Mean Error $p(1|a)$. The standard deviation of errors in estimating ancestry-a specific allele frequency is SD Error $p(1|a)$. A pictoral display of estimates and true parameter values is given in Figure 1.9. Case 2 of the three-ancestry approach was used for all SNPs.

$p(1\|A)$	$p(1\|E)$	$p(1\|N)$	Mean Error $p(1\|A)$	SD Error $p(1\|A)$	Mean Error $p(1\|E)$	SD Error $p(1\|E)$	Mean Error $p(1\|N)$	SD Error $p(1\|N)$
0	0.5	0.9	1.25E-02	8.49E-03	-9.22E-03	4.33E-02	-2.04E-01	3.99E-02
0.1	0.5	0.8	-9.40E-06	2.68E-02	-2.94E-03	4.48E-02	-4.70E-04	3.67E-02
0.4	0.5	0.6	1.47E-03	4.82E-02	-1.64E-03	5.03E-02	-6.15E-04	4.89E-02
0.5	0.5	0.5	7.44E-04	5.27E-02	-7.34E-04	5.08E-02	1.70E-03	5.16E-02
0.1	0.6	0.6	1.99E-03	2.94E-02	-1.58E-04	4.61E-02	-1.10E-03	4.66E-02
0.4	0.5	0.5	4.38E-04	5.07E-02	-3.95E-04	5.08E-02	-1.60E-03	5.07E-02
0.5	0.9	0	-9.42E-03	4.32E-02	-3.45E-03	2.63E-02	1.19E-02	8.53E-03
0.9	0	0.5	-2.71E-03	2.75E-02	1.22E-02	8.90E-03	-1.05E-02	4.32E-02

Table 1.8: Co-Occurrence of Meeting the Condition that Every $X_{ga} > 0$, and the Three-Ancestry Approach Yielding an Interior MLE $\hat{\mathbf{p}}$, in Simulated Data. From the lower row, we see that whenever the condition is met, the estimate $\hat{\mathbf{p}}$ is in the interior, consistent with our claim in Section 1.3.1 that having every $X_{ga} > 0$ is sufficient for an MLE to be in the interior. From the right column, we see that it is possible for the estimate to be in the interior, without meeting the condition, i.e. the converse statement (if the estimate is in the interior, then the condition holds) is not true.

	Estimate Not in Interior	Estimate in Interior
Assumption Not Met	593	2407
Assumption Met	0	5000

necessary for an MLE $\hat{\mathbf{p}}$ to be in the interior of the domain.

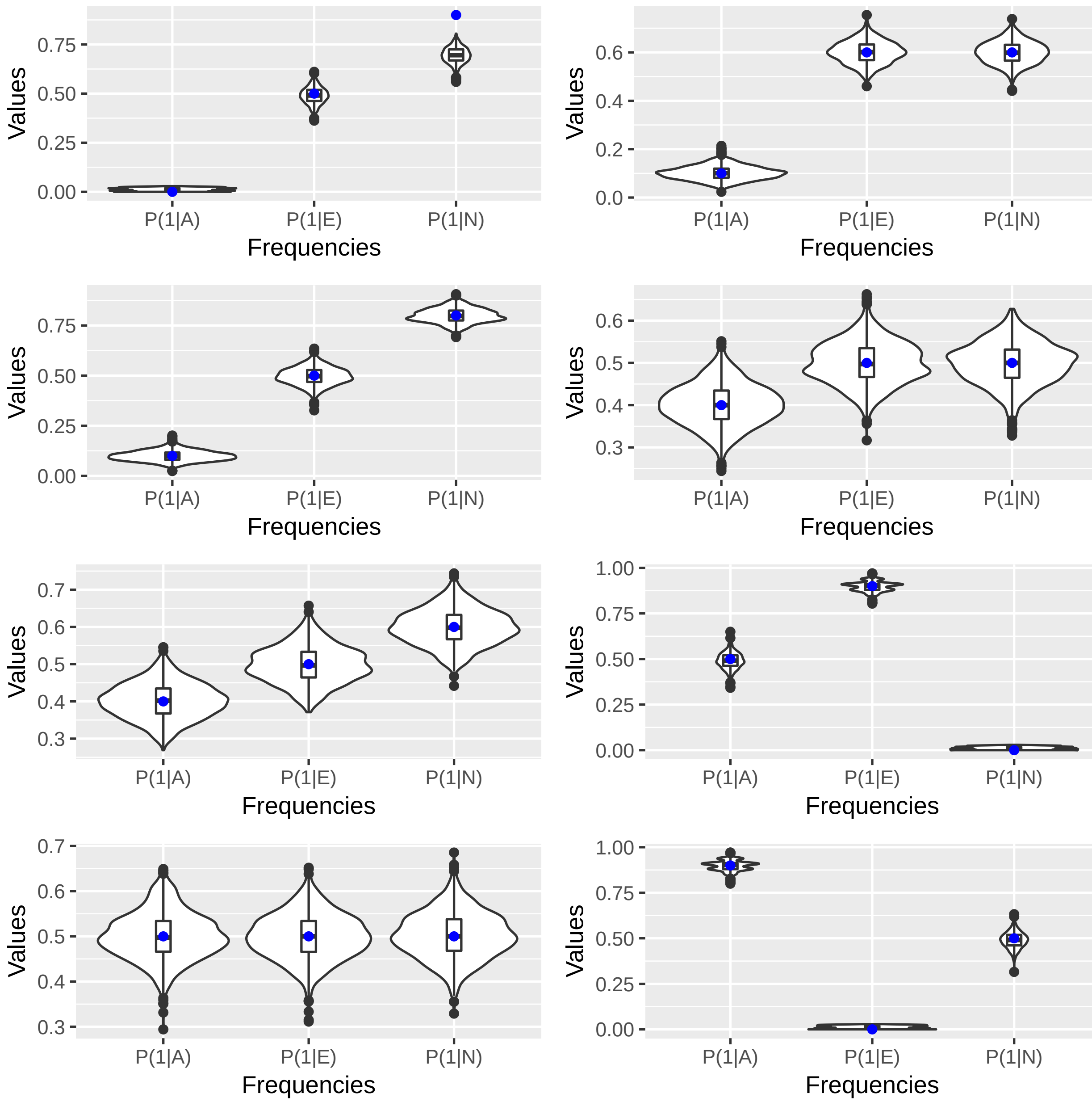

Figure 1.9: Violin plots of ancestry-specific allele frequency estimates (black) obtained by the 3-ancestry approach, with true ancestry-specific allele frequency values used to generate data (blue dot). Eight allele frequency combinations for $[p(1|A), p(1|E), p(1|N)]$ are considered, going from the left column (top to bottom) to the right column (top to bottom) panels: $[0.0, 0.5, 0.9]$, $[0.1, 0.5, 0.8]$, $[0.4, 0.5, 0.6]$, $[0.5, 0.5, 0.5]$, $[0.1, 0.6, 0.6]$, $[0.4, 0.5, 0.5]$, $[0.5, 0.9, 0.0]$, $[0.9, 0.0, 0.5]$. For every allele frequency combination and ancestry-specific allele frequency $p(1|A)$, $p(1|E)$, or $p(1|N)$, estimates (black) appear centered around the true parameter value (blue dot), consistent with low mean error (i.e. bias) seen in Table 1.7.

No SNP required Case 1 of the three-ancestry approach. In other words, Case 2 of the three-ancestry approach was used for all SNPs. No SNP had multiple maximizers from a sub-grid search.

1.4.4 Results: Two-Ancestry Approach Applied to 1000 Genomes Data on the ASW Population

We applied the two-ancestry approach to 141,318 Chr 22 SNPs and 61 African American (ASW) individuals of the 1000 Genomes Phase 1 data. Depending on the case of the two-ancestry approach, a SNP's data can result in one to three real solutions $\{p^*_{0A,i}, i \in \{1,2,3\}\}$ to Equation 1.10. For each case of the two-ancestry approach and each number of possible p^*_{0A} solutions to Equation 1.11, we randomly selected a Chr 22 SNP for which we employed that case and obtained that number of p^*_{0A} solutions. Table 1.9 illustrates these selected SNPs' counts $n_{gg',aa'}$'s of genotype gg' and ancestry pair aa' combinations, and shows that many cases of the two-ancestry approach were used to analyze these data. Table 1.10 gives the same SNPs' p^*_{0A} solutions to Equation 1.11. This table illustrates the two-ancestry approach described in Section 1.3.2 by displaying the p^*_{0A} component of stationary points $\mathbf{p}^* = [p^*_{0A}, p^*_{0E}, p^*_{1A}, p^*_{1E}]$, within-bounds stationary points, and MLEs (the within-bounds stationary point with highest log likelihood value amongst within-bounds stationary points) for SNPs that span a variety of cases considered in the two-ancestry approach.

In Table 1.10, rsIDs rs182927015, rs137437, rs2412981, rs73162032, rs8140818, and rs28553294 each have a unique solution $\mathbf{p}^*$, the MLE $\hat{\mathbf{p}}$.

rsIDs rs1297593, rs1044085, and rs77200745 each have multiple solutions, with only one solution within bounds imposed by marginal frequency estimates, i.e. $0 \leq p^*_{ga} \leq min(\hat{p}_{g.}, \hat{p}_{.a})$ for all $g \in \{0,1\}$ and $a \in \{A, E\}$. The one, within-bounds solution is the MLE. For rs1297593 for example, the MLE $\hat{\mathbf{p}}$ has component $\hat{p}_{0A}$ that is 0.

rsIDs rs132365, rs5753697, rs112153936, and rs116822375 each have multiple solutions, with two solutions within bounds imposed by marginal frequency estimates. Evaluating the log likelihood function at both solutions, we can identify which solution results in a higher likelihood value, and so identify an MLE. For rsID rs132365 for example, the MLE $\hat{\mathbf{p}}$ has component $\hat{p}_{0A}$ that is 0. These SNPs demonstrate that there are cases in real data where our analytical two-ancestry approach has certain advantages over the EM algorithm. When the log likelihood function has multiple stationary points that are not all MLEs, the EM algorithm may converge to a stationary point that is not an MLE, depending on starting value. Our analytical approach does not rely on a lucky starting value to deliver an MLE. With EM, one can try multiple starting values to improve the chance of finding an MLE. However, iterating a system of four equations multiple times (once per starting value) can be more computationally intensive than our analytical approach. In the two-ancestry case, one can reduce luck involved in the choice of the EM algorithm's starting value by plotting the 1-dimensional log likelihood function as a function of p_{0A}, visually inspecting this plot to come up with an approximate value for a $\hat{p}_{0A}$ belonging to an MLE $\hat{\mathbf{p}}$, calculating the other $\hat{p}_{0E}$, $\hat{p}_{1A}$, and $\hat{p}_{1E}$ components corresponding to the approximate $\hat{p}_{0A}$, and using the approximate $\hat{\mathbf{p}}$ as a starting value for the EM algorithm. However, our analytical approach can be automated, as it does not necessitate the creation and human inspection of such a plot, so requires less time to deliver an MLE.

Incidentally, and perhaps counterintuively, in cases where only 1 $X_{ga} = 0$, the stationary point with component $p^*_{ga} = 0$ is not necessarily an MLE, which is true of SNPs rs5753697 and rs116822375. Table 1.9 indicates that rs5753697 has count $X_{0E} = 0$ and rs116822375 has count $X_{1E} = 0$. However, rs5753697 does not have MLE $\hat{\mathbf{p}} = [\hat{p}_{0A}, \hat{p}_{0E}, \hat{p}_{1A}, \hat{p}_{1E}]$ with component $\hat{p}_{0E} = 0$, and rs116822375 does not have MLE $\hat{\mathbf{p}}$ with component $\hat{p}_{1E} = 0$: In Table 1.10, one can use relations $\hat{p}_{0E} = \hat{p}_{0.} - \hat{p}_{0A}$, $\hat{p}_{1A} = \hat{p}_{.A} - \hat{p}_{0A}$, and $\hat{p}_{1E} = 1 - \hat{p}_{0.} - \hat{p}_{.A} + \hat{p}_{0A}$, to see this. This result is consistent with a comment in Weir and Cockerham (1979): "There is no way of predicting, in general, if the quadratics will provide valid roots, or if they do, which of the valid roots is the ML estimate. The likelihoods must still be evaluated for each valid root," where "valid" means a root within the domain of the likelihood function.

Table 1.9: Number of SNPs per Case of the Two-Ancestry Approach Applied to 141,318 Chr 22 SNPs in the 1000 Genomes ASW Dataset, and Representative SNPs per Case. Case stands for the way solutions p^*_{0A}'s to Equation 1.11 were determined. For each number of solutions and case of the two-ancestry approach, we randomly selected a Chr 22 SNP for which we employed that case and obtained that number of p^*_{0A} solutions. Multiple SNPs with the same case are shown because they have different numbers of p^*_{0A} solutions, as can be seen from Table 1.10. We give the SNP's rsID, base pair position (bp), and number $n_{gg',aa'}$ of 61 ASW individuals in each genotype ancestry pair category gg', aa' where $gg' \in \{0/0, 0/1, 1/1\}$ and $aa' \in \{A/A, A/E, E/E\}$. Column "N SNPs per Case" gives the number of 141,318 SNPs per case. Cases that were not employed on any SNP's $n_{gg',aa'}$ counts are absent from the table. Cases "Cubic Dis. <0" and "Cubic Dis. >0" stand for cases that involved solving a cubic equation with a negative discriminant and positive discriminant respectively.

rsID	bp	$n_{00,AA}$	$n_{00,AE}$	$n_{00,EE}$	$n_{01,AA}$	$n_{01,AE}$	$n_{01,EE}$	$n_{11,AA}$	$n_{11,AE}$	$n_{11,EE}$	Case	N SNPs per Case
rs182927015	30728120	28	27	5	1	0	0	0	0	0	$n_{01,AE}=0$	34,509
rs1297593	28169521	0	0	0	0	6	4	31	18	2	$X_{0A}=0$	769
rs137437	33594193	0	0	1	0	7	3	31	16	3	$X_{0A}=0$	769
rs132365	29599486	0	0	0	0	2	0	31	23	5	$X_{0A}=X_{0E}=0$	182
rs2412981	30600065	1	0	0	9	3	0	19	24	5	$X_{0E}=0$	582
rs5753697	31967532	0	0	0	1	1	0	29	25	5	$X_{0E}=0$	582
rs1044085	32014719	30	20	3	0	6	2	0	0	0	$X_{1A}=0$	4,552
rs73162032	33373326	30	19	3	0	5	4	0	0	0	$X_{1A}=0$	4,552
rs112153936	32100119	30	25	5	0	1	0	0	0	0	$X_{1A}=X_{1E}=0$	15,755
rs116822375	31604422	29	25	5	1	1	0	0	0	0	$X_{1E}=0$	41,945
rs8140818	33228722	20	21	5	8	7	0	0	0	0	$X_{1E}=0$	41,945
rs77200745	30719320	28	22	3	1	5	2	0	0	0	Cubic Dis. <0	4,815
rs28553294	32680887	20	21	3	7	7	2	1	0	0	Cubic Dis. >0	38,209

Table 1.10: Results of Applying the Two-Ancestry Approach to Representative Chr 22 SNPs in the 1000 Genomes ASW Dataset. We give each representative SNP's rsID, base pair position (bp), and real p^*_{0A} solutions to Equation 1.11. Columns $p^*_{0A,2}$ and $p^*_{0A,3}$ have empty entries when a SNP did not have a second or third real solution. A dagger † means the $p^*_{0A,i}$ value is a component of solution vector $\mathbf{p}^*_i$ that is within the domain of the log likelihood function, i.e. $0 \leq p^*_{ga} \leq min(\hat{p}_{g.}, \hat{p}_{.a})$, for all $g \in \{0,1\}$ and $a \in \{A, E\}$. A bold $p^*_{0A,i}$ value means the corresponding solution vector $\mathbf{p}^*_i$ is an MLE. These SNPs' $n_{gg',aa'}$ counts are given in Table 1.9.

rsID	bp	$p^*_{0A,1}$	$p^*_{0A,2}$	$p^*_{0A,3}$	$\hat{p}_{0.}$	$\hat{p}_{.A}$
rs182927015	30728120	**0.6885†**			0.9918	0.6967
rs1297593	28169521	**0.0000†**	0.1803	0.1311	0.0820	0.7049
rs137437	33594193	**0.0000†**			0.0984	0.6967
rs132365	29599486	**0.0000†**	0.0164†	0.2213	0.0164	0.7131
rs2412981	30600065	**0.1148†**			0.1148	0.6967
rs5753697	31967532	0.0164†	0.2030	**0.0142†**	0.0164	0.7049
rs1044085	32014719	**0.7049†**	0.6230	0.4918	0.9344	0.7049
rs73162032	33373326	**0.6885†**			0.9262	0.6885
rs112153936	32100119	**0.7049†**	0.6967†	0.4959	0.9918	0.7049
rs116822375	31604422	0.6885†	**0.6907†**	0.5019	0.9836	0.7049
rs8140818	33228722	**0.5656†**			0.8770	0.6885
rs77200745	30719320	**0.6857†**	0.5062	0.6073	0.9344	0.6967
rs28553294	32680887	**0.5851†**			0.8525	0.6885

1.4.5 Results: Three-Ancestry Approach Applied to 1000 Genomes Data on the PUR Population

Table 1.11 shows that both Case 1 and Case 2 of the 3-ancestry approach are applicable to 1000 Genomes Phase 1 SNP data. For each case, a SNP's observed counts $n_{gg',aa'}$'s of genotype gg' and ancestry pair aa' combinations and $[\hat{p}(1|A), \hat{p}(1|E)]$ allele frequency MLEs are given as examples, should the reader be interested in reproducing estimates. Figure 1.10 plots $\hat{p}(1|A)$, $\hat{p}(1|E)$, and $\hat{p}(1|N)$ estimates, illustrating that most estimates are close to 0.

There were no SNPs with multiple maximizers from grid search.

Table 1.11: Results of Applying the Three-Ancestry Approach to Certain SNPs in the 1000 Genomes PUR Data. For two Chr 22 SNPs, we give rsID, and number $n_{gg',aa'}$ of 55 PUR individuals with each genotype gg' and ancestry pair aa' combination, where $gg' \in \{0/0, 0/1, 1/1\}$, and $aa' \in \{A/A, E/A, E/E, N/A, N/E, N/N\}$. We also give maximum likelihood estimates $\hat{p}_{0A}$ and $\hat{p}_{0E}$. From $\hat{p}_{0A}$, $\hat{p}_{0E}$, and marginal estimates $\hat{p}_{0.}$, $\hat{p}_{.A}$, and $\hat{p}_{.E}$, we obtain estimates for the other p_{ga}'s, i.e. $\hat{p}_{0N} = \hat{p}_{0.} - \hat{p}_{0A} - \hat{p}_{0E}$, $\hat{p}_{1A} = \hat{p}_{.A} - \hat{p}_{0A}$, $\hat{p}_{1E} = \hat{p}_{.E} - \hat{p}_{0E}$, and $\hat{p}_{1N} = 1 - \hat{p}_{0.} - \hat{p}_{.A} - \hat{p}_{.E} + \hat{p}_{0A} + \hat{p}_{0E}$. Case stands for the way the $\hat{p}_{0A}$ and $\hat{p}_{0E}$ estimates were obtained, as described in Section 1.3.3. Case 1 means $\hat{p}_{0A}$ and $\hat{p}_{0E}$ estimates were obtained analytically. Case 2 means $\hat{p}_{0A}$ and $\hat{p}_{0E}$ estimates were obtained using a grid search of the log likelihood function considered as a function $l(p_{0A}, p_{0E})$ of only two parameters.

<table>
<tr><th colspan="10">Example of a SNP with a Particular Case</th><th rowspan="2">Case</th><th rowspan="2">N SNPs / Case</th></tr>
<tr><th>rsID</th><th>Counts $n_{gg',aa'}$</th><th>A/A</th><th>A/E</th><th>E/E</th><th>N/A</th><th>N/E</th><th>N/N</th><th>$\hat{p}_{0A}$</th><th>$\hat{p}_{0E}$</th></tr>
<tr><td rowspan="3">rs114039316</td><td>0/0</td><td>2</td><td>5</td><td>36</td><td>1</td><td>10</td><td>1</td><td rowspan="3">0.091</td><td rowspan="3">0.791</td><td rowspan="3">1</td><td rowspan="3">712</td></tr>
<tr><td>0/1</td><td>0</td><td>0</td><td>0</td><td>0</td><td>0</td><td>0</td></tr>
<tr><td>1/1</td><td>0</td><td>0</td><td>0</td><td>0</td><td>0</td><td>0</td></tr>
<tr><td rowspan="4">rs5993559</td><td>Counts $n_{gg',aa'}$</td><td>A/A</td><td>E/A</td><td>E/E</td><td>N/A</td><td>N/E</td><td>N/N</td><td>$\hat{p}_{0A}$</td><td>$\hat{p}_{0E}$</td><td rowspan="4">2</td><td rowspan="4">288</td></tr>
<tr><td>0/0</td><td>0</td><td>0</td><td>13</td><td>0</td><td>0</td><td>0</td><td rowspan="3">0.010</td><td rowspan="3">0.430</td></tr>
<tr><td>0/1</td><td>1</td><td>1</td><td>15</td><td>0</td><td>6</td><td>0</td></tr>
<tr><td>1/1</td><td>1</td><td>4</td><td>8</td><td>1</td><td>4</td><td>1</td></tr>
</table>

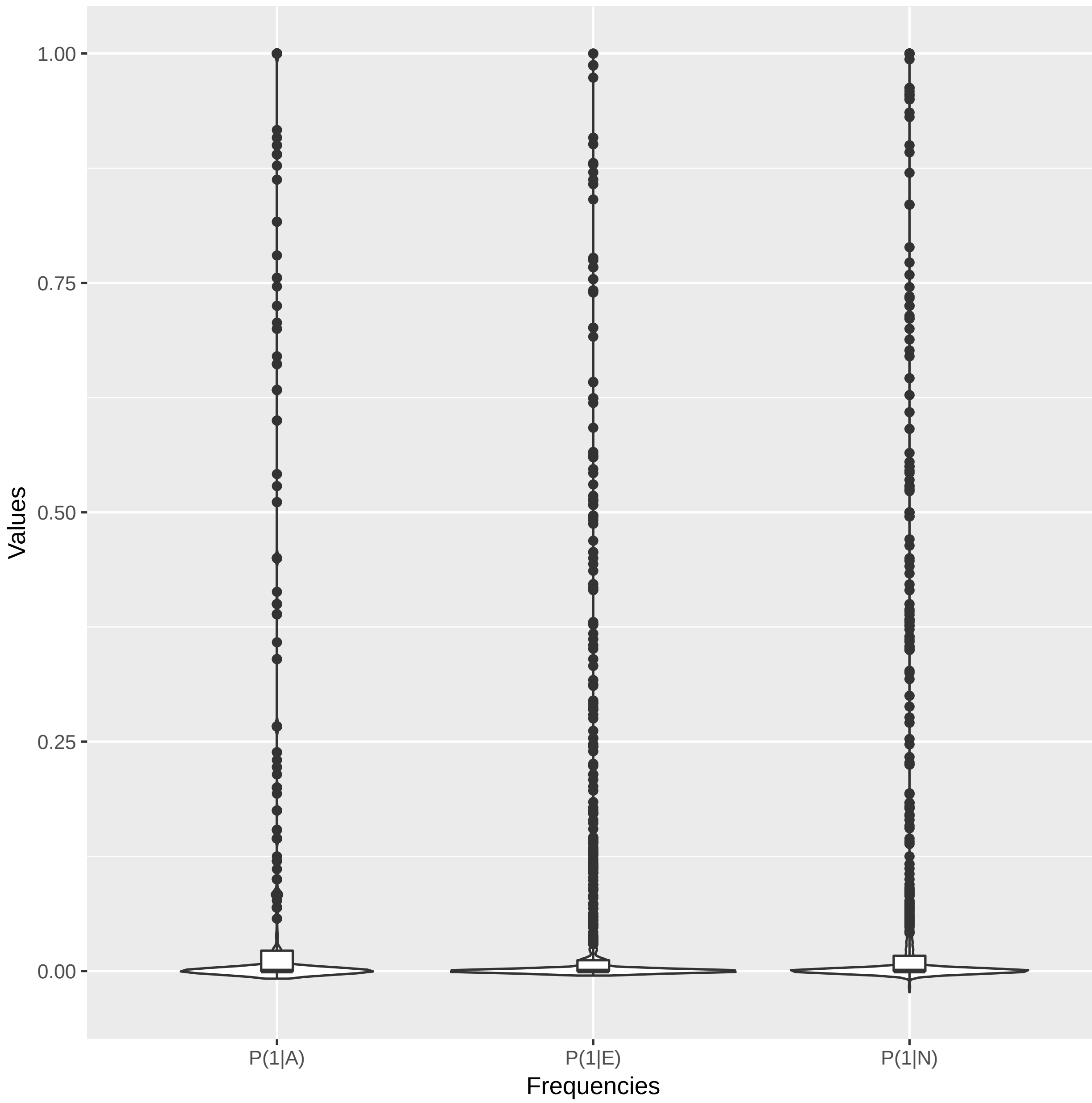

Figure 1.10: Violin plots of ancestry-specific allele frequency estimates (black) obtained by applying the three-ancestry approach to ancestry pairs and genotypes for 1,000 Chr 22 SNPs in Phase 1 PUR data.

1.4.6 Results: Comparing the Three-Ancestry Approach to the ASAFE EM Algorithm with 1000 Genomes PUR Data

Figure 1.11 plots our novel three-ancestry approach estimates against ASAFE EM algorithm estimates for $1,000$ SNPs in the 1000 Genomes PUR data. Our novel three-ancestry approach estimates did not differ much from ASAFE EM algorithm estimates.

We did not detect a SNP with a multimodal log likelihood function, since our novel three-ancestry approach estimates were close to ASAFE EM algorithm estimates. We detected SNPs satisfying Assumption 1 sufficient for a unimodal log likelihood function, and amongst these, found that the Case 1 SNP of Table 1.11 had a boundary MLE. This SNP's ASAFE EM algorithm estimate $\mathbf{p}^*$ had elements $\hat{p}_{0A} = 0.09090909$, $\hat{p}_{0E} = 0.79090909$, $\hat{p}_{0N} = 0.11818182$, $\hat{p}_{1A} = 0.00000001$, $\hat{p}_{1E} = 0.00000001$, and $\hat{p}_{1N} = 0.00000001$. This SNP's extended three-ancestry approach estimate $\hat{\mathbf{p}}$ had elements $\hat{p}_{0A} = 0.09090909$, $\hat{p}_{0E} = 0.79090909$, $\hat{p}_{0N} = 0.11818182$, $\hat{p}_{1A} = 0$, $\hat{p}_{1E} = 0$, and $\hat{p}_{1N} = 0$. The estimates are similar, as we expected, because this SNP satisfies Assumption 1 ($n_{01,AE} = n_{01,NA} = n_{01,NE} = 0$) sufficient for a unimodal log likelihood, so a stationary point, which an EM algorithm yields, is an MLE, which the Case 1 analytical formula of our novel three-ancestry approach yields. These estimates are different, as we expected, in that our extended three-ancestry approach gives exact boundary estimates of 0 for $\hat{p}_{1A}$, $\hat{p}_{1E}$, and $\hat{p}_{1N}$, while the ASAFE EM algorithm implementation yields 10^{-8}. This difference arises because our extended three-ancestry approach involves an analytical solution, while the ASAFE EM algorithm implementation pushes each k-th iterate's $p_{k,ga}$ from the M step off boundary values of 0 and 1 by 10^{-8}.

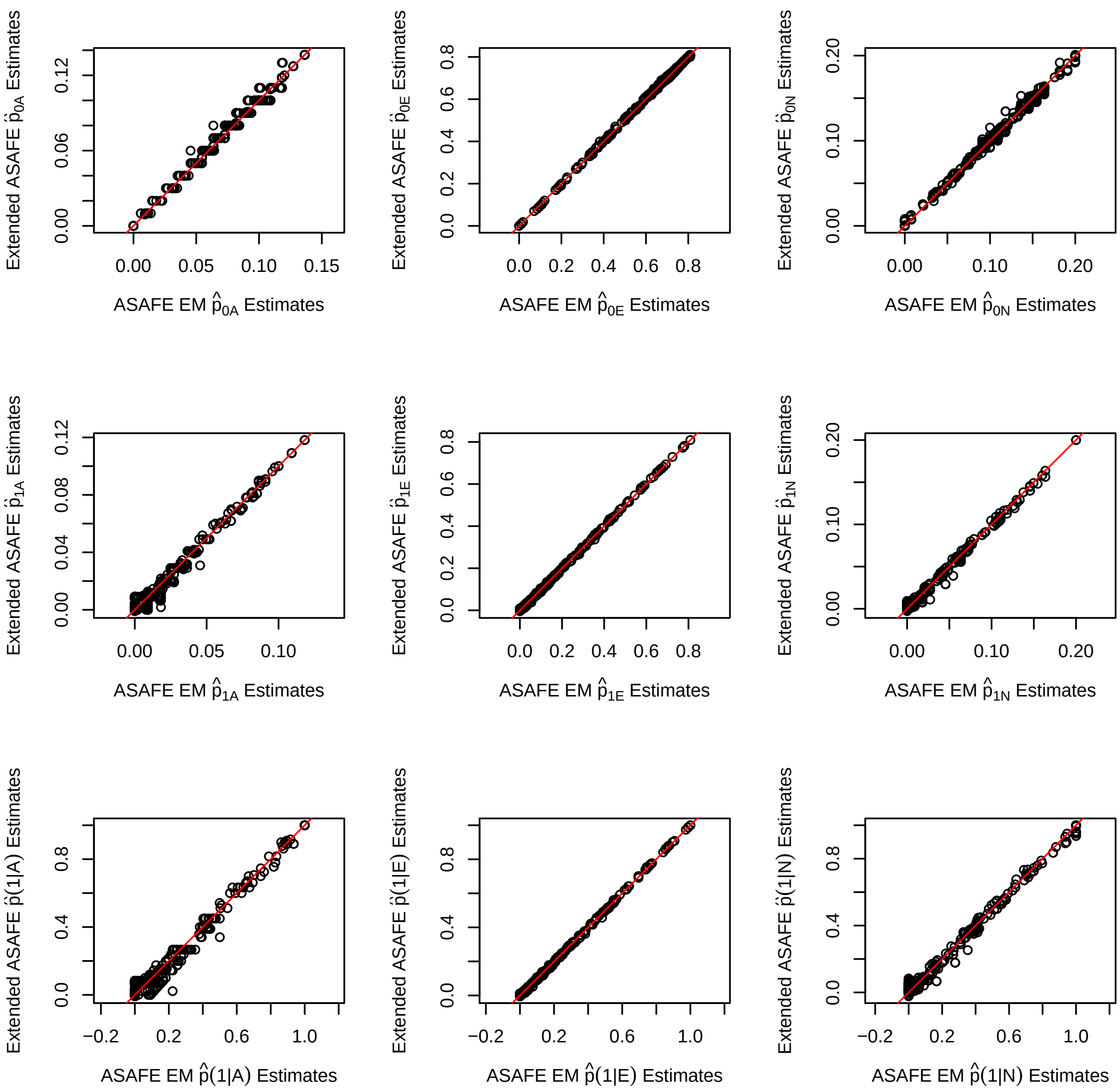

Figure 1.11: Plots of Our Extended ASAFE Three-Ancestry Approach Estimates $\hat{p}_{0A}$, $\hat{p}_{0E}$, $\hat{p}_{0N}$, $\hat{p}_{1A}$, $\hat{p}_{1E}$, $\hat{p}_{1N}$, $\hat{p}(1|A), \hat{p}(1|E)$, and $\hat{p}(1|N)$ vs ASAFE EM Algorithm Estimates for $1,000$ SNPs in 1000 Genomes PUR Data

1.4.7 Results: Using ASAFE Allele Frequency Estimates for Population Structure

Using ancestry-specific allele frequency estimates from Section 1.4.5 as described in Section 1.3.10, we obtained the following β_i estimates for African, European, and Native American ancestries respectively: 1.75×10^{-1} , 2.11×10^{-1}, 2.28×10^{-1}. Bootstrap 95% confidence intervals were respectively $(0.10, 0.24)$, $(0.16, 0.26)$, and $(0.16, 0.29)$. We would expect that due to greater genetic diversity in Africa, θ_i for Africans is lower than θ_i for Europeans and for Native Americans. From the definition of β_i, $\beta_i := (\theta_i - \theta_B)/(1 - \theta_B)$, we would then expect β_i for Africans to be less than β_i for Europeans and β_i for Native Americans. Our estimates are consistent with this expectation.

1.5 Discussion

As more GWASs are performed in admixed populations such as African Americans and Hispanics, it is becoming increasingly important to estimate ancestry-specific allele frequencies for bi-allelic markers. Here, we extend an existing approach, ASAFE (Zhang et al., 2016), that solves this estimation problem for three ancestries using an EM algorithm. We list methodological extensions below:

1. We provide a likelihood framework for this estimation problem for any finite number K of ancestries and two alleles

2. We make use of marginal frequency estimates $\hat{p}_{g.}$'s and $\hat{p}_{.a}$'s to reduce the dimensionality of the problem, from maximizing a likelihood function of $(2K - 1)$ free p_{ga} parameters to maximizing a likelihood function of $(2 - 1) * (K - 1) = (K - 1)$ free p_{ga} parameters. This reduction inspires our two- and three-ancestry approaches.

3. We provide estimation approaches for two ancestries as well as three ancestries, rather than only three ancestries

4. We provide observed Fisher information-based standard error estimators $\hat{SD}(\hat{p}_{ga}), g \in$

$\{0,1\}, a \in \{A, E\}$ and $\hat{SD}(\hat{p}(g=1|a)), a \in \{A, E\}$ for two ancestries. Our simulation studies showed that their performance was poor when the number of individuals n was 100. NA variance estimates resulted because SNPs were generated with many 0 counts. Performance was much better when the number of individuals n was $1,000$ and the true parameter $\mathbf{p}$ was in the interior of the parameter space. NA standard error estimates were avoided with $n = 1,000$, because no SNP was generated to have so many 0 $n_{gg',aa'}$ counts that the inverse information matrix was undefined. Since our standard error estimators were derived with standard asymptotic theory that assumes an interior true parameter $\mathbf{p}$, other standard error estimators derived by considering the asymptotics literature for boundary-valued true parameters (e.g. the work of Feng and McCulloch (1992) and Self and Liang (1987)) would perhaps have better performance in simulation settings with a boundary $\mathbf{p}$.

5. We show two ways, by setting Lagrangian partial derivatives to 0 and by deriving EM convergence equations, that points $\{\mathbf{p}^*\}$ satisfying a system of equations motivating our two-ancestry approach must be stationary points. The ascent property of the EM algorithm further shows that they are either saddle points or local maximizers.

6. All cases of the two-ancestry approach and Case 1 of the three-ancestry approach involve closed-form solutions for all stationary points of the log likelihood function, which are more accurate than stationary points obtained via the EM algorithm, whose accuracy depends on the number of EM iterations. Case 2 of the three-ancestry approach involves a grid search whose accuracy in obtaining an MLE - if a grid point does not correspond exactly to an MLE $\hat{\mathbf{p}}$ - depends on the grid interval, analogous to how the EM algorithm's accuracy depends on the number of EM iterations.

7. Our closed-form solutions and grid search allow for exact zeroes in estimates of joint allele-g, ancestry-a probabilities $\hat{p}_{ga}$'s in $\hat{\mathbf{p}}$, whereas the ASAFE EM implementation did not allow for exact boundary estimates. Given the existence of rare, population-

specific alleles uncovered by sequencing data, it is plausible for an MLE $\hat{\mathbf{p}}$ to have boundary components.

8. Our two-ancestry approach and Case 1 of our three-ancestry approach provide analytical formulas for all stationary points, and Case 2 of our three-ancestry approach examines all stationary points to the nearest user-specified ϵ. By examining all stationary points, our approaches can find an MLE when it is one of multiple stationary points, bypassing the luck involved in an EM algorithm, which returns one of possibly multiple stationary points, depending on starting value (Wu, 1983). Indeed, we have demonstrated in 1000 Genomes data (Table 1.10) that it is possible for a SNP's counts $n_{gg',aa'}$'s of genotype gg' and ancestry pair aa' combinations to result in a log likelihood function with multiple stationary points, but only one MLE.

   To reduce luck involved in the EM algorithm, one could try multiple starting values, but iterating $2K$ equations to convergence multiple times could require more computation time to deliver an MLE than our analytical formulas or grid search. One could also inspect a plot of the 1-dimensional log likelihood function (in the case of two ancestries) or the 2-dimensional log likelihood function (in the case of three ancestries) to inform the choice of starting value for EM, but this requires more human time to deliver an MLE than our analytical formulas or grid search, which can be automated.

9. We have proven that when all genotype ancestry pair counts $n_{gg',aa'}$'s are zero for categories with heterozygous genotype and ancestry pair with differing ancestries , the negative log likelihood function is convex, so a stationary point must be an MLE. Under this condition, our extended two-ancestry approach (which examines all stationary points to obtain an MLE), three-ancestry approach Case 1 (which yields a stationary point), and the ASAFE EM algorithm (which yields a stationary point) yield the unique MLE.

10. We have shown in Section 1.3.1 that when, for every ga-allele, there is a count $n_{gg',aa'}$

consistent with carrying that ga-allele with certainty, so that the number X_{ga} of gametes carrying allele g of ancestry a with certainty is positive, an MLE must lie in the interior of the parameter space. Under this condition, if one uses the ASAFE EM algorithm implementation, one needn't worry about missing an MLE on the boundary, because an MLE cannot be there. We illustrated that this condition is sufficient, but not necessary, for an MLE to be in the interior of the domain (Tables 1.8, 1.6).

We showed that on simulated data, under a variety of simulation settings, our two and three-ancestry approaches yielded ancestry-specific allele frequency estimators $\hat{p}(1|a)$ with low bias and variance, since mean errors (bias) and standard deviation of errors were both small (Tables 1.4 and 1.7, Figures 1.2 and 1.9). For three ancestries, we suspect that a smaller grid interval in the three-ancestry approach could further reduce bias.

We demonstrated our two-ancestry method on African American 1000 Genomes data, and found that many cases of our two-ancestry approach were used for at least one SNP (Table 1.9). We also demonstrated our three-ancestry method on Puerto Rican 1000 Genomes data (Table 1.11), and found that all cases of our three-ancestry approach were used for at least one SNP. Our three-ancestry estimates show that the frequency of the 1 allele on African, European, and Native American backgrounds is low for most SNPs (Figure 1.10). Finally, we exemplified the usage of ancestry-specific allele frequencies by using them to estimate population-specific F_{ST}.

1.6 Appendix

1.6.1 URLs to 1000 Genomes Data

For the two-ancestry real data application, we obtained Chr 22 marker genotypes from the 1000 Genomes FTP site at this URL:

`ftp://ftp.1000genomes.ebi.ac.uk/vol1/ftp/phase1/`
`analysis_results/shapeit2_phased_haplotypes/`

```
ALL.chr22.SHAPEIT2_integrated_phase1_v3.
20101123.snps_indels_svs.genotypes.all.vcf.gz
```

For the two-ancestry real data application, we obtained ancestry calls for ASW individuals from the 1000 Genomes FTP site at this URL:

```
ftp://ftp-trace.ncbi.nih.gov/1000genomes/ftp/
phase1/analysis_results/ancestry_deconvolution/
ASW_phase1_ancestry_deconvolution.zip
```

For the three-ancestry real data application, we obtained Chr 22 SNP genotypes from the 1000 Genomes FTP site at this URL:

```
ftp://ftp.1000genomes.ebi.ac.uk/vol1/ftp/phase1/
analysis_results/shapeit2_phased_haplotypes/
ALL.chr22.SHAPEIT2_integrated_phase1_v3.
20101123.snps_indels_svs.genotypes.all.vcf.gz
```

This file contains 494,328 SNPs for 1,099 individuals, including 55 PUR individuals.

For the three-ancestry real data application, we obtained ancestry calls for 55 PUR individuals from the 1000 Genomes FTP site at this URL:

```
ftp://ftp-trace.ncbi.nih.gov/1000genomes/ftp/
phase1/analysis_results/ancestry_deconvolution/
PUR_phase1_ancestry_deconvolution.zip
```

All ancestry calls and genotypes mentioned in this Section were used by The 1000 Genomes Project Consortium (2012).

1.6.2 *When is the Negative Log Likelihood Function Convex, so that a Stationary Point is an MLE?*

The aim of this section is to show conditions on the genotype gg' and ancestry pair aa' counts $n_{gg',aa'}$'s under which the negative log likelihood function $-l(\mathbf{p})$ is a convex function of $\mathbf{p}$ on the interior of the domain D_{int} of the probability simplex domain D of the log likelihood function $l(\mathbf{p})$, so that a stationary point of $l(\mathbf{p})$ is a global maximizer of $l(\mathbf{p})$, i.e. an MLE.

Definition 3.1 of Boyd and Vandenberghe (2004)'s book defines a convex function:

Definition 1.6.1. Convex Function. "A function f: $\mathbf{R}^n \rightarrow \mathbf{R}$ is convex if $\mathbf{dom}f$ is a convex set, and for all $\mathbf{x}, \mathbf{y} \in \mathbf{dom}f$, and θ with $0 \leq \theta \leq 1$, we have $f((\theta)(\mathbf{x}) + (1-\theta)(\mathbf{y})) \leq (\theta)f(\mathbf{x}) + (1-\theta)f(\mathbf{y})$."

If the negative log likelihood function $-l(\mathbf{p})$ is a convex function of $\mathbf{p}$ on the domain D_{int}, then a point $\mathbf{p}^*$ where the gradient of the negative log likelihood function $-l(\mathbf{p})$ evaluates to 0 is not simply a stationary point (so either a local minimizer, saddle point, or local maximizer), but a global minimizer of $-l(\mathbf{p})$. Since $\mathbf{p}^*$ is also a point where the gradient of the log likelihood $l(\mathbf{p})$ evaluates to 0, a local minimizer of $-l(\mathbf{p})$ is a local maximizer of $l(\mathbf{p})$, and a global minimizer of $-l(\mathbf{p})$ is a global maximizer (i.e. MLE) of $l(\mathbf{p})$, we can re-state the last sentence as follows: If the negative log likelihood function $-l(\mathbf{p})$ is a convex function of $\mathbf{p}$ on D_{int}, then a point $\mathbf{p}^*$ where the gradient of the log likelihood function $l(\mathbf{p})$ evaluates to 0 is not simply a stationary point, but a global maximizer of $l(\mathbf{p})$. In other words, if the negative log likelihod function is a convex function of $\mathbf{p}$ on D_{int}, then when solving the system of equations obtained by setting partial derivatives of the log likelihood function with Lagrangian term to zero, or obtaining an EM convergence point using the EM algorithm, any resultant stationary point $\mathbf{p}^*$ in D_{int} is an MLE.

We require further facts about convex functions to show that $-l(\mathbf{p})$ is a convex function of $\mathbf{p}$ under certain conditions on the counts $\{n_{gg',aa'}\}$. Boyd and Vandenberghe give the following Lemma 1.6.1 on page 79:

Lemma 1.6.1. If f_1, ..., f_m are convex functions and w_1, ..., w_m are nonnegative scalars, then the function $f = w_1 f_1 + ... + w_m f_m$ is also a convex function.

Assume for the moment that every $-log(p_{gg',aa'}(\mathbf{p}))$ is a convex function of $\mathbf{p}$ on the interior D_{int} of the domain. Then the negative log likelihood function with constant term removed is $-l(\mathbf{p}) + ln(C) = \sum_{gg'} \sum_{aa'} n_{gg',aa'}(-ln(p_{gg',aa'}(\mathbf{p})))$, a sum of convex functions of $\mathbf{p}$ with nonnegative weights being the counts $\{n_{gg',aa'}\}$. By Lemma 1.6.1, $-l(\mathbf{p}) + ln(C)$ is a convex function of $\mathbf{p}$ on D_{int}.

Recall from mapping 1.1 that the probability $p_{gg',aa'}(\mathbf{p})$ that an individual has genotype gg' and ancestry pair aa' takes the form $(c)(p_{ga}p_{g'a'})$, where $c = 1$ if $g = g'$ and $a = a'$ and $c = 2$ if $g = g'$ and $a \neq a'$, or $g \neq g'$ and $a = a'$, or the form $(2)(p_{ga}p_{g'a'}) + (2)(p_{ga'}p_{g'a})$, if $g \neq g'$ and $a \neq a'$. We will show that when $p_{gg',aa'}(\mathbf{p}) = (c)(p_{ga}p_{g'a'})$, $-ln(p_{gg',aa'}(\mathbf{p}))$ is a convex function of $\mathbf{p}$, but when $p_{gg',aa'}(\mathbf{p}) = (2)(p_{ga}p_{g'a'}) + (2)(p_{ga'}p_{g'a})$, $-ln(p_{gg',aa'}(\mathbf{p}))$ is not a convex function of $\mathbf{p}$.

Consider $-l(\mathbf{p}) + ln(C) = \sum_{gg'aa' \in S_{gg',aa'}} (n_{gg',aa'})(-ln(p_{gg',aa'}(\mathbf{p})))$. If every $n_{gg',aa'}$ count that gets multiplied to a factor $-ln(p_{gg',aa'}(\mathbf{p}))$, where $p_{gg',aa'}(\mathbf{p})$ is of the form $(2)(p_{ga}p_{g'a'}) + (2)(p_{ga'}p_{g'a})$, is 0, then $-l(\mathbf{p}) + ln(C)$ is a nonnegative weighted sum of convex functions $-ln(p_{gg',aa'}(\mathbf{p}))$, so is a convex function of $\mathbf{p}$ on D_{int}. Note that the genotype gg' and ancestry pair aa' combinations that would make $-l(\mathbf{p})$ convex if their counts $n_{gg',aa'}$ were zero are the combinations corresponding to a heterozygous genotype and ancestry pair with differing ancestries. Because this condition on the counts guarantees convexity of the negative log likelihood function, we emphasize its importance as an Assumption:

Assumption 1 (Assumption Sufficient for a Stationary Point to be an MLE)**.** Every $n_{gg',aa'}$ count that, in the log likelihood function, gets multiplied to a factor $-ln(p_{gg',aa'}(\mathbf{p}))$, where $p_{gg',aa'}(\mathbf{p})$ is of the form $(2)(p_{ga}p_{g'a'}) + (2)(p_{ga'}p_{g'a})$, is 0. In other words, every category with heterozygous genotype and ancestry pair with differing ancestries has count $n_{gg',aa'} = 0$.

For two ancestries, this assumption is $n_{01,AE} = 0$. For three ancestries, this assumption is $n_{01,AE} = n_{01,NA} = n_{01,NE} = 0$.

However, if at least one $n_{gg',aa'}$ count that gets multiplied to a $-ln(p_{gg',aa'}(\mathbf{p}))$, where $p_{gg',aa'}(\mathbf{p})$ is of the form $(2)(p_{ga}p_{g'a'})+(2)(p_{ga'}p_{g'a})$, is positive, then $-l(\mathbf{p})$ is not a nonnegative weighted sum of convex functions on D_{int}, so we make no statement about whether $-l(\mathbf{p})$ is a convex function of $\mathbf{p}$ on D_{int} in this case.

It remains to show that when $p_{gg',aa'}(\mathbf{p}) = (c)(p_{ga}p_{g'a'})$, $-ln(p_{gg',aa'}(\mathbf{p}))$ *is* a convex function of $\mathbf{p}$ on the interior D_{int} of the domain, but when $p_{gg',aa'}(\mathbf{p}) = (2)(p_{ga}p_{g'a'}) + (2)(p_{ga'}p_{g'a})$, $-ln(p_{gg',aa'}(\mathbf{p}))$ *is not* a convex function of $\mathbf{p}$ on D_{int}. We will prove these Claims 1 and 2 below.

To prove these two claims, we use these facts stated by Boyd and Vandenberghe (2004): "We now assume that f is twice differentiable, that is, its Hessian or second derivative $\nabla^2 f$ exists at each point in $dom f$, which is open. Then f is convex **if and only if** [emphasis added] $dom f$ is convex and its Hessian is positive semidefinite: **for all** [emphasis added] $x \in dom f$, $\nabla^2 f(x) \succeq 0$. For a function on $\mathbf{R}$, this reduces to the simple condition $f''(x) \geq 0$ (and $\mathbf{dom} f$ convex, i.e., an interval)." Boyd and Vandenberghe (2004) define a positive semidefinite matrix like so, "Suppose $A \in \boldsymbol{S}^n$, i.e., A is a real symmetric n $\times$ n matrix... If A satisfies $x^T Ax \geq 0$ for all x, we say that A is *positive semidefinite*." In Claim 1, we will argue that a function is convex on D_{int} because the function's Hessian is PSD at every point in D_{int}. In Claim 2, we will argue that a function is not convex on D_{int} because there exists a point in D_{int} at which the function's Hessian is not PSD.

Claim 1. $-ln(p_{gg',aa'}(\mathbf{p}))$ is a convex function of $\mathbf{p}$ on the interior of the domain D_{int} when $p_{gg',aa'}(\mathbf{p}) = (c)(p_{ga}p_{g'a'})$, where $c = 1$ if $g = g'$ and $a = a'$, and $c = 2$ if $g \neq g'$ and $a = a'$, or $g = g'$ and $a \neq a'$.

Proof of Claim 1. Consider the function $-ln(p_{gg',aa'}(\mathbf{p})) = -ln((c)(p_{ga}p_{g'a'}))$, where $c = 1$ if $g = g'$ and $a = a'$, and $c = 2$ if $g \neq g'$ and $a = a'$, or $g = g'$ and $a \neq a'$. Without loss of generality, consider the three-ancestry case, where $\mathbf{p} = [p_{0A}, p_{0E}, p_{0N}, p_{1A}, p_{1E}, p_{1N}]$. For two ancestries or more than three ancestries, there would simply be fewer or additional zeroes in vectors and matrices in the following proofs, corresponding to fewer or additional elements of vector $\mathbf{p}$.

First, consider $p_{ga} = p_{g'a'} = p_{0A}$, so $c = 1$. In this case, $-ln(p_{gg',aa'}(\mathbf{p})) = -ln(p_{ga}^2) = -2ln(p_{ga})$. The first derivative is $-2/p_{ga}$. The second derivative is $2/p_{ga}^2 \geq 0$ for all $\mathbf{p}$, so $-ln(p_{gg',aa'}(\mathbf{p}))$ is a convex function.

Next, consider $p_{ga} = p_{0A}$ and $p_{g'a'} = p_{0E}$, so $c = 2$. Then we have:

$$\nabla - ln(p_{gg',aa'}(\mathbf{p})) := \begin{bmatrix} \frac{\partial - ln(cp_{0A}p_{0E})}{\partial p_{0A}} \\ \frac{\partial - ln(cp_{0A}p_{0E})}{\partial p_{0E}} \\ \frac{\partial - ln(cp_{0A}p_{0E})}{\partial p_{0N}} \\ \frac{\partial - ln(cp_{0A}p_{0E})}{\partial p_{1A}} \\ \frac{\partial - ln(cp_{0A}p_{0E})}{\partial p_{1E}} \\ \frac{\partial - ln(cp_{0A}p_{0E})}{\partial p_{1N}} \end{bmatrix} = \begin{bmatrix} (\frac{-1}{cp_{0A}p_{0E}})(cp_{0E}) = -p_{0A}^{-1} \\ (\frac{-1}{cp_{0A}p_{0E}})(cp_{0A}) = -p_{0E}^{-1} \\ 0 \\ 0 \\ 0 \\ 0 \end{bmatrix}$$

and

$$\nabla^2 - ln(p_{gg',aa'}(\mathbf{p})) = \begin{bmatrix} p_{0A}^{-2} & 0 & 0 & 0 & 0 & 0 \\ 0 & p_{0E}^{-2} & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 \end{bmatrix}.$$

Recall the definition for a matrix $\mathbf{A}$ to be positive semidefinite (PSD): The matrix $\mathbf{A}$ is $n \times n$, symmetric, and $\mathbf{x}^T\mathbf{A}\mathbf{x} \geq 0$ for all $\mathbf{x} \in \mathbf{R}^n$. Now take $\mathbf{x} = [x_1, x_2, x_3, x_4, x_5, x_6] \in \mathbf{R}^6$.

$$\begin{aligned} \mathbf{x}^T\nabla^2(-ln(p_{gg',aa'}(\mathbf{p})))\mathbf{x} &= x_1^2 p_{0A}^{-2} + x_2^2 p_{0E}^{-2} \\ &\geq 0. \end{aligned}$$

By definition of PSD, $\nabla^2(-ln(p_{gg',aa'}(\mathbf{p})))$ is a PSD matrix for all $\mathbf{p} \in D_{int}$. Hence, $-ln(p_{gg',aa'}(\mathbf{p}))$ is a convex function of $\mathbf{p}$ on D_{int}.

Claim 2. $-ln(p_{gg',aa'}(\mathbf{p}))$ is not a convex function of $\mathbf{p}$ on the interior D_{int}, when $p_{gg',aa'}(\mathbf{p}) = (2)(p_{ga}p_{g'a'}) + (2)(p_{ga'}p_{g'a})$, where $g \neq g'$ and $a \neq a'$.

Proof of Claim 2. Consider the function $-ln(p_{gg',aa'}(\mathbf{p})) = -ln((2)(p_{ga}p_{g'a'}) + (2)(p_{ga'}p_{g'a}))$, where $g \neq g'$ and $a \neq a'$. Without loss of generality, consider the three-ancestry case, where $\mathbf{p} = [p_{0A}, p_{0E}, p_{0N}, p_{1A}, p_{1E}, p_{1N}]$, and let $p_{ga} = p_{0A}$, $p_{g'a'} = p_{1E}$, $p_{ga'} = p_{0E}$, and $p_{g'a} = p_{1A}$. For two ancestries or more than three ancestries, there would simply be fewer or additional zeroes in vectors and matrices, corresponding to fewer or additional elements of vector $\mathbf{p}$.

We would like to show that $-ln(p_{gg',aa'}(\mathbf{p})) = -ln((2)(p_{0A}p_{1E}) + (2)(p_{0E}p_{1A}))$ is not a convex function of $\mathbf{p}$ on D_{int}. We have:

$$\nabla - ln(p_{gg',aa'}(\mathbf{p})) = \begin{bmatrix} \frac{\partial - ln(p_{gg',aa'}(\mathbf{p}))}{\partial p_{0A}} \\ \frac{\partial - ln(p_{gg',aa'}(\mathbf{p}))}{\partial p_{0E}} \\ \frac{\partial - ln(p_{gg',aa'}(\mathbf{p}))}{\partial p_{0N}} \\ \frac{\partial - ln(p_{gg',aa'}(\mathbf{p}))}{\partial p_{1A}} \\ \frac{\partial - ln(p_{gg',aa'}(\mathbf{p}))}{\partial p_{1E}} \\ \frac{\partial - ln(p_{gg',aa'}(\mathbf{p}))}{\partial p_{1N}} \end{bmatrix} = \begin{bmatrix} -(p_{0A}p_{1E} + p_{0E}p_{1A})^{-1}(p_{1E}) \\ -(p_{0A}p_{1E} + p_{0E}p_{1A})^{-1}(p_{1A}) \\ 0 \\ -(p_{0A}p_{1E} + p_{0E}p_{1A})^{-1}(p_{0E}) \\ -(p_{0A}p_{1E} + p_{0E}p_{1A})^{-1}(p_{0A}) \\ 0 \end{bmatrix}$$

and

$$\nabla^2 - ln(p_{gg',aa'}(\mathbf{p})) = \begin{bmatrix} A_{11} & A_{12} & 0 & A_{13} & A_{14} & 0 \\ A_{21} & A_{22} & 0 & A_{23} & A_{24} & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 \\ A_{31} & A_{32} & 0 & A_{33} & A_{34} & 0 \\ A_{41} & A_{42} & 0 & A_{43} & A_{44} & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 \end{bmatrix}.$$

To save computation in a grid search which we will describe, we define matrix $\boldsymbol{A}(\mathbf{p})$ as

$$\begin{bmatrix} A_{11} & A_{12} & A_{13} & A_{14} \\ A_{21} & A_{22} & A_{23} & A_{24} \\ A_{31} & A_{32} & A_{33} & A_{34} \\ A_{41} & A_{42} & A_{43} & A_{44} \end{bmatrix} = \begin{bmatrix} \frac{p_{1E}^2}{(p_{0A}p_{1E}+p_{0E}p_{1A})^2} & \frac{p_{1A}p_{1E}}{(p_{0A}p_{1E}+p_{0E}p_{1A})^2} & \frac{p_{0E}p_{1E}}{(p_{0A}p_{1E}+p_{0E}p_{1A})^2} & \frac{-p_{0E}p_{1A}}{(p_{0A}p_{1E}+p_{0E}p_{1A})^2} \\ \frac{p_{1A}p_{1E}}{(p_{0A}p_{1E}+p_{0E}p_{1A})^2} & \frac{p_{1A}^2}{(p_{0A}p_{1E}+p_{0E}p_{1A})^2} & \frac{-p_{0A}p_{1E}}{(p_{0A}p_{1E}+p_{0E}p_{1A})^2} & \frac{p_{0A}p_{1A}}{(p_{0A}p_{1E}+p_{0E}p_{1A})^2} \\ \frac{p_{0E}p_{1E}}{(p_{0A}p_{1E}+p_{0E}p_{1A})^2} & \frac{-p_{0A}p_{1E}}{(p_{0A}p_{1E}+p_{0E}p_{1A})^2} & \frac{p_{0E}^2}{(p_{0A}p_{1E}+p_{0E}p_{1A})^2} & \frac{p_{0A}p_{0E}}{(p_{0A}p_{1E}+p_{0E}p_{1A})^2} \\ \frac{-p_{0E}p_{1A}}{(p_{0A}p_{1E}+p_{0E}p_{1A})^2} & \frac{p_{0A}p_{1A}}{(p_{0A}p_{1E}+p_{0E}p_{1A})^2} & \frac{p_{0A}p_{0E}}{(p_{0A}p_{1E}+p_{0E}p_{1A})^2} & \frac{p_{0A}^2}{(p_{0A}p_{1E}+p_{0E}p_{1A})^2} \end{bmatrix}$$

$$= \frac{1}{(p_{0A}p_{1E}+p_{0E}p_{1A})^2} \begin{bmatrix} p_{1E}^2 & p_{1A}p_{1E} & p_{0E}p_{1E} & -p_{0E}p_{1A} \\ p_{1A}p_{1E} & p_{1A}^2 & -p_{0A}p_{1E} & p_{0A}p_{1A} \\ p_{0E}p_{1E} & -p_{0A}p_{1E} & p_{0E}^2 & p_{0A}p_{0E} \\ -p_{0E}p_{1A} & p_{0A}p_{1A} & p_{0A}p_{0E} & p_{0A}^2 \end{bmatrix}.$$

Because for any $\mathbf{x} = [x_1, x_2, x_3, x_4, x_5, x_6] \in \mathbf{R}^6$ and $\mathbf{y} = [x_1, x_2, x_4, x_5]$, $\mathbf{x}^T\ \nabla^2(-ln(p_{gg',aa'}(\mathbf{p})))\mathbf{x}$ $= \mathbf{y}^T\mathbf{A}(\mathbf{p})\mathbf{y}$, the Hessian matrix $\nabla^2(-ln(p_{gg',aa'}(\mathbf{p})))$ is PSD at a point $\mathbf{p}^*$ if and only if its submatrix $\mathbf{A}(\mathbf{p})$ is PSD at this point.

We performed a grid search of the domain $D_{int} := \{\mathbf{p} \in (0,1)\times(0,1)\times(0,1)\times(0,1)\times(0,1)\times(0,1)\}$ such that $p_{0A}+p_{0E}+p_{0N}+p_{1A}+p_{1E}+p_{1N} = 1\}$. For each point on the grid, we constructed the matrix $(p_{0A}p_{1E}+p_{0E}p_{1A})^2\boldsymbol{A}(\mathbf{p})$:

$$\begin{bmatrix} p_{1E}^2 & p_{1A}p_{1E} & p_{0E}p_{1E} & -p_{0E}p_{1A} \\ p_{1A}p_{1E} & p_{1A}^2 & -p_{0A}p_{1E} & p_{0A}p_{1A} \\ p_{0E}p_{1E} & -p_{0A}p_{1E} & p_{0E}^2 & p_{0A}p_{0E} \\ -p_{0E}p_{1A} & p_{0A}p_{1A} & p_{0A}p_{0E} & p_{0A}^2 \end{bmatrix}.$$

For any $\mathbf{p} \in D_{int}$, the matrix $(p_{0A}p_{1E}+p_{0E}p_{1A})^2\boldsymbol{A}(\mathbf{p})$ is PSD if and only if matrix $\mathbf{A}(\mathbf{p})$ is PSD. Our grid search yielded many points $\mathbf{p} \in D_{int}$ where matrix $(p_{0A}p_{1E}+p_{0E}p_{1A})^2\boldsymbol{A}(\mathbf{p})$ was not PSD. One such point was $\mathbf{p}^* = [0.1, 0.1, 0.1, 0.1, 0.1, 0.5]$. The eigen() function in R computed eigenvalues of matrix $(p_{0A}p_{1E}+p_{0E}p_{1A})^2\boldsymbol{A}(\mathbf{p}^*)$ to be 0.02, 0.02, 0.02, and -0.02. Since one of these eigenvalues was negative, the matrix $(p_{0A}p_{1E}+p_{0E}p_{1A})^2\boldsymbol{A}(\mathbf{p})$ was not PSD at this particular point $\mathbf{p}^*$ in D_{int}, i.e. $(p_{0A}p_{1E}+p_{0E}p_{1A})^2\boldsymbol{A}(\mathbf{p}^*)$ was not PSD, so $\mathbf{A}(\mathbf{p}^*)$ was not PSD.

Since we just showed that $\mathbf{A}(\mathbf{p})$ is not PSD for a particular $\mathbf{p}^* = [0.1, 0.1, 0.1, 0.1, 0.1, 0.5]$, we conclude that $\nabla^2(-ln(p_{gg',aa'}(\mathbf{p})))$ is also not PSD when evaluated at this $\mathbf{p}^*$, so $\nabla^2(-ln(p_{gg',aa'}(\mathbf{p})))$ is not PSD for all $\mathbf{p}$ in the interior D_{int}, so $-ln(p_{gg',aa'}(\mathbf{p}))$ is not a convex function on D_{int}.

1.6.3 Observed Fisher Information-Based SEs for the Case of Two Ancestries

We use a SE estimator based on the observed Fisher information (Efron and Hinkley, 1978). We provide derivations of information-based SE formulas in this section.

1.6.3.1 Log Likelihood Function

Recall, the log likelihood function is the following, as a function of p_{0A}, p_{0E}, p_{1A}, and p_{1E}:

$$
\begin{aligned}
l(\mathbf{p}) := &(n_{00,AA})ln(p_{0A}^2) + (n_{00,AE})ln(2p_{0E}p_{0A})\\
&+ (n_{00,EE})ln(p_{0E}^2) + (n_{01,AA})ln(2p_{1A}p_{0A})\\
&+ (n_{01,AE})ln(2p_{1E}p_{0A} + 2p_{1A}p_{0E})\\
&+ (n_{01,EE})ln(2p_{1E}p_{0E}) + (n_{11,AA})ln(p_{1A}^2)\\
&+ (n_{11,AE})ln(2p_{1E}p_{1A}) + (n_{11,EE})ln(p_{1E}^2)\\
&= (X_{0A})ln(p_{0A}) + (X_{0E})ln(p_{0E})\\
&+ (X_{1A})ln(p_{1A}) + (X_{1E})ln(p_{1E})\\
&+ (n_{01,AE})ln(2p_{1E}p_{0A} + 2p_{1A}p_{0E})\\
&+ (n_{00,AE} + n_{01,AA} + n_{01,EE} + n_{11,AE})ln(2),
\end{aligned}
$$

where

$$X_{0A} = (2n_{00,AA} + n_{00,AE} + n_{01,AA})$$
$$X_{0E} = (2n_{00,EE} + n_{00,AE} + n_{01,EE})$$
$$X_{1A} = (2n_{11,AA} + n_{01,AA} + n_{11,AE})$$
$$X_{1E} = (2n_{11,EE} + n_{01,EE} + n_{11,AE}).$$

Since the $\{p_{ga}\}$ sum to 1, we express p_{1E} as $1 - p_{0A} - p_{0E} - p_{1A}$, so that the log likelihood function is a function of three independent parameters:

$$\begin{aligned}
l(\vec{p}) &= (X_{0A})ln(p_{0A}) + (X_{0E})ln(p_{0E}) \\
&+ (X_{1A})ln(p_{1A}) + (X_{1E})ln(1 - p_{0A} - p_{0E} - p_{1A}) \\
&+ (n_{01,AE})ln(2(1 - p_{0A} - p_{0E} - p_{1A})p_{0A} + 2p_{1A}p_{0E}) \\
&+ (n_{00,AE} + n_{01,AA} + n_{01,EE} + n_{11,AE})ln(2) \\
&= (X_{0A})ln(p_{0A}) + (X_{0E})ln(p_{0E}) \\
&+ (X_{1A})ln(p_{1A}) + (X_{1E})ln(1 - p_{0A} - p_{0E} - p_{1A}) \\
&+ (n_{01,AE})ln((2p_{0A} - 2p_{0A}^2 - 2p_{0A}p_{0E} - 2p_{0A}p_{1A}) + 2p_{1A}p_{0E}) \\
&+ (n_{00,AE} + n_{01,AA} + n_{01,EE} + n_{11,AE})ln(2).
\end{aligned}$$

1.6.3.2 First Derivatives of the Log Likelihood Function

The first derivatives of the log likelihood function are

$$\dot{l}(\mathbf{p}) := [\frac{\partial l(\mathbf{p})}{\partial p_{0A}}, \frac{\partial l(\mathbf{p})}{\partial p_{0E}}, \frac{\partial l(\mathbf{p})}{\partial p_{1A}}]$$

where

$$\frac{\partial l(\mathbf{p})}{\partial p_{0A}} = (X_{0A})(p_{0A})^{-1}$$
$$- (X_{1E})(1 - p_{0A} - p_{0E} - p_{1A})^{-1}$$
$$+ (n_{01,AE})(2p_{0A} - 2p_{0A}^2 - 2p_{0A}p_{0E} - 2p_{0A}p_{1A} + 2p_{1A}p_{0E})^{-1}(2 - 4p_{0A} - 2p_{0E} - 2p_{1A}),$$
$$\frac{\partial l(\mathbf{p})}{\partial p_{0E}} = (X_{0E})(p_{0E})^{-1}$$
$$- (X_{1E})(1 - p_{0A} - p_{0E} - p_{1A})^{-1}$$
$$+ (n_{01,AE})(2p_{0A} - 2p_{0A}^2 - 2p_{0A}p_{0E} - 2p_{0A}p_{1A} + 2p_{1A}p_{0E})^{-1}(-2p_{0A} + 2p_{1A}), \text{ and}$$
$$\frac{\partial l(\mathbf{p})}{\partial p_{1A}} = (X_{1A})(p_{1A})^{-1}$$
$$- (X_{1E})(1 - p_{0A} - p_{0E} - p_{1A})^{-1}$$
$$+ (n_{01,AE})(-2p_{0A} + 2p_{0E})(2p_{0A} - 2p_{0A}^2 - 2p_{0A}p_{0E} - 2p_{0A}p_{1A} + 2p_{1A}p_{0E})^{-1}.$$

1.6.3.3 Observed Fisher Information, and Second Derivatives of the Log Likelihood Function

The observed Fisher information matrix is $-\ddot{l}(\mathbf{p})|_{\hat{\mathbf{p}}}$, the matrix of negative second derivatives of the log likelihood function evaluated at the MLE $\hat{\mathbf{p}}$. The second derivative matrix is symmetric:

$$\begin{bmatrix} \frac{\partial^2 l(\mathbf{p})}{\partial p_{0A}\partial p_{0A}} & \frac{\partial^2 l(\mathbf{p})}{\partial p_{0A}\partial p_{0E}} & \frac{\partial^2 l(\mathbf{p})}{\partial p_{0A}\partial p_{1A}} \\ \frac{\partial^2 l(\mathbf{p})}{\partial p_{0E}\partial p_{0A}} & \frac{\partial^2 l(\mathbf{p})}{\partial p_{0E}\partial p_{0E}} & \frac{\partial^2 l(\mathbf{p})}{\partial p_{0E}\partial p_{1A}} \\ \frac{\partial^2 l(\mathbf{p})}{\partial p_{1A}\partial p_{0A}} & \frac{\partial^2 l(\mathbf{p})}{\partial p_{1A}\partial p_{0E}} & \frac{\partial^2 l(\mathbf{p})}{\partial p_{1A}\partial p_{1A}} \end{bmatrix} = \begin{bmatrix} \frac{\partial^2 l(\mathbf{p})}{\partial p_{0A}\partial p_{0A}} & \frac{\partial^2 l(\mathbf{p})}{\partial p_{0A}\partial p_{0E}} & \frac{\partial^2 l(\mathbf{p})}{\partial p_{0A}\partial p_{1A}} \\ \ldots & \frac{\partial^2 l(\mathbf{p})}{\partial p_{0E}\partial p_{0E}} & \frac{\partial^2 l(\mathbf{p})}{\partial p_{0E}\partial p_{1A}} \\ \ldots & \ldots & \frac{\partial^2 l(\mathbf{p})}{\partial p_{1A}\partial p_{1A}} \end{bmatrix}.$$

Hence, we only need to compute six second derivatives:

$$\frac{\partial^2 l(\mathbf{p})}{\partial^2 p_{0A}} = (-1)(X_{0A})(p_{0A})^{-2}$$
$$+ (X_{1E})(1 - p_{0A} - p_{0E} - p_{1A})^{-2}(-1)$$
$$+ (n_{01,AE})\Big[[(-1)(2p_{0A} - 2p_{0A}^2 - 2p_{0A}p_{0E} - 2p_{0A}p_{1A} + 2p_{1A}p_{0E})^{-2}][2 - 4p_{0A} - 2p_{0E} - 2p_{1A}]^2$$
$$+ [(2p_{0A} - 2p_{0A}^2 - 2p_{0A}p_{0E} - 2p_{0A}p_{1A} + 2p_{1A}p_{0E})^{-1}][-4]\Big],$$

$$\frac{\partial^2 l(\mathbf{p})}{\partial p_{0A}\partial p_{0E}} = -(X_{1E})(1 - p_{0A} - p_{0E} - p_{1A})^{-2}$$
$$+ (n_{01,AE})\Big[[(-1)(2p_{0A} - 2p_{0A}^2 - 2p_{0A}p_{0E} - 2p_{0A}p_{1A} + 2p_{1A}p_{0E})^{-2}(-2p_{0A} + 2p_{1A})]$$
$$[2 - 4p_{0A} - 2p_{0E} - 2p_{1A}] + [2p_{0A} - 2p_{0A}^2 - 2p_{0A}p_{0E} - 2p_{0A}p_{1A} + 2p_{1A}p_{0E}]^{-1}[-2]\Big],$$

$$\frac{\partial^2 l(\mathbf{p})}{\partial p_{0A}\partial p_{1A}} = -(X_{1E})(1 - p_{0A} - p_{0E} - p_{1A})^{-2}$$
$$+ (n_{01,AE})$$
$$\Big[[(-1)(2p_{0A} - 2p_{0A}^2 - 2p_{0A}p_{0E} - 2p_{0A}p_{1A} + 2p_{1A}p_{0E})^{-2}(-2p_{0A} + 2p_{0E})]$$
$$[(2 - 4p_{0A} - 2p_{0E} - 2p_{1A})]$$
$$+ [(2p_{0A} - 2p_{0A}^2 - 2p_{0A}p_{0E} - 2p_{0A}p_{1A} + 2p_{1A}p_{0E})^{-1}][-2]\Big],$$

$$\frac{\partial^2 l(\mathbf{p})}{\partial p_{0E}\partial p_{0E}} = -(X_{0E})(p_{0E})^{-2}$$
$$- (X_{1E})(1 - p_{0A} - p_{0E} - p_{1A})^{-2}$$
$$+ (n_{01,AE})\Big[[(-1)(2p_{0A} - 2p_{0A}^2 - 2p_{0A}p_{0E} - 2p_{0A}p_{1A} + 2p_{1A}p_{0E})^{-2}(-2p_{0A} + 2p_{1A})]$$
$$(-2p_{0A} + 2p_{1A})\Big],$$

$$\frac{\partial^2 l(\mathbf{p})}{\partial p_{0E}\partial p_{1A}} = (-X_{1E})(1-p_{0A}-p_{0E}-p_{1A})^{-2}$$
$$+ (n_{01,AE})\Big[[(-1)(2p_{0A}-2p_{0A}^2-2p_{0A}p_{0E}-2p_{0A}p_{1A}+2p_{1A}p_{0E})^{-2}(-2p_{0A}+2p_{0E})]$$
$$[-2p_{0A}+2p_{1A}]$$
$$+ [(2p_{0A}-2p_{0A}^2-2p_{0A}p_{0E}-2p_{0A}p_{1A}+2p_{1A}p_{0E})^{-1}][2]\Big], \text{ and}$$

$$\frac{\partial^2 l(\mathbf{p})}{\partial p_{1A}\partial p_{1A}} = (-1)(X_{1A})(p_{1A})^{-2}$$
$$- (X_{1E})(1-p_{0A}-p_{0E}-p_{1A})^{-2}$$
$$+ (n_{01,AE})(-2p_{0A}+2p_{0E})(-1)(2p_{0A}-2p_{0A}^2-2p_{0A}p_{0E}-2p_{0A}p_{1A}+2p_{1A}p_{0E})^{-2}(-2p_{0A}+2p_{0E}).$$

We checked expressions for second derivatives by deriving expressions multiple times by hand and comparing those expressions to expressions obtained with Mathematica.

The observed Fisher information based on n individuals' data is

$$I_n([\hat{p}_{0A},\hat{p}_{0E},\hat{p}_{1A}]) := -\ddot{l}([p_{0A},p_{0E},p_{1A}])|_{[\hat{p}_{0A},\hat{p}_{0E},\hat{p}_{1A}]}$$
$$= -\begin{bmatrix} \frac{\partial^2 l(\vec{p})}{\partial p_{0A}\partial p_{0A}} & \frac{\partial^2 l(\vec{p})}{\partial p_{0A}\partial p_{0E}} & \frac{\partial^2 l(\vec{p})}{\partial p_{0A}\partial p_{1A}} \\ \cdots & \frac{\partial^2 l(\vec{p})}{\partial p_{0E}\partial p_{0E}} & \frac{\partial^2 l(\vec{p})}{\partial p_{0E}\partial p_{1A}} \\ \cdots & \cdots & \frac{\partial^2 l(\vec{p})}{\partial p_{1A}\partial p_{1A}} \end{bmatrix} |_{[p_{0A},p_{0E},p_{1A}]=[\hat{p}_{0A},\hat{p}_{0E},\hat{p}_{1A}]}.$$

1.6.3.4 Employing Asymptotic Normality of the MLE $[\hat{p}_{0A},\hat{p}_{0E},\hat{p}_{1A}]$ to Derive $Var(\hat{P}(1|a)), a \in \{A,E\}$

Standard theory about asymptotic normality of the MLE assumes that the true parameter is in the interior of the parameter space (e.g. theorems concluding asymptotic normality in Sections 5.3, 5.5, and 5.6 of Vaart (1998)'s book). For example, Theorem 5.23 of Vaart

(1998)'s book assumes the true parameter θ_0 is in "an open subset of Euclidean space" and Theorem 5.39 assumes the true parameter θ_0 is an "inner point" of parameter space $\Theta \subset \mathbb{R}^k$.

Applying standard theory, for an interior true parameter $[p_{0A}, p_{0E}, p_{1A}]$, asymptotic normality of the MLE means $\sqrt{n}([\hat{p}_{0A}, \hat{p}_{0E}, \hat{p}_{1A}]$ - $[p_{0A}, p_{0E}, p_{1A}]) \rightarrow_d N([0,0,0], I([p_{0A}, p_{0E}, p_{1A}]))$, where $I([p_{0A}, p_{0E}, p_{1A}])$ is the Fisher information $-E(\ddot{l}([p_{0A}, p_{0E}, p_{1A}]))$ for one observation. This convergence in distribution means the MLE $[\hat{p}_{0A}, \hat{p}_{0E}, \hat{p}_{1A}]$ is approximately distributed $N([p_{0A}, p_{0E}, p_{1A}], I_n([p_{0A}, p_{0E}, p_{1A}])^{-1})$, where $I_n([p_{0A}, p_{0E}, p_{1A}]) = nI([p_{0A}, p_{0E}, p_{1A}])$, for every $[p_{0A}, p_{0E}, p_{1A}]$ and a large number n of individuals (page 63 of Vaart (1998)'s book). Hence, $Var([\hat{p}_{0A}, \hat{p}_{0E}, \hat{p}_{1A}]) \approx I_n([p_{0A}, p_{0E}, p_{1A}])^{-1}$. We estimate $I_n([p_{0A}, p_{0E}, p_{1A}])$ with the observed Fisher information $I_n([\hat{p}_{0A}, \hat{p}_{0E}, \hat{p}_{1A}])$ given in Section 1.6.3.3, obtaining estimator $\hat{Var}([\hat{p}_{0A}, \hat{p}_{0E}, \hat{p}_{1A}]) \approx I_n([\hat{p}_{0A}, \hat{p}_{0E}, \hat{p}_{1A}])^{-1}$.

Perhaps $\sqrt{n}([\widehat{logit(p_{0A})}, \widehat{logit(p_{0E})}, \widehat{logit(p_{1A})}]$ - $[logit(p_{0A}), logit(p_{0E}), logit(p_{1A})])$ converges to a normal distribution faster than $\sqrt{n}([\widehat{p}_{0A}, \widehat{p}_{0E}, \widehat{p}_{1A}]$ - $[p_{0A}, p_{0E}, p_{1A}])$, because each $\widehat{logit(p)}_{ga}$ takes values on the whole real line. We did not derive the asymptotic distribution of $\sqrt{n}([\widehat{logit(p_{0A})}, \widehat{logit(p_{0E})}, \widehat{logit(p_{1A})}]$ - $[logit(p_{0A}), logit(p_{0E}), logit(p_{1A})])$, because we were more interested in obtaining standard error estimates for MLEs on the p_{ga} scale.

To compute $I_n([\hat{p}_{0A}, \hat{p}_{0E}, \hat{p}_{1A}])^{-1}$, we implemented an analytical formula for the inverse of a 3 x 3 matrix. For a 3 x 3 matrix

$$\begin{bmatrix} a & b & c \\ b & d & e \\ c & e & f \end{bmatrix},$$

the inverse matrix is

$$\frac{1}{-adf + ae^2 + b^2 f - 2bce + c^2 d} \begin{bmatrix} e^2 - df & bf - ce & cd - be \\ bf - ce & c^2 - af & ae - bc \\ cd - be & ae - bc & b^2 - ad \end{bmatrix}.$$

Alternatively, we could have used R's solve function to compute a matrix inverse.

To obtain a variance estimator for $\hat{p}(1|a) = g(\hat{p}_{0a}, \hat{p}_{1a}) := \frac{\hat{p}_{1a}}{\hat{p}_{0a}+\hat{p}_{1a}}$, we employ Theorem 1.3 of Colley (2012)'s book to approximate $\hat{p}(1|a)$ by its first order Taylor polynomial about $[E(\hat{p}_{0a}), E(\hat{p}_{1a})]$: $\hat{p}(1|a) \approx E(\hat{p}_{1a})/\Big(E(\hat{p}_{0a}) + E(\hat{p}_{1a})\Big) + c\Big(\hat{p}_{0a} - E(\hat{p}_{0a})\Big) + d\Big(\hat{p}_{1a} - E(\hat{p}_{1a})\Big)$, where we define

$$
\begin{aligned}
c &:= \Big(- \frac{E(\hat{p}_{1a})}{(E(\hat{p}_{0a}) + E(\hat{p}_{1a}))^2}\Big) \text{ and} \\
d &:= \Big(\frac{1}{E(\hat{p}_{0a}) + E(\hat{p}_{1a})} - \frac{E(\hat{p}_{1a})}{(E(\hat{p}_{0a}) + E(\hat{p}_{1a}))^2}\Big).
\end{aligned}
$$

It follows that

$$
\begin{aligned}
Var(\hat{p}(1|a)) &\approx Var\Big[c\Big(\hat{p}_{0a}\Big) + d\Big(\hat{p}_{1a}\Big)\Big] \\
&= Cov\Big[c\hat{p}_{0a} + d\hat{p}_{1a}, c\hat{p}_{0a} + d\hat{p}_{1a}\Big] \\
&= c^2 Var(\hat{p}_{0a}) + 2cdCov(\hat{p}_{0a}, \hat{p}_{1a}) + d^2 Var(\hat{p}_{1a}), \text{for } a \in \{A, E\}.
\end{aligned}
$$

By asymptotic normality of the MLE, when the sample size is large and the true parameter $[p_{0A}, p_{0E}, p_{1A}]$ is in the interior of the parameter space, for each $g \in \{0, 1\}$ and $a \in \{A, E\}$, $E(\hat{p}_{ga}) \approx p_{ga}$, an unknown parameter which we estimate with $\hat{p}_{ga}$. We obtain numeric values for c and d by substituting $\hat{p}_{ga}$ in for $E(\hat{p}_{ga})$ in expressions for c and d. Also note that we derived the variance-covariance matrix for vector $[\hat{p}_{0A}, \hat{p}_{0E}, \hat{p}_{1A}]$, which does not include $\hat{p}_{1E}$. To obtain $Cov(\hat{p}_{0E}, \hat{p}_{1E})$ and $Var(\hat{p}_{1E})$, consider that $\hat{p}_{1E} = 1 - \hat{p}_{0A} - \hat{p}_{0E} - \hat{p}_{1A}$. Then

$$
\begin{aligned}
Cov(\hat{p}_{0E}, \hat{p}_{1E}) &= Cov(\hat{p}_{0E}, 1 - \hat{p}_{0A} - \hat{p}_{0E} - \hat{p}_{1A}) \\
&= -Cov(\hat{p}_{0E}, \hat{p}_{0A}) - Cov(\hat{p}_{0E}, \hat{p}_{0E}) - Cov(\hat{p}_{0E}, \hat{p}_{1A}),
\end{aligned}
$$

and

$$
\begin{aligned}
Var(\hat{p}_{1E}) &= Var(1 - \hat{p}_{0A} - \hat{p}_{0E} - \hat{p}_{1A}) \\
&= Cov(1 - \hat{p}_{0A} - \hat{p}_{0E} - \hat{p}_{1A}, 1 - \hat{p}_{0A} - \hat{p}_{0E} - \hat{p}_{1A}) \\
&= Cov(-\hat{p}_{0A} - \hat{p}_{0E} - \hat{p}_{1A}, -\hat{p}_{0A} - \hat{p}_{0E} - \hat{p}_{1A}) \\
&= Cov(\hat{p}_{0A}, \hat{p}_{0A}) + Cov(\hat{p}_{0A}, \hat{p}_{0E}) + Cov(\hat{p}_{0A}, \hat{p}_{1A}) \\
&+ Cov(\hat{p}_{0E}, \hat{p}_{0A}) + Cov(\hat{p}_{0E}, \hat{p}_{0E}) + Cov(\hat{p}_{0E}, \hat{p}_{1A}) \\
&+ Cov(\hat{p}_{1A}, \hat{p}_{0A}) + Cov(\hat{p}_{1A}, \hat{p}_{0E}) + Cov(\hat{p}_{1A}, \hat{p}_{1A}).
\end{aligned}
$$

1.6.4 Bootstrapping Over Loci

Population-i-specific F_{ST} $\beta_i = (\theta_i - \theta_B)/(1 - \theta_B)$ is a function of coancestry coefficients θ_i and θ_B. The population-i-specific F_{ST} MME $\hat{\beta}_i$ is a function of population-i, $i \in \{1, ..., r\}$ sample allele frequencies $\{\tilde{p}_{il}\}$ at loci $l \in \{1, ..., L\}$.

There are two sources of variation in $\tilde{p}_{il}$, genetic sampling - repetitions of the evolutionary process generate, from an ancestral population, different individuals' alleles in populations $i \in \{1, ..., r\}$ - and statistical sampling - repeated sampling of n individuals from a population generates different sampled individuals' alleles.

Others have previously given that $Var(\tilde{p}_{il}) = (\pi_l)(1 - \pi_l)(p_l)(1 - p_l)\Big[\theta_i + (1 - \theta_i)/(2n_i)\Big]$ (Buckleton et al., 2016), where π_l is the reference allele frequency in an ancestral population at locus l, and n_i is the number of individuals sampled from population i. The first term arises from genetic sampling and the second term arises from statistical sampling. When the sample of n_i individuals from a population i of individuals is large ($n_i \rightarrow \infty$), genetic sampling dominates statistical sampling.

We considered each locus l's vector $[\hat{p}(1|A)_l, \hat{p}(1|E)_l, \hat{p}(1|N)_l]$ of population $i \in \{A, E, N\}$ allele 1 frequency estimates to be an evolutionary replicate. In the context of three ancestries A, E, and N, $\tilde{p}_{il} \in \{\hat{p}(1|A)_l, \hat{p}(1|E)_l, \hat{p}(1|N)_l\}$. We bootstrapped over loci's vectors to capture variation in $\tilde{p}_{il}$ (and so variation in $\hat{\beta}_i$ which is a function of $\tilde{p}_{il}$'s) due to genetic sampling.

We did not bootstrap over individuals because statistical sampling contributes negligibly to $Var(\tilde{p}_{il})$.

Furthermore, resampling strategies for coancestry coefficients were discussed by Dodds (1986), who found that bootstrapping over loci yielded better estimates of a coancestry coefficient estimator's variance than bootstrapping over individuals: "The performance of the resampling techniques for an estimator of the coancestry coefficient is investigated using a simulation study. It is found that resampling independent loci (genetic sampling) yields good estimates of the variance, whereas resampling individuals (statistical sampling) yields poor estimates." Since $\beta_i := (\theta_i - \theta_B)/(1 - \theta_B)$, defined as two coancestry coefficients θ_i and θ_B relative to each other, is a coancestry coefficient θ_i when $\theta_B = 0$, we expected that bootstrapping over loci would similarly be preferable to bootstrapping over individuals.

Chapter 2

ESTIMATION OF INDIVIDUAL INBREEDING COEFFICIENTS

2.1 Abstract

Many population genetic activities, ranging from evolutionary studies to association mapping to forensic identification, rely on estimates of individual inbreeding coefficients and between-individual coancestry coefficients. We examined existing method of moments estimators (MMEs) and maximum likelihood estimators (MLEs) introduced as estimators of an individual j's total inbreeding coefficient F_j (i.e. Hall et al.'s simple method of moments estimator, GCTA^{I} MME, GCTA^{III} MME, and Hall et al.'s first maximum likelihood estimator) or as an estimator of within-population inbreeding coefficient β_j (i.e. Weir and Goudet's MME), as well as our novel likelihood estimator of β_j. For each MME, we considered the estimator that combines over loci using the ratio of means (RM) method (a weighted estimator), and the mean of ratios (MR) method (an unweighted estimator). We bootstrapped over loci to provide interval estimates for MMEs and MLEs. We evaluated all methods, except Hall et al.'s simple MME, on simulated data. We found that, despite previously published claims of unbiasedness for some evaluated estimators, none of these estimators were unbiased for F_j, and the Weir and Goudet MME RM had very low bias for β_j. We showed that evaluated estimators were better at estimating certain non-F_j parameters (which we refer to as "pedigree parameters" or "parametric pedigree values") than at estimating F_j, in that they had lower squared bias and lower MSE for these other parameters than for F_j, and tracked more closely with these other parameters than with F_j. In other words, estimators intended to estimate F_j have been mischaracterized as F_j estimators, and are actually better at estimating other parameters. Lastly, we recommend the Weir and Goudet MME RM for

estimation of β_j.

2.2 *Introduction*

2.2.1 Notation, Definitions, and Basic Model

Chapter 2 addresses inbreeding estimation of individuals within a single population, and it stresses the distinction between within-population and total inbreeding. Unless the context specifies otherwise, we use "allele" to mean an allele copy in an individual (so for example, a diploid individual has two alleles), "reference allele" or "alternative allele" to mean an allelic type, the particular value an allele copy can take (e.g. nucleotides A, C, T, or G), and "allele frequency" to mean the relative frequency of an allelic type amongst some collection of individuals' alleles.

We consider n individuals randomly sampled with equal probabilities from each of r populations (via a process we call statistical sampling). These r populations have descended from an ancestral reference population (via a process we call genetic sampling). We use "ancestral population" and "ancestral reference population" synonymously. Figure 2.1 depicts the ancestral population, genetic sampling to obtain r descendent time-t populations, and statistical sampling to obtain samples from a population. In the ancestral population, the reference allele frequency is π. In a population descended from the ancestral population, the reference allele frequency is p. In a random sample of alleles from a descended population, the reference allele frequency is $\tilde{p}$. We subscript these quantities with l to refer to frequencies at a particular locus l.

Following the notation of Weir (1996) (e.g. page 40), it is convenient to indicate the allelic type of allele k ($k = 1, 2$) for individual j in a population by x_{jk}. We define $x_{jk} = 1$ if that allele is the designated type, often called the reference type, and $x_{jk} = 0$ otherwise. The reference allele dosage X is the sum of the x's for the two alleles carried by an individual. The product $x_{jk}x_{jk'}, k \neq k'$ will be 1 when the individual is homozygous for the designated allele type, and taking expected value $\mathcal{E}_W$ over individuals within (W) the population, we have

$$\mathcal{E}_W(x_{jk}x_{jk'}) \;=\; p^2 + p(1-p)f. \tag{2.1}$$

Here, f is defined as the within-population inbreeding coefficient and p is the proportion of the designated allele type in the population. We can refer to p as the actual population proportion, for the population under study, of the designated allele type. The quantity f is not a probability, and it can be negative. It is the intraclass correlation coefficient of the indicator variables $x_{jk}, x_{jk'}, k \neq k'$ since $\mathcal{E}_W(x_{jk}) = p$ and $\mathrm{Var}_W(x_{jk}) = p(1-p)$.

Inbreeding is generally taken to mean identity by descent, meaning that the two alleles an individual has at a locus have a single ancestral allele in a reference population. There is variation in actual identity by descent for pairs of alleles between (B) replicates of the evolutionary process from the reference to the current population, and also variation in actual allele proportions over replicate samples from the current population. The between-replicate variance of actual allele proportions is

$$\mathrm{Var}_B(p) \;=\; \pi(1-\pi)\theta,$$

so, as discussed by Weir and Hill (2002),

$$\mathcal{E}_B(p^2) = \pi^2 + \pi(1-\pi)\theta, \tag{2.2}$$

and π is $\mathcal{E}_B(p)$. The quantity θ is positive and can be regarded as the probability of identity by descent of any two alleles in the population.

Making use of Equation 2.2, the total (T) expectation of Equation 2.1 is

$$
\begin{aligned}
\mathcal{E}_T(x_{jk}x_{jk'}) &= \mathcal{E}_B[\mathcal{E}_W(x_{jk}x_{jk'})] \\
&= \mathcal{E}_B[p^2 + p(1-p)(f)] \\
&= \pi^2 + \pi(1-\pi)[1-(1-f)(1-\theta)] \\
&= \pi^2 + \pi(1-\pi)F.
\end{aligned}
$$

to introduce the total inbreeding coefficient F by $(1-F) = (1-f)(1-\theta)$ as described by Wright (1922). This equation $(1-F) = (1-f)(1-\theta)$ implies $f = (F-\theta)/(1-\theta)$. As this coefficient F reflects variation among evolutionary replicates of the current population, the "genetic sampling" of Weir (1996), it cannot be estimated from data in a single population.

This discussion of within-population and total inbreeding coefficients can be refined to apply to individuals: the average of individual-specific within-population inbreeding coefficients f_j is f, and the average of individual-specific total inbreeding coefficients F_j is F. The f, F notation was employed by Cockerham (1973) and has been used since then by Weir, e.g. (Weir, 1996). Cockerham and Weir (1986) used β for quantities such as $(F-\theta)/(1-\theta)$ and this was also used by Weir and Goudet (2017). In this dissertation we will continue using β: in this chapter $\beta_j = f_j$ is the within-population inbreeding coefficient for individual j and it along with F_j are the parameters of interest.

Values of F_j and β_j can be calculated from pedigrees by path-counting (Wright, 1922). The total inbreeding coefficient F_j is $\sum(0.5)^n(1+F_J)$, where n is the number of individuals in the pedigree linking the parents of j to their common ancestor J, whose total inbreeding coefficient was F_J. The sum is taken over common ancestors and pathways to each common ancestor. The same procedure provides the coancestry, or kinship, coefficient $\theta_{jj'}$ for pairs of individuals j and j': the probability an allele in one individual j and an allele in another individual j' are identical by descent (ibd), i.e. descended from the same allele copy in an ancestral population. The quantity $\theta_S = \sum_{j=1}^{n}\sum_{j'=1,j'\neq j}^{n}(\theta_{jj'})/(n^2-n)$ is the average coancestry coefficient, averaging $\theta_{jj'}$'s between distinct individuals j and j', where both individuals come from a sample of n individuals from a population.

To summarize, this chapter is concerned with estimating F_j and β_j for individual j in some population, where

$$\beta_j \quad = \quad \frac{F_j - \theta_S}{1 - \theta_S}.$$

Values of these parameters for individuals in a sample from a population can be calculated from the pedigrees of those individuals: F_j is the total inbreeding coefficient for individual j and θ_S is the average coancestry coefficient for all pairs of individuals in the sample. We will compare several estimators introduced as estimators of F_j or β_j, and point out that several of these have been mischaracterized as being unbiased estimators of F_j, which we ultimately cannot estimate without bias due to unknown ancestral reference population allele frequencies $\{\pi_l\}$. As Ritland (1996) suggested, however, the values of an estimator of β_j may be close to F_j for samples where the average coancestry is small. The closeness of these two quantities will improve with sample size only if larger samples have smaller proportions of related individuals.

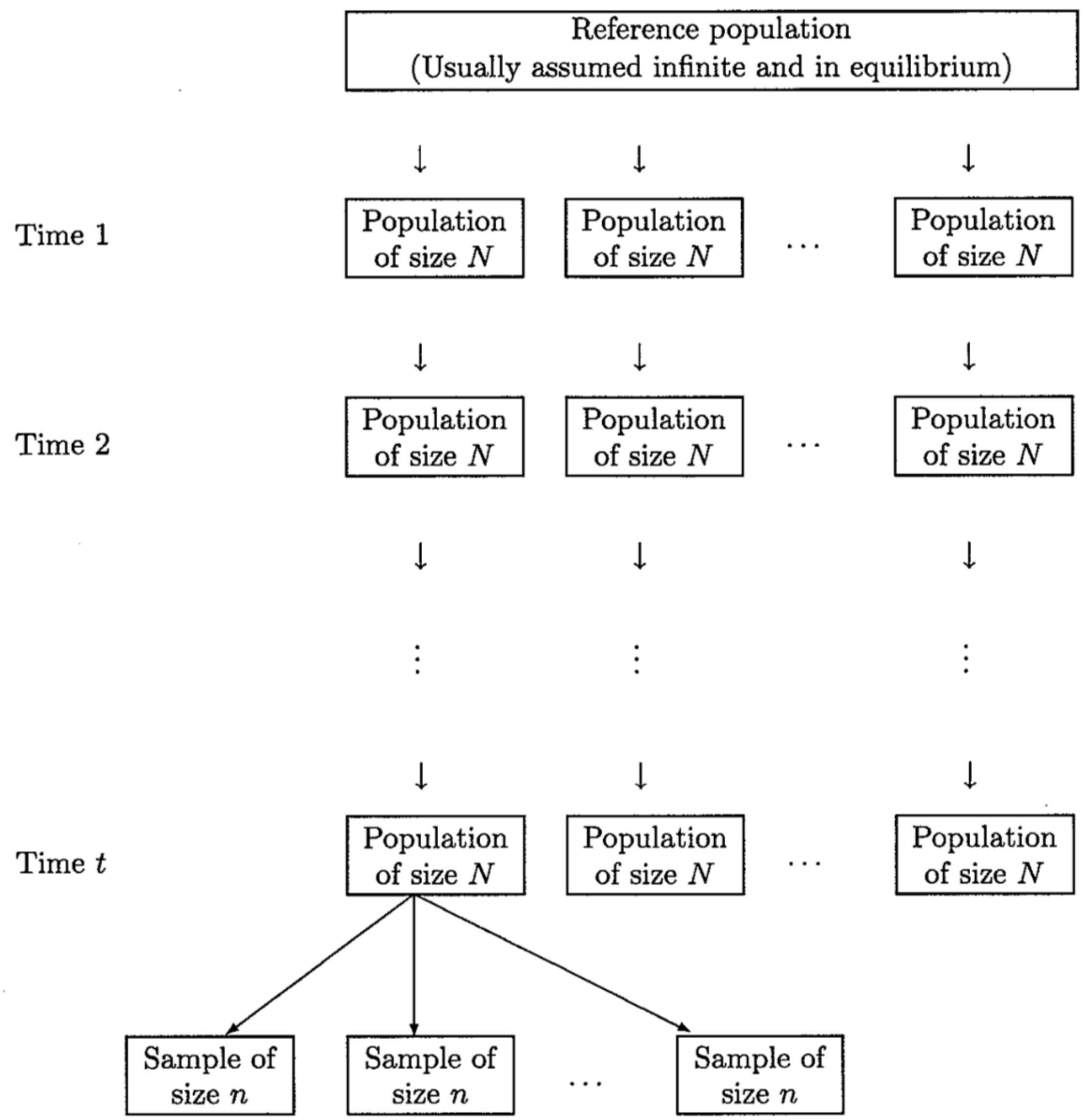

Figure 1.5 Representation of the two processes of genetic and statistical sampling. Genetic sampling causes the populations to differ, whereas statistical sampling causes the samples from one population to differ.

Figure 2.1: Weir (1996)'s genetic model: Populations descend from an ancestral reference population, and random samples of n individuals are obtained from a population. Our data consists of dosages for sample individuals from the same population typed at one or more bi-allelic markers (e.g. bi-allelic SNPs). The figure is copied from Weir (1996).

2.2.2 The Problem and Our Contribution

There is continuing interest in being able to use genetic marker data to estimate the inbreeding coefficient F for individuals. See the introduction of Ritland (1996) who describes applications in ecology and population genetics, including research on kin selection, isolation by distance, and paternity analysis. Quantitative genetic theory shows that both the expected value and the genetic variance of trait values for inbred individuals depend linearly on their inbreeding coefficients (Falconer, 1996). This is taken into account in association mapping, where the genetic relatedness matrix for a set of individuals has elements with estimated values of $(1 + \hat{F})$ on the diagonal. Recently, Yengo et al. (2017) used estimated inbreeding coefficients to characterize inbreeding depression in several human quantitative traits. These authors considered several different estimators of the inbreeding coefficient to estimate inbreeding depression. They recommended a particular method of moments estimator (MME), introduced earlier by Yang et al. (2011) as $\hat{F}_i^{III}$, an estimator we denote GCTA^{III}. A general discussion of moment estimators was given by Weir and Goudet (2017), with particular attention to assumptions about allele frequencies. Whereas Yengo et al. (2017) used sample allele frequencies in place of values in an ancestral population, Weir and Goudet (2017) showed how that leads to estimates biased by the relationships among all individuals used for estimating allele frequencies.

This chapter reviews some of the key literature for individual inbreeding coefficient estimation, compares existing methods for inbreeding coefficient F_j point estimation, evaluates a novel maximum likelihood estimator (MLE) of β_j, and assesses interval estimation and point estimation procedures on simulated data. In comparing existing methods, namely Hall et al.'s simple MME, GCTA^I MME, GCTA^{III} MME, Weir and Goudet's MME, and Hall et al.'s MLE (Hall et al., 2012), and our novel likelihood estimator - where all methods except Weir and Goudet's MME have sample allele frequencies assumed equal to population allele frequencies - we find that none of these methods are unbiased for an individual j's inbreeding coefficient F_j. To our knowledge, the bias for F_j of existing methods introduced as F_j estimators has not been previously acknowledged (Yang et al., 2011). We also find that Weir and Goudet's MME, combining over loci using the ratio of means method described later, has low bias for $\beta_j = (F_j - \theta_S)/(1-\theta_S)$, a monotone increasing function of F_j for fixed θ_S (Weir and Goudet, 2017).

2.3 Methods

2.3.1 Method of Moments Estimation: Combining Estimates Over Loci

The method of moment estimators (MMEs) that follow take the form $\hat{\theta}_l = \hat{a}_l/\hat{b}_l$ for a single locus l. Ways of combining locus-specific estimates over loci include the mean of ratios (MR) i.e. unweighted method, $\sum_{l=1}^{L} \hat{\theta}_l/L$, and the ratio of means (RM) i.e. weighted method, $\left[\sum_{l=1}^{L} \hat{a}_l/L\right]/\left[\sum_{l=1}^{L} \hat{b}_l/L\right] = \sum_{l=1}^{L} \hat{a}_l/\sum_{l=1}^{L} \hat{b}_l$. The software tool Genome-wide Complex Trait Analysis (GCTA) by Yang et al. (2011) uses MR.

The problem with MR is that SNPs with small $\hat{b}_l$ lead to inflated variance of the corresponding estimates. We advocate instead using RM, which is not sensitive to this problem.

2.3.2 Method of Moments Estimator (MME): Hall

If the reference allele frequencies $\{\pi_l\}$ are known, all information about parameter F, which measures departures from Hardy-Weinberg equilibrium (HWE) frequencies (such that $F = 0$ yields HWE frequencies), is captured by whether or not an individual is heterozygous at a locus. If indicator $\tilde{H}_{jo_l}$ is 1 when individual j is heterozygous at locus l, then the observed number $\tilde{H}_{jo}$ of heterozygous loci for individual j is

$$\tilde{H}_{jo} := \sum_{l=1}^{L} \tilde{H}_{jo_l} = \sum_{l=1}^{L} \Big[(2 - X_{jl})(X_{jl})\Big]. \tag{2.3}$$

X_{jl} is the reference allele dosage for individual j at locus l. Counts of heterozygous loci are called heterozygosities.

The total (accounting for genetic and statistical sampling) expected value of $\tilde{H}_{jo}$ is

$$\mathcal{E}_T(\tilde{H}_{jo}) \;=\; \sum_{l=1}^{L} \mathcal{E}_T(\tilde{H}_{jo_l}) = \sum_{l=1}^{L} P(AB_{jl}) = (1 - F_j) \sum_{l} 2\pi_l(1 - \pi_l).$$

Equating observed and expected values, $\tilde{H}_{jo}$ and $\mathcal{E}(\tilde{H}_{jo})$, gives Hall et al. (2012)'s "Simple" estimator:

$$\hat{F}_{j_{\text{Sim.}}} \;=\; 1 - \frac{\tilde{H}_{jo}}{H_e} = 1 - \frac{\sum_{l=1}^{L} \Big[(2 - X_{jl})(X_{jl})\Big]}{\sum_{l=1}^{L}(2\pi_l)(1 - \pi_l)},$$

where $H_e := \sum_l 2(\pi_l)(1 - \pi_l)$ is the number of loci expected to be heterozygous when $F_j = 0$. This estimator $\hat{F}_{j_{\text{Sim.}}}$ is unbiased for F_j.

In practice, however, the ancestral population reference allele probabilities $\{\pi_l\}$ are not known, so we instead substitute sample allele frequencies $\{\tilde{p}_l\}$ in their stead:

$$\hat{\beta}_{j_{\text{Sim.}}} \;=\; 1 - \frac{\sum_{l=1}^{L}(2 - X_{jl})X_{jl}}{2\sum_{l=1}^{L} \tilde{p}_l(1 - \tilde{p}_l)} = 1 - \frac{\tilde{H}_{jo}}{\sum_{l=1}^{L} 2\tilde{p}_l(1 - \tilde{p}_l)} = 1 - \frac{\tilde{H}_{jo}}{\tilde{H}_e}, \tag{2.4}$$

where $\tilde{H}_e$ is H_e with sample allele proportions $\{\tilde{p}_l\}$ in place of unknown ancestral population allele frequencies $\{\pi_l\}$. Estimator $\hat{\beta}_{j_{\text{Sim.}}}$ with $\{\tilde{p}_l\}$ in place of $\{\pi_l\}$ is not unbiased for F_j, since $E(\tilde{p}_l^2) \neq \pi_l^2$ (Equation 2.18). According to Hall et al. (2012), this biased estimator is used in PLINK (Purcell et al., 2007).

For the estimator $\hat{\beta}_{j_{\text{Sim.}}}$, the numerator $\tilde{H}_{jo}$ has a within-population expectation of $\sum_l 2p_l(1-p_l)(1 - f_j)$ and a total expectation of $\sum_l 2\pi_l(1 - \pi_l)(1 - F_j)$. These two expressions rest on genotype probabilities for the two alleles in one individual. For individual j at SNP l, the within-population and total probabilities both alleles are the reference type are respectively

$$\begin{aligned}
\mathcal{E}_W[X_{jl}(X_{jl} - 1)] &= p_l^2 + p_l(1 - p_l)f_j \text{ and} \\
\mathcal{E}_T[X_{jl}(X_{jl} - 1)] &= \pi_l^2 + \pi_l(1 - \pi_l)F_j.
\end{aligned}$$

The denominator of $\hat{\beta}_{j\text{Sim.}}$ rests on the observed genotypes of all individuals in the sample, and it is necessary to consider probabilities of pairs of alleles from different individuals. Randomly drawn alleles from individuals j, j' are both the reference type at SNP l with probabilities

$$\begin{aligned}
\mathcal{E}_W\left\{\frac{1}{2}[1+(X_{jl}-1)(X_{j'l}-1]\right\} &= p_l^2 + p_l(1-p_l)\phi_{jj'} \text{ and} \\
\mathcal{E}_T\left\{\frac{1}{2}[1+(X_{jl}-1)(X_{jl}-1)]\right\} &= \pi_l^2 + \pi_l(1-\pi_l)\theta_{jj'}.
\end{aligned}$$

These define the within-population coancestry coefficient $\phi_{jj'}$ and the total coancestry coefficient $\theta_{jj'}$ for individuals j and j'.

These expressions lead to the expectations of $2\tilde{p}_l(1-\tilde{p}_l)$,

$$\begin{aligned}
\mathcal{E}_W[2\tilde{p}_l(1-\tilde{p}_l)] &= 2p_l(1-p_l)\left[(1-\phi_S) - \frac{1}{2n}(1+f_W-2\phi_S)\right] \text{ and} \\
\mathcal{E}_T[2\tilde{p}_l(1-\tilde{p}_l)] &= 2\pi_l(1-\pi_l)\left[(1-\theta_S) - \frac{1}{2n}(1+F_W-2\theta_S)\right],
\end{aligned}$$

where ϕ_S, θ_S are the averages over pairs (j, j') of distinct sample individuals of $\phi_{jj'}$ and $\theta_{jj'}$ respectively and f_w, F_W are the averages over sample individuals of f_j and F_j respectively.

Clearly, $\hat{\beta}_{j\text{Sim.}}$ is not providing an estimate of either f_j or F_j, as its values are affected by the average inbreeding and coancestry values for the sample. For large samples, however, the ratio of the expected values of the numerator and denominator is the within-population individual-specific inbreeding coefficient $f_j = \beta_j$, and this is taken to be a parameter of interest.

2.3.3 Method of Moments Estimator (MME): GCTA^I and GCTA^{III}

Other moment estimators, such as that of Yang et al. (2011), use sample allele frequencies $\tilde{p}_l$'s at multiple loci for a sample of n individuals in place of reference allele frequencies p_l's. For a single locus l, their estimator is as follows:

$$\hat{F}_j = \frac{(X_{jl} - 2\tilde{p}_l)^2}{2\tilde{p}_l(1-\tilde{p}_l)} - 1, \tag{2.5}$$

where recall X_{jl} is individual j's reference allele dosage at a locus l and $\tilde{p}_l$ is the sample reference allele frequency at the locus. This is one of the estimators advocated by Yengo et al. (2017) for inbreeding depression studies, and it is used more frequently than the simple MME.

Combining over loci with the ratio of means (i.e. weighted) method, the GCTA^I_w estimator for individual j is the following:

$$\text{GCTA}^I_{jw} = \frac{\sum_{l=1}^{L}(X_{jl} - 2\tilde{p}_l)^2}{2\sum_{l=1}^{L}\tilde{p}_l(1-\tilde{p}_l)} - 1.$$

Combining over loci with the mean of ratios (i.e. unweighted) method, the GCTA^I_u estimator for individual j is the following:

$$\text{GCTA}^I_{ju} = \frac{1}{L}\sum_{l=1}^{L}\frac{(X_{jl} - 2\tilde{p}_l)^2}{2\tilde{p}_l(1-\tilde{p}_l)} - 1.$$

Yang et al. (2011) claimed that GCTA^I_j is unbiased, i.e. $\mathcal{E}(\text{GCTA}^I_j|F_j) = F_j$.

The GCTA approach is motivated by the expressions for the variances of allele dosage. The within-population and total expectations of X_{jl} are $2p_l$ and $2\pi_l$ respectively. The within-population and total variances are

$$\begin{aligned} \text{Var}_W(X_{jl}) = \mathcal{E}_W(X_{jl} - 2p_l)^2 &= 2p_l(1-p_l)(1+f_j) \text{ and} \\ \text{Var}_T(X_{jl}) = \mathcal{E}_T(X_{jl} - 2\pi_l)^2 &= 2\pi_l(1-\pi_l)(1+F_j). \end{aligned}$$

If the $\{p_l\}$ or $\{\pi_l\}$ were known, then the sample variance of allele dosages would give a simple moment estimator of f_j or F_j, and this is implied by the GCTA estimator with $\tilde{p}_l$

replacing p_l or π_l. However, $2\tilde{p}_l$ is not the sample mean dosage for a particular individual j. The numerator $(X_{jl} - 2\tilde{p}_l)^2$ for GCTA does not have the variance of X_{jl} as its expectation. Instead,

$$\mathcal{E}_T[(X_{jl} - 2\tilde{p}_l)^2] = 2\pi_l(1-\pi_l)\left[(1 + F_j - 4\psi_j + 2\theta_S) - \frac{1}{n}\left(2(1 + F_j - 2\psi_j) - (1 + F_W - 2\theta_S)\right)\right],$$

where $\psi_j = \sum_{j'=1, j' \neq j}^{n} \theta_{jj'}/(n-1)$ is the average coancestry of individual j with every other member of the sample of n individuals from the same population, and π_l, F_j, θ_S, and F_W were defined previously in Sections 2.2.1 and 2.3.2. The denominator $2\tilde{p}_l(1 - \tilde{p}_l)$ for GCTA has expectation

$$\mathcal{E}_T[2\tilde{p}_l(1-\tilde{p}_l)] = 2\pi_l(1-\pi_l)\left[(1-\theta_S) - \frac{1}{2n}(1 + F_W - 2\theta_S)\right].$$

We refer to the parameter $(F_j - 4\psi_j + 3\theta_S)/(1 - \theta_S)$ as the "pedigree parameter" or "parametric pedigree value" of GCTA^I (RM or MR).

Yang et al. (2011) also gave an estimator GCTA^{III}, which they claimed was unbiased (i.e. $E(\text{GCTA}^{III}|F_j) = F_j$) with smaller variance than GCTA^I when $0 \leq F_j < 1/3$:

$$\text{GCTA}_j^{III} = \frac{[X_{jl}^2 - (1 + 2\tilde{p}_l)(X_{jl}) + 2\tilde{p}_l^2]}{2(\tilde{p}_l)(1-\tilde{p}_l)}.$$

Combining over loci with the ratio of means (i.e. weighted) method, the GCTA$_{jw}^{III}$ estimator for individual j is the following:

$$\text{GCTA}_{jw}^{III} = \frac{\sum_l [X_{jl}^2 - (1 + 2\tilde{p}_l)(X_{jl}) + 2\tilde{p}_l^2]}{\sum_l 2(\tilde{p}_l)(1-\tilde{p}_l)}.$$

Combining over loci with the mean of ratios (i.e. unweighted) method, the GCTA$_{ju}^{III}$ estimator for individual j is the following:

$$\mathrm{GCTA}^{III}_{ju} = \frac{1}{L}\sum_{l=1}^{L}\frac{[X_{jl}^2 - (1+2\tilde{p}_l)(X_{jl}) + 2\tilde{p}_l^2]}{2(\tilde{p}_l)(1-\tilde{p}_l)}.$$

We refer to the parameter $(F_j - 2\psi_j + \theta_S)/(1-\theta_S)$ as the "pedigree parameter" or "parametric pedigree value" of GCTA^{III}, whether RM or MR.

2.3.4 Method of Moments Estimator (MME): Weir and Goudet

To avoid explicit estimation of allele probabilities, Weir and Goudet adopted a method of moments estimator (MME) for the inbreeding coefficient F_j for individual j in a sample of individuals *relative to* the average coancestry θ_S of all pairs of individuals in the sample. Making the target parameter $\beta_j := (F_j - \theta_S)/(1-\theta_S)$ means that the unknown reference allele frequencies $\{\pi_l\}$ do not need to be estimated. The quantity $\beta_j = f_j$ is the inbreeding coefficient F_j of individual j relative to the average coancestry θ_S of all pairs of individuals in the sample to which j belongs, and this is a parameter under study. This β_j quantity that the MME aims to estimate is related to F_j in that it ranks according to F_j, i.e. $F_j < F_{j'}$ if and only if $\beta_j < \beta_{j'}$. We refer to β_j as the "pedigree parameter" or "parametric pedigree value" of the Weir and Goudet MME, whether RM or MR.

Their estimator was, for a single locus l,

$$\hat{\beta}_j = \frac{\tilde{M}_{jl} - \tilde{M}_{Sl}}{1 - \tilde{M}_{Sl}}, \tag{2.6}$$

where, for a sample of n individuals, the within-individual-j matching proportion is

$$\tilde{M}_{jl} := (X_{jl} - 1)^2. \tag{2.7}$$

The matching proportion $\tilde{M}_{jl}$ is 1 if the two alleles individual j has at the locus l are identical in state, and 0 otherwise. The average between-individual-j matching proportion is

$$\tilde{M}_{Sl} := \sum_{j=1}^{n} \sum_{j'=1, j' \neq j}^{n} [1 + (X_{jl} - 1)(X_{j'l} - 1)]/[2n(n-1)], \tag{2.8}$$

where $[1 + (X_{jl} - 1)(X_{j'l} - 1)]/2$ gives the fraction of pairs of alleles, one allele drawn from individual j and the other allele drawn from individual j' of the same population as j at the same locus l, that are identical in state (ibs), i.e. of the same allelic type. The matching proportion $\tilde{M}_{Sl}$ is the mean of such fractions over all pairs of individuals j and j'.

Combining statistics over L loci using the ratio of means (RM) i.e. weighted method, we have:

$$\begin{aligned} \hat{\beta}_{jw} &= \frac{\sum_{l=1}^{L} (\tilde{M}_{jl} - \tilde{M}_{Sl})}{\sum_{l=1}^{L} (1 - \tilde{M}_{Sl})} \\ &= 1 - \frac{\sum_{l=1}^{L} (1 - \tilde{M}_{jl})}{\sum_{l=1}^{L} (1 - \tilde{M}_{Sl})}. \end{aligned} \tag{2.9}$$

Alternatively, one can combine over loci using an unweighted average:

$$\hat{\beta}_{ju} = \frac{1}{L} \sum_{l=1}^{L} \frac{\tilde{M}_{jl} - \tilde{M}_{Sl}}{1 - \tilde{M}_{Sl}}.$$

2.3.4.1 Connection Between Weir and Goudet MME and Hall's Simple MME

In Section 2.6.4, we show these relations between matching proportions and heterozygosities:

$$\begin{aligned} &\sum_{l=1}^{L} (1 - \tilde{M}_{jl}) = \tilde{H}_{jo} \text{ and} \\ &\sum_{l=1}^{L} (1 - \tilde{M}_{Sl}) = \frac{n}{n-1} \left(\tilde{H}_e - \frac{1}{2n} \tilde{H}_o \right), \end{aligned} \tag{2.10}$$

where $\tilde{H}_e$ is $\sum_{l=1}^{L} (2\tilde{p}_l(1-\tilde{p}_l))$, and $\tilde{H}_o$ is the average of the $\tilde{H}_{jo}$ over individuals $j \in \{1, ..., n\}$. Recall from Definition 2.3 that $\tilde{H}_{jo}$ is the number of heterozygous loci in individual j, and $\tilde{M}_{jl}$ is an indicator for whether the two alleles in individual j at locus l are identical in state.

Using Equations 2.4 and 2.9 that give formulas for estimators, relations 2.10 between matching proportions and heterozygosities, moments of matching proportions derived in Appendix 2.6.3, the Strong Law of Large Numbers and the Continuous Mapping Theorem, we proved in Appendix 2.6.5 that the difference between estimators $\hat{\beta}_{j\text{Sim.}}$ and $\hat{\beta}_{jw}$ converges almost surely to 0 as $L \to \infty$ and $n \to \infty$, so long as $[\tilde{M}_{jl}, \tilde{M}_{Sl}], l \in \{1, ..., L\}$ are independent and identically distributed (implying π_l takes the same value π for all l, $\mathcal{E}(1-\tilde{M}_{jl})$ is the same quantity $\mathcal{E}(1-\tilde{M}_j)$ for all l, and $\mathcal{E}(1-\tilde{M}_{Sl})$ is the same quantity $\mathcal{E}(1-\tilde{M}_S)$ for all l) and these two conditions hold: $\mathcal{E}(1-\tilde{M}_j) \neq 0$ and $\mathcal{E}(1-\tilde{M}_S) \neq 0$. From Equation 2.23, $\mathcal{E}(1-\tilde{M}_{jl}) = 2\pi_l(1-\pi_l)(1-F_j)$, and Equation 2.25, $\mathcal{E}(1-\tilde{M}_{Sl}) = 2\pi_l(1-\pi_l)(1-\theta_S)$, these two conditions are equivalent to $\pi \neq 0$, $\pi \neq 1$, $F_j \neq 1$, and $\theta_S \neq 1$.

2.3.4.2 Adjusting $\hat{\beta}_j$ to Obtain Nonnegative Estimates

If $\hat{\beta}_j$ estimates are used in place of F_j's in diagonal entries $(1+F_j)$'s of a genetic relatedness matrix (GRM), negative estimates could result in a non-positive semidefinite GRM. To obtain nonnegative estimates to substitute in for F_j's in a GRM, one can define the following adjusted estimator: $\hat{\beta}_j^{adj} := (\hat{\beta}_j - \hat{\beta}_0)/(1-\hat{\beta}_0)$, where $\hat{\beta}_0$ is the mean $\hat{\beta}_j$, averaged over the least inbred individuals' in the sample, those with the smallest $\hat{\beta}_j$ estimates, assumed to be those with $F_j = 0$.

2.3.5 Likelihood Approach: First Likelihood Model of Hall et al. (2012) (Genotype Likelihood for MLEg)

If the observed genotype for an individual j at locus l is G_{jl}, then the likelihood as a function of inbreeding coefficient F_j is

$$L(F_j) \quad = \quad \prod_{l=1}^{L} \Pr(G_{jl}|F_j), \tag{2.11}$$

where genotype probabilities at each locus are functions of the inbreeding coefficient, as shown in Equations 2.12,

$$\begin{aligned}
\Pr(AA_j) &= \pi^2 + \pi(1-\pi)(F_j) \\
\Pr(AB_j) &= 2\pi(1-\pi)(1-F_j) \\
\Pr(BB_j) &= (1-\pi)^2 + \pi(1-\pi)(F_j),
\end{aligned} \tag{2.12}$$

and genotypes at different loci are assumed to be independent.

Substituting expressions from these three equations into Equation 2.11 and noting that $(\pi_l)^2 + (\pi_l)(1-\pi_l)(F_j) = (\pi_l)(\pi_l + (1-\pi_l)(F_j))$ and $(1-\pi_l)^2 + (\pi_l)(1-\pi_l)(F_j) = (1-\pi_l)((1-\pi_l) + (\pi_l)(F_j))$, we obtain expressions for the log likelihood function:

$$\begin{aligned}
\ln[L(F_j)] &= \sum_{l=1}^{L} \ln(P(G_{jl}|F_j)) \\
&= \sum_{l=1}^{L} \{I_{AA_{jl}} \ln[(\pi_l)^2 + (\pi_l)(1-\pi_l)(F_j)]\} \\
&+ \{I_{AB_{jl}} \ln[2\pi_l(1-\pi_l)(1-F_j)]\} \\
&+ \{I_{BB_{jl}} \ln[(1-\pi_l)^2 + (\pi_l)(1-\pi_l)(F_j)]\} \\
&= \sum_{l=1}^{L} \{I_{AA_{jl}} [\ln(\pi_l) + \ln(\pi_l + (1-\pi_l)(F_j))]\} \\
&+ \{I_{AB_{jl}} [\ln(2) + \ln(\pi_l(1-\pi_l)) + \ln(1-F_j)]\} \\
&+ \{I_{BB_{jl}} [\ln(1-\pi_l) + \ln((1-\pi_l) + (\pi_l)(F_j))]\} \\
&=_{C+} \sum_{l=1}^{L} \{I_{AA_{jl}} [\ln(\pi_l + (1-\pi_l)(F_j))]\} \\
&+ \{I_{AB_{jl}} [\ln(1-F_j)]\} \\
&+ \{I_{BB_{jl}} [\ln((1-\pi_l) + (\pi_l)(F_j))]\},
\end{aligned}$$

where the indicator I_Z is 1 when Z is true and 0 otherwise, and $=_{C+}$ means we have kept only terms of $\ln[L(F_j)]$ that depend on F_j, so that our expression is equivalent to $\ln[L(F_j)]$

up to an added constant. Evaluating $\ln[L(F_j)]$ up to an added constant rather than fully can save computation time.

The first likelihood method of Hall et al. assumes the $\{\pi_l\}$ are known and equal to sample allele frequencies $\{\tilde{p}_l\}$. If the reference allele frequencies $\{\pi_l\}$ are known, then the likelihood is a function of a single parameter F_j. In evaluating $\ln[L(F_j)]$, we consider indicator functions first, before evaluating the ln(.) function, so that only when an individual has a certain genotype, does the term corresponding to that genotype contribute to the log likelihood evaluation. If an indicator function is 1, then we consider evaluating the ln(.) factor that it gets multiplied to. If the ln(.) factor has an argument (either $(\tilde{p}_l) + (1 - \tilde{p}_l)(F_j)$, $(1 - F_j)$, or $(1 - \tilde{p}_l) + (\tilde{p}_l)(F_j)$) that is negative when evaluated at an F_j value, we consider that F_j value outside the domain of $\ln[L(F_j)]$. An analytical solution for MLE $\hat{F}_j$ is not available, but numerical methods to obtain $\hat{F}_j$ are available, such as an iterative EM algorithm, or a one-dimensional grid search over F_j.

Hall et al. use an EM algorithm for an individual with inbreeding coefficient F_j. We prefer the grid search over the EM algorithm, because the grid search can return a global maximizer rather than just a stationary point, and because a 1-dimensional grid search can involve less computation than performing an iterative procedure until convergence, once per starting value.

We implemented a grid search as follows:

1. Set a variable log_Like_Fj_MLE that records our latest value for the log likelihood $\ln(L(\hat{F}_j))$ evaluated at the MLE $\hat{F}_j$ to $-Inf$.

2. For loop through F_j values from -1 to 1 (boundaries imposed by Equations 2.12) by an increment of 0.001.

For each F_j value, do steps 3-4.

3. Set variable log_Like_at_Fj_value to 0 and increment its value by summing over L terms $\ln(P(G_{jl}|F_j))$ up to an added constant, $l \in \{1, ..., L\}$

To evaluate $\ln(P(G_{jl}|F_j))$ up to an added constant, for each SNP l, determine $X_l : 2, 1, 0$ for AA, AB, and BB genotypes. Compute the argument of $\ln(P(G_{jl}|F_j))$ up to an added constant as $(\tilde{p}_l) + (1 - \tilde{p}_l)(F_j)$, $(1 - F_j)$, and $(1 - \tilde{p}_l) + (\tilde{p}_l)(F_j)$ respectively.

If the argument is less than or equal to zero, return to Step 2 and consider the next F_j value. If the argument is greater than 0, take its natural log $\ln(P(G_{jl}|F_j))$ up to an added constant, and add that to log_Like_at_Fj_value.

After iterating over all SNPs, variable log_Like_at_Fj_value should equal $\ln(L(F_j))$ up to an added constant.

4. If variable log_Like_Fj_MLE is less than log_Like_at_Fj_value (i.e. $\ln(L(F_j))$), then replace log_Like_Fj_MLE with log_Like_at_Fj_value and set a variable Fj_MLE that is our latest value for MLE $\hat{F}_j$ to the F_j value.

5. After looping over F_j values, we should have variable Fj_MLE equal to MLE $\hat{F}_j$, and have variable log_Like_Fj_MLE be the log likelihood evaluated at the MLE.

The problem with replacing allele frequencies π_l's with sample values $\tilde{p}_l$'s is that sample values have variances that depend on the inbreeding coefficients $\{F_j\}$ of all members of the sample and coancestry coefficients $\{\theta_{jj'}\}$ of all pairs (j, j') of members of the sample, which can be seen from Equation 2.18 for $\mathcal{E}(\tilde{p}^2)$. If n individuals are sampled from a population, then the sample allele frequency is

$$\tilde{p} \;=\; \frac{1}{2n}\sum_{j=1}^{n} X_j = \frac{1}{2n}\sum_{j=1}^{n}(x_{j1} + x_{j2}),$$

where the first subscript j of x_{jk} indexes the individual and the second subscript indexes the allele within that individual. Each x_{jk} has expected value π so $\tilde{p}$ also has expected value π. The expected value of $\tilde{p}^2$ is Equation 2.18 derived in the Appendix:

$$\mathcal{E}(\tilde{p}^2) = \pi^2 + \pi(1-\pi)\theta_S + \frac{1}{2n}\pi(1-\pi)(1 + F_W - 2\theta_S).$$

Here we have used the summary measures

$$F_W = \frac{1}{n}\sum_{j=1}^{n} F_j \text{ and}$$

$$\theta_S = \frac{1}{n(n-1)}\sum_{j=1}^{n}\sum_{j'=1,j'\neq j}^{n} \theta_{jj'}.$$

While we do not have an explicit formula for the expectation of Hall et al. (2012)'s MLE for the model with sample frequencies equal to population frequencies, we surmise that because its employs $\tilde{p}_l^2$ in its likelihood function (Equation 2.11), and $\mathcal{E}(\tilde{p}_l^2)$ is a function of inbreeding coefficients and coancestry coefficients involving other (non-j) individuals in the sample, Hall et al. (2012)'s MLE is also biased for F_j.

We call the MLE $\hat{F}_j$ from this model Hall MLE 1 (because the likelihood model is Hall et al. (2012)'s first), or MLEg (because the likelihood model involves genotype probabilities). We later refer to β_j as the "pedigree parameter" or "parametric pedigree value" of this MLE.

2.3.6 Likelihood Approach: Our Likelihood Model (Heterozygosity Likelihood for MLEh)

In our likelihood model, we combine homozygous genotype categories into a single category. The likelihood function is

$$\ln[L(F_j)] = \sum_{l=1}^{L}\{I_{\text{Heterozygous}_{\text{jl}}} \ln[2(\pi_l)(1-\pi_l)(1-F_j)]\}$$
$$+ \{I_{\text{Homozygous}_{\text{jl}}} \ln[1-2(\pi_l)(1-\pi_l)(1-F_j)]\},$$

where $I_{\text{Heterozygous}_{\text{jl}}}$ indicates whether individual j's genotype at locus l is heterozygous (1 if so, 0 otherwise) and $I_{\text{Homozygous}_{\text{jl}}}$ indicates whether the genotype is homozygous (1 if so, 0 otherwise). Our indicators of heterozygous or homozygous genotype are connected to Hall et al. (2012)'s via the following relations: $I_{\text{Homozygous}_{\text{jl}}} = I_{AA_{jl}} + I_{BB_{jl}}$ and $I_{\text{Heterozygous}_{\text{jl}}} = I_{AB_{jl}}$.

Intuition for collapsing genotype categories is the following: When we estimate an inbreeding coefficient F_j from genotype data, we are inferring identity by descent (ibd) from identity by state (ibs). In other words, the relevant information for ibd inference is whether an individual's two alleles are ibs (i.e. of the same allelic type), and particular genotypes are irrelevant.

We use grid search to obtain an MLE for F_j from this model, analogous to the procedure given for Hall et al. (2012)'s first likelihood model.

To emphasize the difference between Hall et al. (2012)'s first likelihood model and our model, we call an MLE $\hat{F}_j$ from our model MLEh, "h" for "heterozygosity," and we call an MLE $\hat{F}_j$ from Hall et al. (2012)'s first likelihood model MLEg, "g" for "genotype." We later refer to β_j as the "pedigree parameter" or "parametric pedigree value" of MLEh.

2.3.7 Application to Simulated Data

The *sim* package was used by J. Goudet (personal communication) to construct a base population with SNPs in linkage blocks of 20 Morgans. A set of 20 founders was selected for a defined pedigree, and 79,069 SNPs were chosen independently to provide genotypes for 115 descendants in this pedigree. Only data for the non-founders were used in the analyses reported here. There are 17 individuals with inbreeding coefficient of 0.125, and 98 with inbreeding coefficient of zero, so average inbreeding was $F_W = 0.0185$ for the sample of 115 individuals. For the 6,555 pairs of distinct individuals, the coancestries (and numbers of pairs with each value) were 0 (6340), 1/32 (1146), 1/16 (2176), 3/32 (786), 1/8 (1398), 5/32 (228), 3/16 (302), 1/4 (584) and 5/16 (150). The average coancestry $\theta_S = 0.0538$ was greater than the average inbreeding F_W.

We applied the following methods to the full data set: GCTA^{I} MME (weighted and unweighted), GCTA^{III} MME (weighted and unweighted), Weir and Goudet MME (weighted and unweighted), Hall et al.'s first genotype likelihood model's MLE (MLEg), and our heterozygosity likelihood model's MLE (MLEh).

To obtain interval estimates for point estimator $\hat{\theta}$ an MME (i.e. Weir and Goudet RM and MR, GCTA^I MME RM and MR, GCTA^{III} MME RM and MR) or $\hat{\theta}$ an MLE (i.e. Hall MLE 1 also known as MLEg, and MLEh), we performed the following bootstrap procedure for each estimator $\hat{\theta}$ of an individual-specific parameter:

(1) Sample L columns of $[\vec{X}_{.1}, ..., \vec{X}_{.l}, ..., \vec{X}_{.L}]$, where $\vec{X}_{.l}$ is the vector of n individuals' dosages at locus l, with replacement, obtaining bootstrap sample $[\vec{X}^*_{.1}, ..., \vec{X}^*_{.l}, ..., \vec{X}^*_{.L}]$.

(2) Calculate $[\hat{\theta}_1, ..., \hat{\theta}_n]$ for n individuals on the bootstrap sample.

For $\hat{\theta}$ an MME, we repeated Steps (1)-(2) $B = 1,000$ times to get $1,000$ $M\hat{M}E_j$ estimates per individual j. For $\hat{\theta}$ an MLE, due to the computational burden of grid search, we used $B = 80$. We took the $[0.025, 0.975]$ quantiles of bootstrap estimates to form a 95% confidence interval (CI) for $\hat{\theta}$. We calculated coverage as the fraction of $n = 115$ CIs that covered the parameter of interest (either F_j or the estimator's pedigree parameter), where an interval [lower endpoint, upper endpoint] was considered to cover a parameter if lower endpoint $\leq$ parameter $\leq$ upper endpoint.

2.4 Results

We evaluated MME and MLE point estimators and their bootstrap interval estimators on simulated data.

2.4.1 Squared Bias and Mean Squared Error (MSE)

Observing Table 2.1, we see that every estimator's parametric pedigree MSE is lower than its F_j MSE, and every estimator's parametric pedigree Bias^2 is lower than its F_j Bias^2, indicating that these estimators are better estimators of their pedigree parameters than of F_j.

Of the estimators with pedigree parameter β_j, we list methods in order of β_j MSE, from lowest to highest: Weir and Goudet $\hat{\beta}_{jw}$ Ratio of Means MME, our MLEh, Hall et al. (2012)'s MLEg, and Weir and Goudet $\hat{\beta}_{ju}$ Mean of Ratios MME. We note that Weir and Goudet's MME RM had low squared bias for β_j. Since we derived that the difference between the weighted Weir and Goudet $\hat{\beta}_{jw}$ MME and Hall et al.'s simple MME converges almost surely to 0 when $n \to \infty$, $L \to \infty$, and $F_j \neq 1$, $\theta_S \neq 1$, and for all l, $\pi_l \neq 0$, $\pi_l \neq 1$, in samples with a large number n of individuals and a large number L of loci, Hall et al.'s simple MME should also have low bias for β_j.

Estimator	Bias^2 F_j	MSE F_j	Bias^2 Parametric Pedigree Value	MSE Parametric Pedigree Value
Weir and Goudet $\hat{\beta}_{jw}$ Ratio of Means MME	3.18×10^{-3}	3.21×10^{-3}	2.49×10^{-7}	3.41×10^{-5}
Weir and Goudet $\hat{\beta}_{ju}$ Mean of Ratios MME	2.75×10^{-3}	2.58×10^{-2}	1.16×10^{-5}	2.28×10^{-2}
GCTA^{I}_{jw} Ratio of Means MME	3.69×10^{-3}	8.30×10^{-3}	5.97×10^{-5}	9.41×10^{-5}
GCTA^{I}_{ju} Mean of Ratios MME	3.22×10^{-3}	4.29×10^{-2}	1.36×10^{-4}	1.88×10^{-2}
GCTA^{III}_{jw} Ratio of Means MME	3.69×10^{-3}	4.86×10^{-3}	1.50×10^{-7}	1.93×10^{-5}
GCTA^{III}_{ju} Mean of Ratios MME	3.22×10^{-3}	3.95×10^{-3}	1.26×10^{-5}	2.11×10^{-4}
MLEg	3.04×10^{-3}	3.80×10^{-3}	4.94×10^{-7}	8.46×10^{-4}
MLEh	3.72×10^{-3}	3.94×10^{-3}	2.63×10^{-5}	2.62×10^{-4}

Table 2.1: Bias^2 and MSE of MMEs, Hall et al. (2012)'s MLEg, and Our MLEh on Simulated Data. Bias^2 was $[\sum_{j=1}^{115}(\hat{\theta}_j - \theta_j)/115]^2$ and Mean Squared Error (MSE) was $\sum_{j=1}^{115}(\hat{\theta}_j - \theta_j)^2/115$ for estimator $\hat{\theta}_j$ and two parameters θ_j, individual j inbreeding coefficient F_j and the estimator's parametric pedigree value.

2.4.2 Coverages

Estimator	F_j Coverage	Parametric Pedigree Value Coverage
Weir and Goudet $\hat{\beta}_{jw}$ Ratio of Means MME	0.00	9.39×10^{-1}
Weir and Goudet $\hat{\beta}_{ju}$ Mean of Ratios MME	1.13×10^{-1}	1.57×10^{-1}
GCTA^{I}_{jw} Ratio of Means MME	2.61×10^{-2}	7.13×10^{-1}
GCTA^{I}_{ju} Mean of Ratios MME	6.09×10^{-2}	1.91×10^{-1}
GCTA^{III}_{jw} Ratio of Means MME	4.35×10^{-2}	9.57×10^{-1}
GCTA^{III}_{ju} Mean of Ratios MME	4.35×10^{-2}	3.83×10^{-1}
MLEg	1.74×10^{-1}	5.83×10^{-1}
MLEh	6.96×10^{-2}	9.65×10^{-1}

Table 2.2: 95% Nominal Bootstrap Confidence Interval Coverages of Methods for Pedigree F_j and Each Method's Pedigree Parameter. Weir and Goudet MMEs, MLEg, and MLEh all had the same pedigree parameter β_j. GCTA^{I} and GCTA^{III} MMEs' pedigree parameters were given in Section 2.3.3. Coverage was the fraction of $n = 115$ intervals that covered the parameter.

Coverages of 95% confidence intervals of estimators for F_j and their pedigree parameters are given in Table 2.2. No estimator had coverage close to 95% for F_j. For Weir and Goudet MMEs (RM and MR), MLEg, and MLEh, the pedigree parameter was the same quantity, β_j. We see that MLEh and Weir and Goudet RM had coverages close to 95% for β_j, and comparing likelihood methods, MLEh had coverage closer to 95% than MLEg. Comparing MMEs, RM had higher pedigree parameter coverage than MR.

2.4.3 Plots of Estimates vs Parameter Values F_j and β_j Determined from the Pedigree

Figure 2.2 plots GCTA^{I} MME estimates, combining over loci using the ratio of means (RM) and mean of ratios (MR) methods, for $n = 115$ simulated individuals versus individuals' pedigree F_j values and parametric pedigree $(F_j - 4\psi_j + 3\theta_S)/(1-\theta_S)$ values. We see that estimates track more closely with parametric pedigree values than with pedigree F_j values, which is surprising, since GCTA^I MMEs (RM and MR) were intended to estimate F_j, not the estimator's pedigree parameter. Closer tracking with the pedigree parameter is consistent with the estimators having higher coverage for their pedigree parameter than for F_j, shown in Table 2.2, as well as lower Bias^2 for their pedigree parameter than for F_j, shown in Table 2.1. Figure 2.3 gives the same Figure 2.2 with y-axis constrained to [-0.15,0.15].

Figure 2.4 plots GCTA^{III} MME estimates, combining over loci using the ratio of means (RM) and mean of ratios (MR) methods, for the same $n = 115$ simulated individuals versus individuals' pedigree F_j values and parametric pedigree $(F_j - 2\psi_j + \theta_S)/(1-\theta_S)$ values. We see that estimates tracked more closely with parametric pedigree values than with pedigree F_j values, which is surprising, since GCTA^{III} MMEs (RM and MR) were intended to estimate F_j, not the estimator's pedigree parameter. Closer tracking with the pedigree parameter is consistent with the estimators having higher coverage for their pedigree parameter than for F_j, shown in Table 2.2, as well as lower Bias^2 for their pedigree parameter than for F_j, shown in Table 2.1. Figure 2.4's y-axis is essentially constrained to [-0.15,0.15].

Figure 2.5 plots Weir and Goudet $\hat{\beta}_j$ MME estimates, combining over loci using the ratio of means (RM) and mean of ratios (MR) methods, for the same $n = 115$ individuals versus individuals' pedigree F_j and parametric pedigree β_j values. We see that estimates tracked more closely with parametric pedigree β_j values than with pedigree F_j values, which is not surprising, since the Weir and Goudet $\hat{\beta}_j$ MME was designed to estimate β_j, not F_j. Closer tracking with the pedigree parameter is consistent with the estimators having higher coverage for their pedigree parameter than for F_j, shown in Table 2.2, as well as lower Bias^2 for their pedigree parameter than for F_j, shown in Table 2.1. Figure 2.6 gives the same Figure 2.5 with y-axis constrained to [-0.15, 0.15].

Figure 2.7 plots point estimates for Hall Model 1 (i.e. genotype likelihood MLEg), where sample frequencies are assumed equal to population frequencies, for the same $n = 115$ individuals versus individuals' pedigree F_j and parametric pedigree β_j values. We see that estimates tracked more closely with parametric pedigree β_j values than with pedigree F_j values, consistent with higher coverage for β_j than for F_j, shown in Table 2.2, as well as lower Bias^2 for β_j than for F_j, shown in Table 2.1. Figure 2.8 gives the same Figure 2.7 with y-axis constrained to [-0.15,0.15].

Figure 2.9 plots the heterozygosity MLE (MLEh) point estimates, where sample frequencies are assumed equal to population frequencies, for the same $n = 115$ individuals versus individuals' pedigree F_j and parametric pedigree β_j values. We see that estimates tracked more closely with parametric pedigree β_j values than with pedigree F_j values, consistent with higher coverage for β_j than for F_j, shown in Table 2.2, as well as lower Bias^2 for β_j than for F_j, shown in Table 2.1. Figure 2.10 gives the same Figure 2.9 with y-axis constrained to [-0.15,0.15].

Comparing all these Figures 2.3, 2.4, 2.6, 2.8, and 2.10 on the same scale, we see that none of these estimators tracked closely with F_j. Comparing Figures 2.6 and 2.10, we see that the Weir and Goudet MME RM had narrower bootstrap confidence intervals than MLEh.

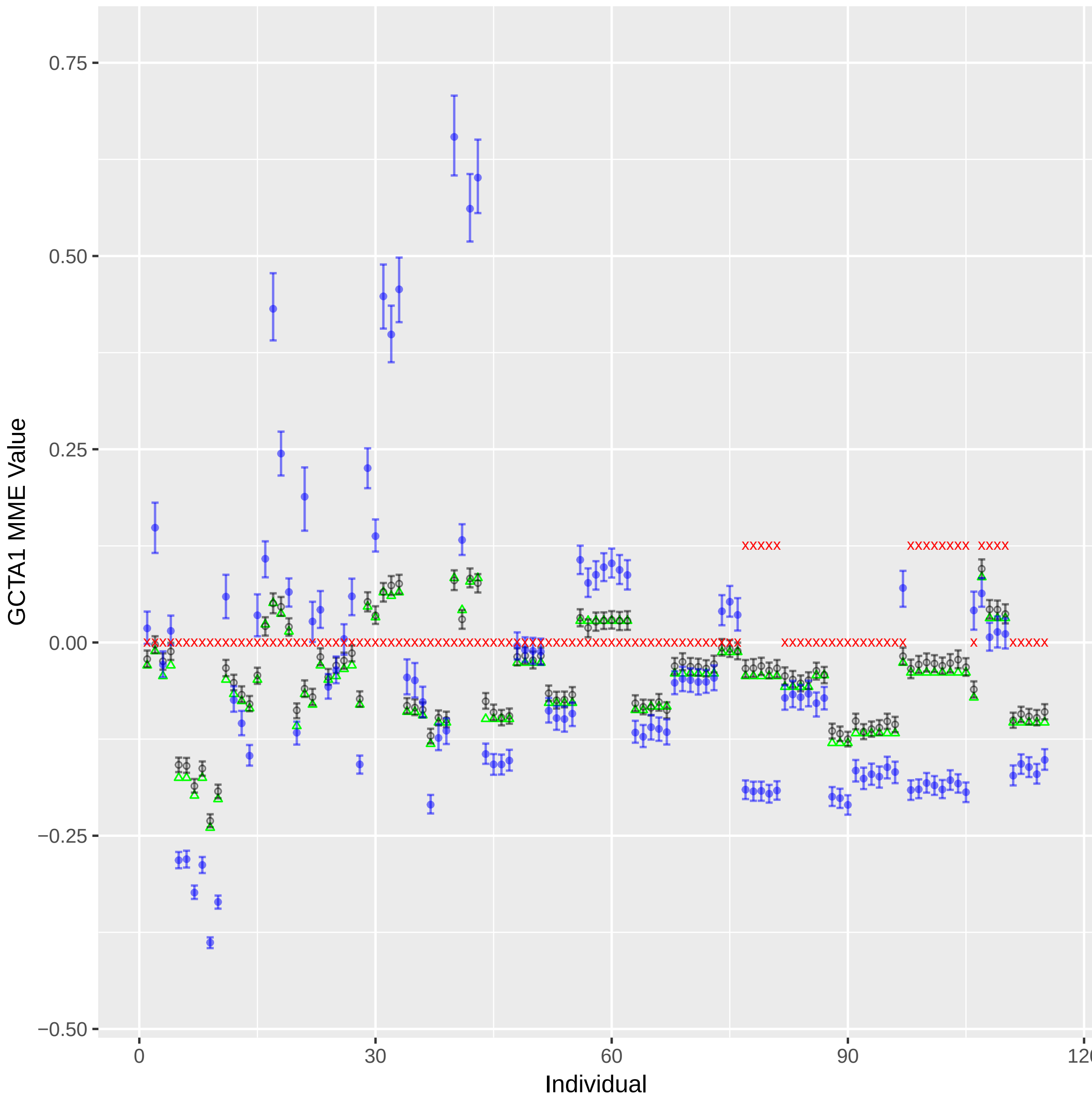

Figure 2.2: GCTA^I MME (RM and MR) Estimates for 115 Individuals and their Pedigree F_j and Parametric Pedigree Values. **Black, Hollow Circle**: GCTA^I MME RM point estimates and 95% bootstrap confidence intervals. **Blue, Solid Circle**: GCTA^I MME MR point estimates and 95% bootstrap confidence intervals. **Red, Cross**: F_j parameter determined from pedigree. **Green, Hollow Triangle**: Parameter $(F_j - 4\psi_j + 3\theta_S)/(1-\theta_S)$ determined from pedigree.

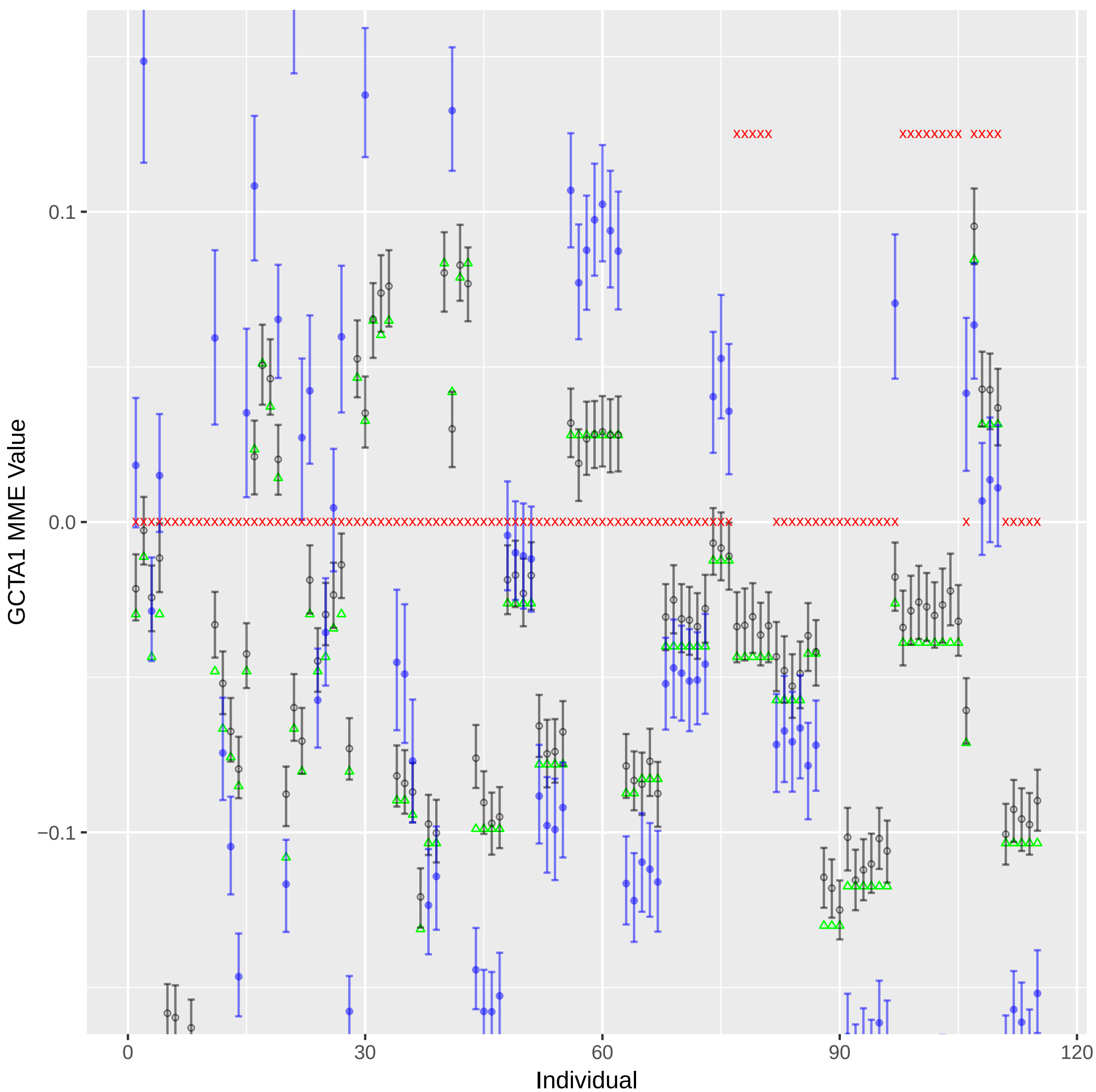

Figure 2.3: GCTA^I MME (RM and MR) Estimates for 115 Individuals and their Pedigree F_j and Parametric Pedigree Values with y-axis constrained to [-0.15,0.15]. **Black, Hollow Circle**: GCTA^I MME RM point estimates and 95% bootstrap confidence intervals. **Blue, Solid Circle**: GCTA^I MME MR point estimates and 95% bootstrap confidence intervals. **Red, Cross**: F_j parameter determined from pedigree. **Green, Hollow Triangle**: Parameter $(F_j - 4\psi_j + 3\theta_S)/(1 - \theta_S)$ determined from pedigree.

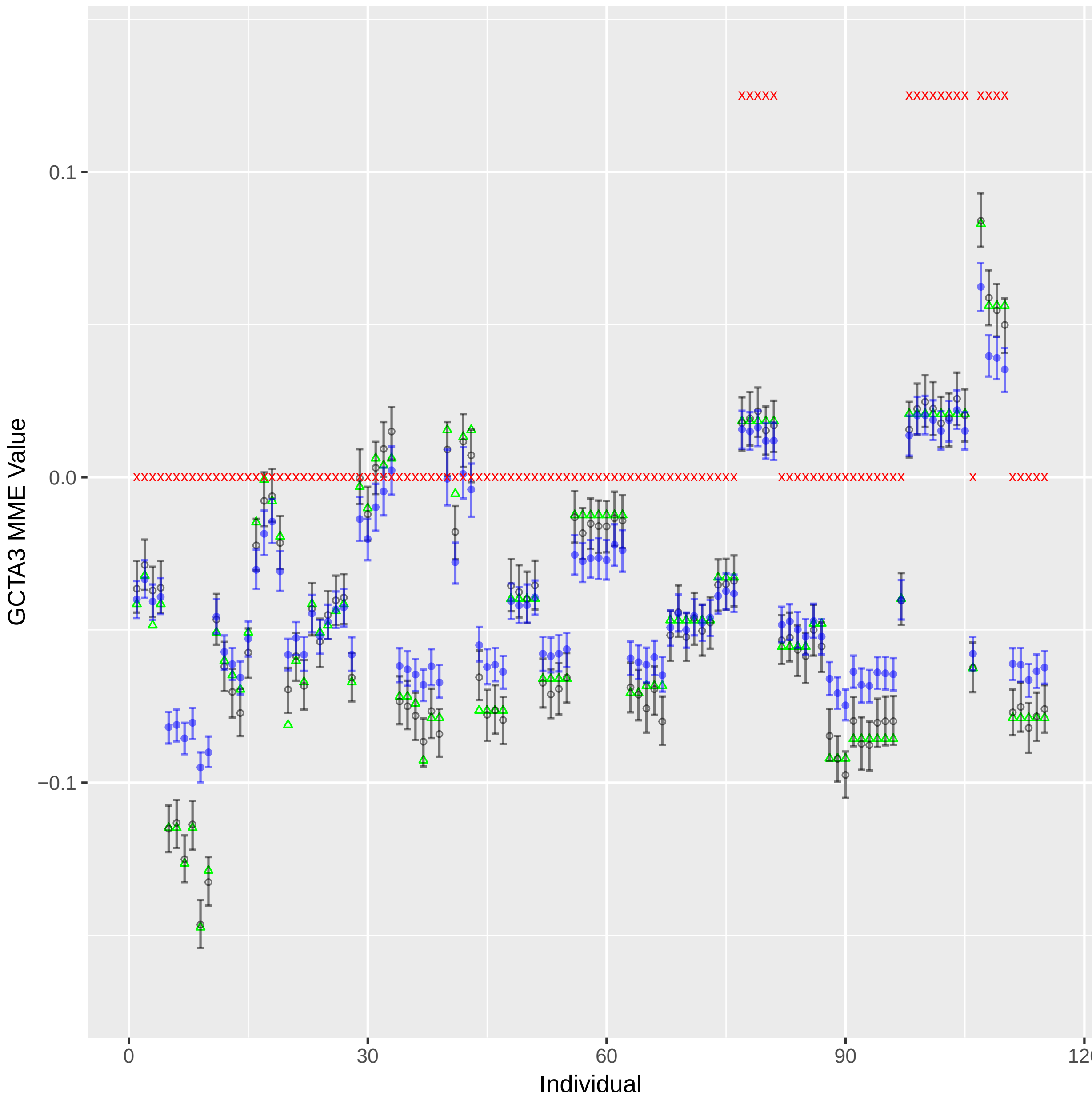

Figure 2.4: GCTA^{III} MME (RM and MR) Estimates for 115 Individuals and their Pedigree F_j and Parametric Pedigree Values. **Black, Hollow Circle**: GCTA^{III} MME RM point estimates and 95% bootstrap confidence intervals. **Blue, Solid Circle**: GCTA^{III} MME MR point estimates and 95% bootstrap confidence intervals. **Red, Cross**: F_j parameter determined from pedigree. **Green, Hollow Triangle**: Parameter $(F_j - 2\psi_j + \theta_S)/(1 - \theta_S)$ determined from pedigree.

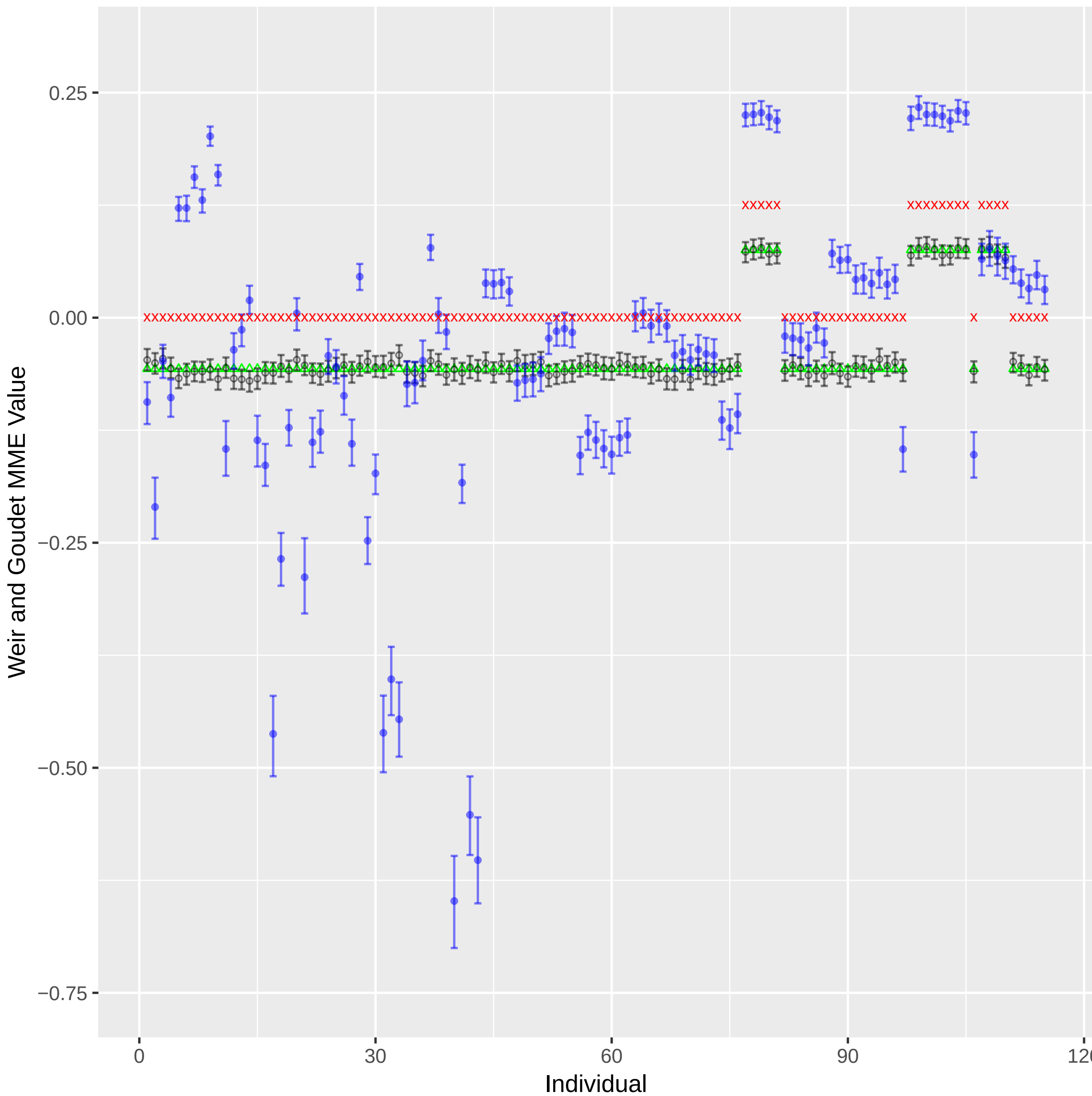

Figure 2.5: Weir and Goudet MME (RM and MR) Estimates for 115 Individuals and their Pedigree F_j and β_j Values. **Black, Hollow Circle**: Weir and Goudet MME RM point estimates and 95% bootstrap CI's. **Blue, Solid Circle**: Weir and Goudet MME MR point estimates and 95% bootstrap CI's. **Red, Cross**: F_j parameter determined from pedigree. **Green, Hollow Triangle**: Parameter β_j determined from pedigree.

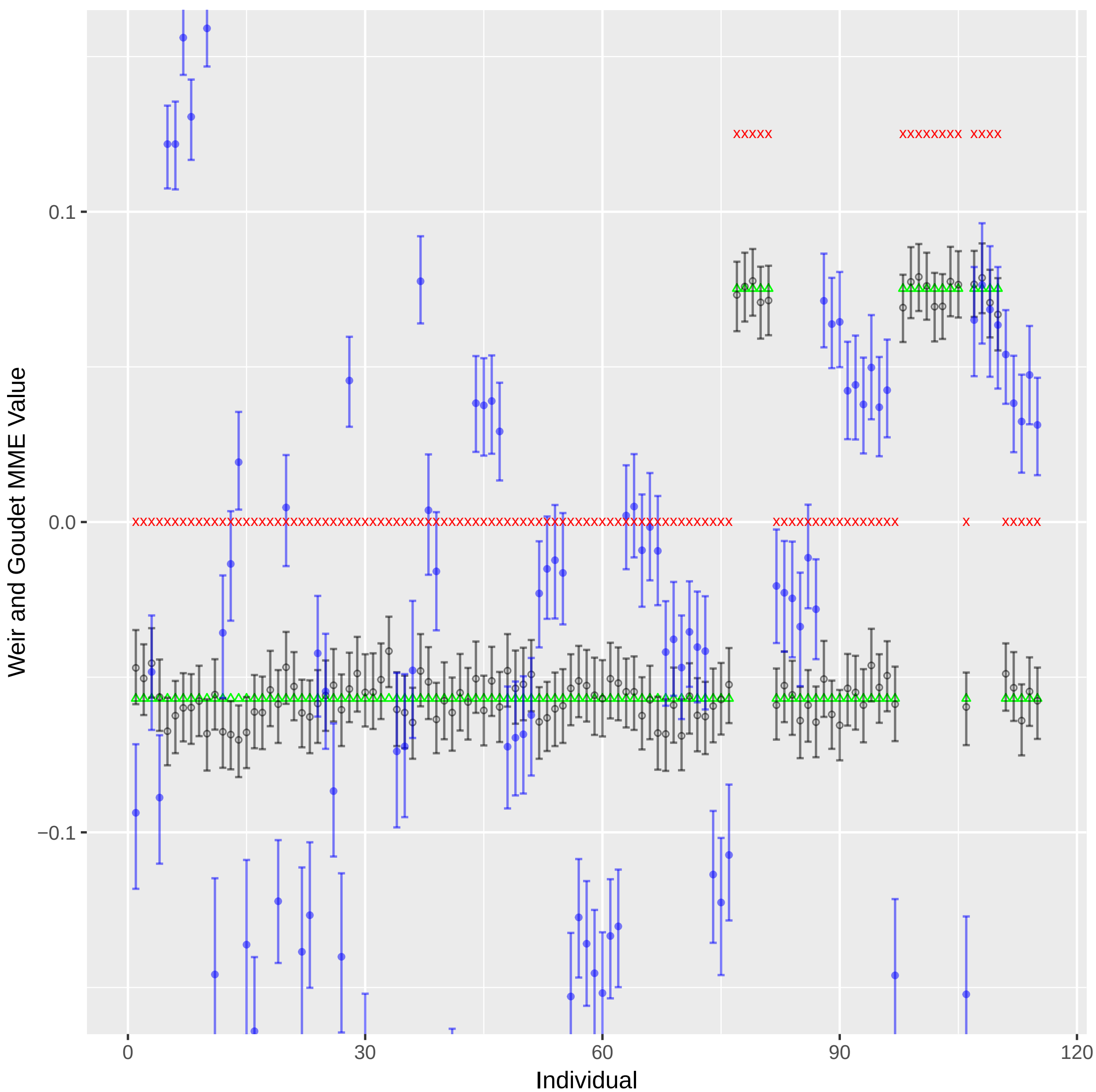

Figure 2.6: Weir and Goudet MME (RM and MR) Estimates for 115 Individuals and their Pedigree F_j and β_j Values with y-axis constrained to [-0.15,0.15]. **Black, Hollow Circle**: Weir and Goudet MME RM point estimates and 95% bootstrap CI's. **Blue, Solid Circle**: Weir and Goudet MME MR point estimates and 95% bootstrap CI's. **Red, Cross**: F_j parameter determined from pedigree. **Green, Hollow Triangle**: Parameter β_j determined from pedigree.

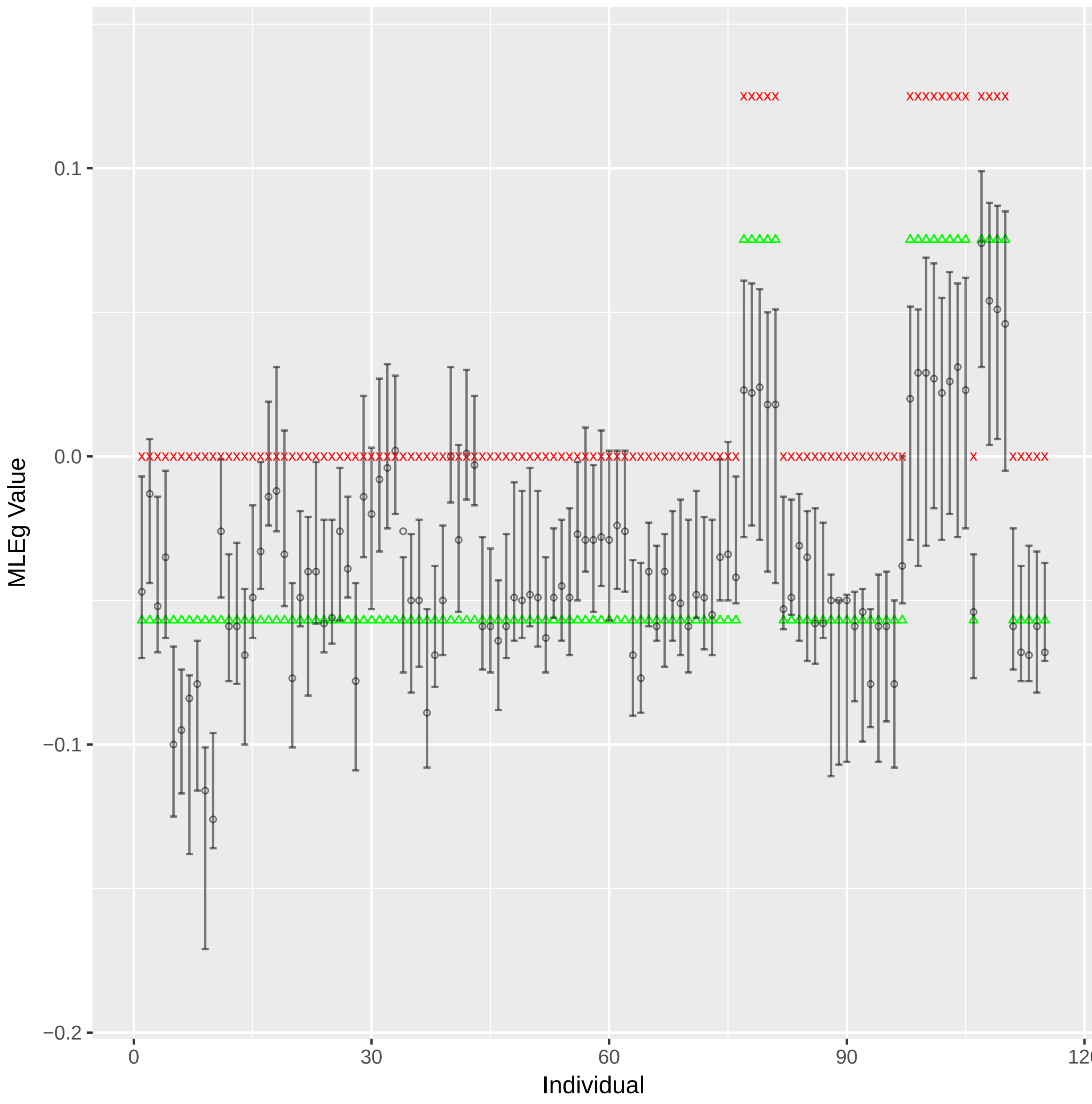

Figure 2.7: Hall MLE 1 (i.e. MLEg) Estimates for 115 Individuals and their Pedigree F_j and β_j Values. **Black, Hollow Circle**: Hall MLE 1 (i.e. MLEg) point estimates and 95% bootstrap confidence intervals. **Red, Cross**: F_j parameter determined from pedigree. **Green, Hollow Triangle**: Parameter β_j determined from pedigree.

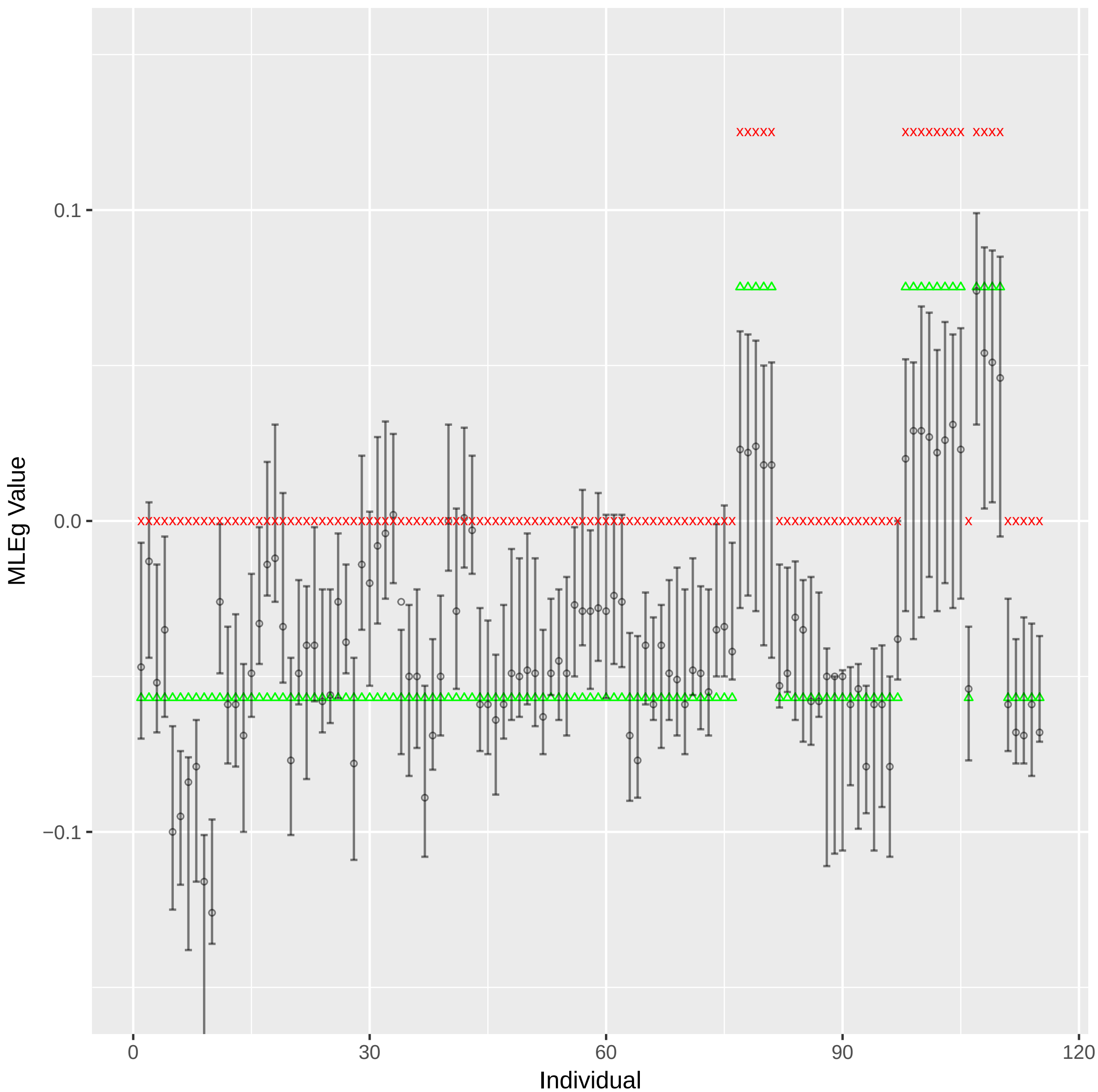

Figure 2.8: Hall MLE 1 (i.e. MLEg) Estimates for 115 Individuals and their Pedigree F_j and β_j Values with y-axis constrained to [-0.15,0.15]. **Black, Hollow Circle**: Hall MLE 1 (i.e. MLEg) point estimates and 95% bootstrap confidence intervals. **Red, Cross**: F_j parameter determined from pedigree. **Green, Hollow Triangle**: Parameter β_j determined from pedigree.

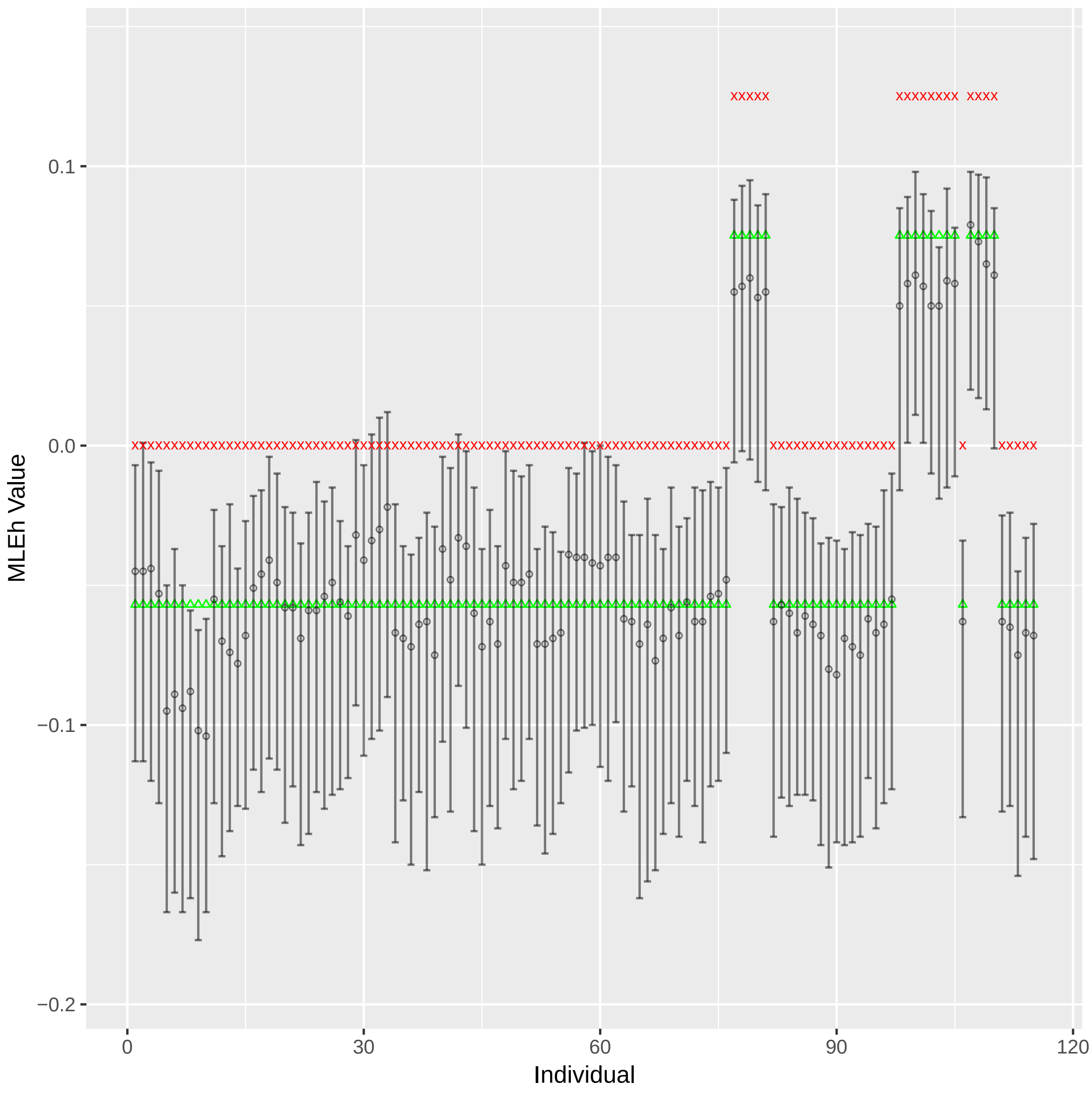

Figure 2.9: MLEh Estimates for 115 Individuals and their Pedigree F_j and β_j Values. **Black, Hollow Circle**: MLEh point estimates and 95% bootstrap confidence intervals. **Red, Cross**: F_j parameter determined from pedigree. **Green, Hollow Triangle**: Parameter β_j determined from pedigree.

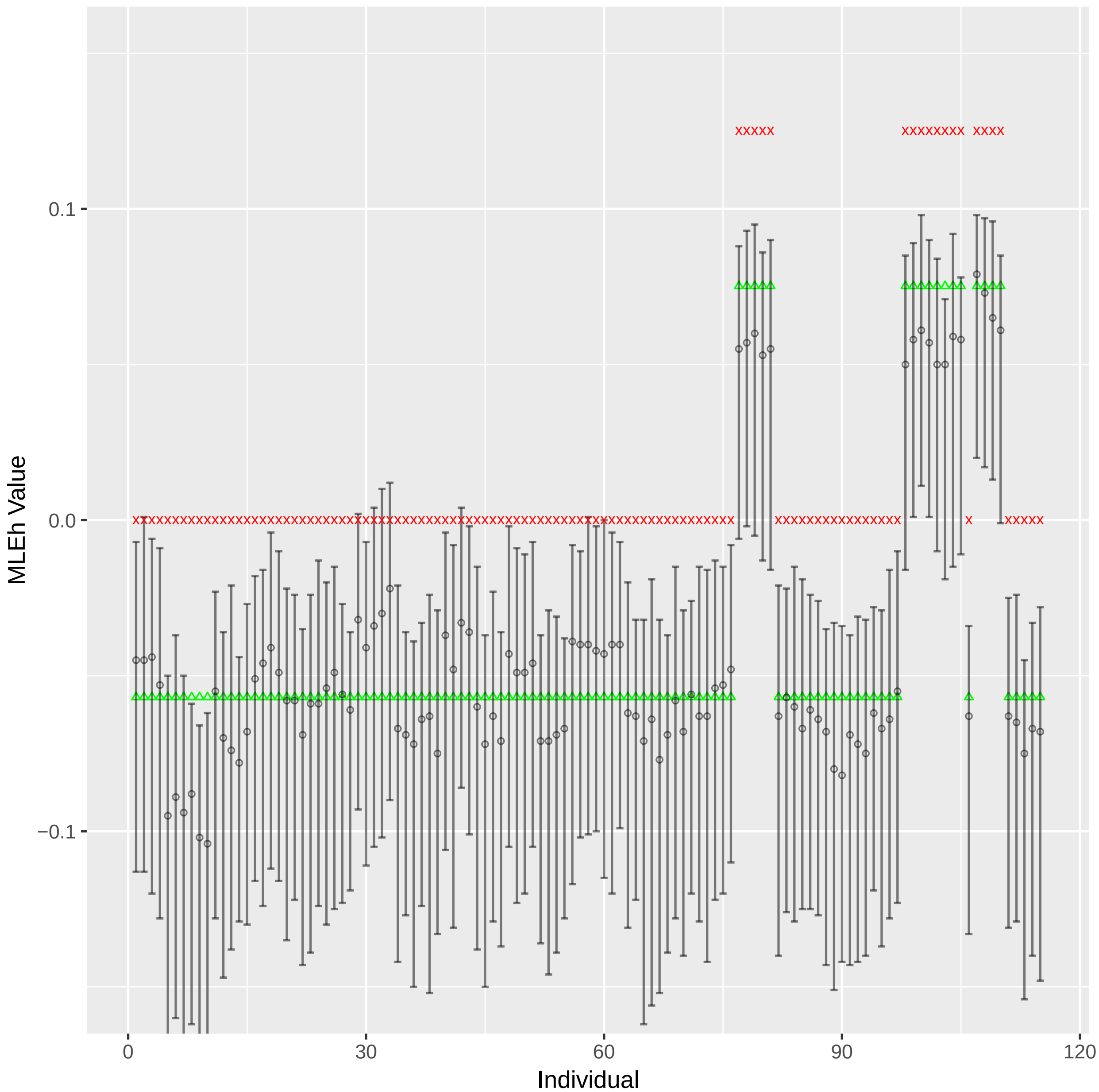

Figure 2.10: MLEh Estimates for 115 Individuals and their Pedigree F_j and β_j Values with y-axis constrained to [-0.15,0.15]. **Black, Hollow Circle**: MLEh point estimates and 95% bootstrap confidence intervals. **Red, Cross**: F_j parameter determined from pedigree. **Green, Hollow Triangle**: Parameter β_j determined from pedigree.

2.5 Conclusion

We have reviewed estimators that previous authors introduced as estimators of F_j or β_j: the weighted Weir and Goudet $\hat{\beta}_{jw}$ (ratio of means) MME and unweighted Weir and Goudet $\hat{\beta}_{ju}$ (mean of ratios) MME were introduced as β_j estimators, and the weighted GCTA$^{I}_{jw}$ (ratio of means) MME, unweighted GCTA$^{I}_{ju}$ (mean of ratios) MME, weighted GCTA$^{III}_{jw}$ (ratio of means) MME, unweighted GCTA$^{III}_{ju}$ (mean of ratios) MME, Hall et al.'s simple MME, and Hall MLE 1 (MLEg) were introduced as F_j estimators. We introduced a new likelihood estimator for β_j, which we call MLEh.

We have shown that the second moment of each sample allele frequency is not simply a function of a single individual j's inbreeding coefficient, but a function of inbreeding coefficients of all n individuals in the sample as well coancestry coefficients between all distinct pairs of individuals in the sample. In Hall et al.'s first model, one assumes in the likelihood function that sample allele frequencies $\tilde{p}_l$'s are equal to population frequencies π_l's, and $\tilde{p}_l^2$ arises in the likelihood function, which results in bias of the MLE $\hat{F}_j$ for individual j's inbreeding coefficient F_j. We give a grid search algorithm for obtaining Hall et al.'s first model's MLE $\hat{F}_j$, and apply it to simulated data for $n = 115$ individuals, showing that MLE estimates do not track closely with pedigree F_j values, consistent with the MLE being biased for F_j.

We find that none of these existing estimators are unbiased for F_j on simulated data, and none of these estimators tracked closely with pedigree F_j parameter values. The bias of these estimators for F_j can be explained by their imperfect information or lack of information about ancestral population reference allele frequencies π_l's. Sample allele frequencies $\tilde{p}_l$'s present in Hall et al.'s simple MME, GCTA^{I} and GCTA^{III} MMEs, Hall et al.'s first model's likelihood function, and our MLEh's likelihood function are imperfect proxies of π_l's, as illustrated by $\mathcal{E}(\tilde{p}_l^2)$ not being π_l^2, but a function of inbreeding coefficients and coancestry coefficients involving all sample individuals. Weir and Goudet's MME does not employ a sample allele frequency $\tilde{p}_l$, and does not have any other source of information about π_l's. As E. Thompson put it, "There is no absolute measure of ibd; ibd is always relative to some ancestral reference population." (Thompson, 2013) Without knowing reference allele frequencies π_l's for this ancestral reference population, we cannot estimate inbreeding coefficients F_j's without bias.

From Bias^2, MSE, coverage, and tracking of estimates, we see that all estimators are better at estimating their pedigree parameters (β_j for Weir and Goudet MME RM and MR, MLEg, and MLEh, $(F_j - 4\psi_j + 3\theta_S)/(1-\theta_S)$ for GCTA^I RM and MR, and $(F_j - 2\psi_j + \theta_S)/(1-\theta_S)$ for GCTA^{III} RM and MR) than at estimating F_j. Of the estimators with pedigree parameters β_j, the Weir and Goudet MME RM and MLEh both had close to 95% coverage for β_j in our simulation study, and had the lowest β_j MSE values. Our MLEh is an improvement over the existing MLEg, in that MLEh had lower β_j MSE and coverage much closer to 95%. However, the Weir and Goudet MME RM had lower β_j Bias^2, lower β_j MSE, and was faster to compute than MLEh, so we advocate usage of the Weir and Goudet $\hat{\beta}_{jw}$ MME for estimation of β_j.

2.6 *Appendix:*

2.6.1 Allele Frequencies

For allele $k = 1, 2$ at locus l in individual j, define indicator variable x_{jkl} to be 1 if it is the reference allele and 0 otherwise. Parameters $\pi_l, F_j, \theta_{jj'}$ were defined in Weir and Hill (2002) as parameters that make the following relations hold:

$$\mathcal{E}(x_{jkl}) = \pi_l, \tag{2.13}$$

$$\mathcal{E}(x_{jkl}^2) = \pi_l, \tag{2.14}$$

$$\mathcal{E}(x_{jkl}x_{jk'l}) = (\pi_l)^2 + (\pi_l)(1-\pi_l)(F_j), \quad k \neq k', \text{ and} \tag{2.15}$$

$$\mathcal{E}(x_{jkl}x_{j'k'l}) = (\pi_l)^2 + (\pi_l)(1-\pi_l)(\theta_{jj'}), \quad j \neq j'. \tag{2.16}$$

Expectations are over replicates of the evolutionary process leading to population j.

Functions of locus l's sample allele proportion $\tilde{p}_l$ have expectations given below. From above Equations 2.13, 2.14, 2.15, and 2.16, we have:

$$\tilde{p}_l := \frac{1}{2n}\Big(\sum_{j=1}^{n}\sum_{k=1}^{2} x_{jkl}\Big)$$
$$\mathcal{E}(\tilde{p}_l) = \pi_l \tag{2.17}$$

$$\tilde{p}_l^2 = \frac{1}{4n^2}\left(\sum_{j=1}^{n}\sum_{k=1}^{2} x_{jkl}^2 + \sum_{j=1}^{n}\sum_{\substack{k=1\\k\neq k'}}^{2}\sum_{k'=1}^{2} x_{jkl}x_{jk'l} + \sum_{\substack{j=1\\j\neq j'}}^{n}\sum_{j'=1}^{n}\sum_{k=1}^{2}\sum_{k'=1}^{2} x_{jkl}x_{j'k'l}\right)$$

$$\begin{aligned}
\mathcal{E}(\tilde{p}_l^2) &= \frac{1}{4n^2}\left\{2n\pi_l + 2\sum_{j=1}^{n}[\pi_l^2 + \pi_l(1-\pi_l)F_j] + 4\sum_{\substack{j=1\\j\neq j'}}^{n}\sum_{j'=1}^{n}[\pi_l^2 + \pi_l(1-\pi_l)\theta_{jj'}]\right\} \\
&= \frac{1}{4n^2}\left\{2n\pi_l + 2n[\pi_l^2 + \pi_l(1-\pi_l)F_W] + 4n(n-1)[\pi_l^2 + \pi_l(1-\pi_l)\theta_S]\right\} \\
&= \pi_l^2 + \pi_l(1-\pi_l)\left[\theta_S + \frac{1}{2n}\left(1 + F_W - 2\theta_S\right)\right].
\end{aligned} \tag{2.18}$$

2.6.2 Allele Dosages

Allele dosages are $X_{jl} = x_{j1l} + x_{j2l}$. Using expectations 2.14 and 2.15, allele dosages have expectations given below.

$$\mathcal{E}(X_{jl}) = \mathcal{E}(x_{j1l} + x_{j2l}) = 2\pi_l \tag{2.19}$$

$$\begin{aligned}
\mathcal{E}(X_{jl}^2) &= \mathcal{E}\Big(\big((x_{j1l} + x_{j2l})(x_{j1l} + x_{j2l})\big)\Big) \\
&= \mathcal{E}(x_{j1l}^2 + 2x_{j1l}x_{j2l} + x_{j2l}^2) \\
&= (\pi_l) + 2\Big((\pi_l)^2 + (\pi_l)(1-\pi_l)F_j\Big) + (\pi_l) \\
&= 2(\pi_l)(1+\pi_l) + 2\pi_l(1-\pi_l)F_j.
\end{aligned} \tag{2.20}$$

Using Equation 2.16, we obtain

$$\begin{aligned}
\mathcal{E}(X_{jl}X_{j'l}) &= \mathcal{E}\Big((x_{j1l} + x_{j2l})(x_{j'1l} + x_{j'2l})\Big) \\
&= \mathcal{E}\Big(x_{j1l}x_{j'1l} + x_{j1l}x_{j'2l} + x_{j2l}x_{j'1l} + x_{j2l}x_{j'2l}\Big) \\
&= 4\Big(\pi_l^2 + \pi_l(1-\pi_l)\theta_{jj'}\Big)
\end{aligned} \tag{2.21}$$

which will be useful in deriving Equation 3.15.

The following will be useful in deriving Equation 3.18:

$$\begin{aligned}
X_{jl}\tilde{p}_l &= (x_{j1l} + x_{j2l})\Big[\sum_{j'=1}^{n}(x_{j'l} + x_{j'2l})/(2n)\Big] \\
&= \frac{1}{2n}\left[(x_{j1l} + x_{j2l})^2 + \sum_{\substack{j'=1 \\ j'\neq j}}^{n}(x_{j1l} + x_{j2l})(x_{j'1l} + x_{j'2l})\right].
\end{aligned}$$

If $\psi_j = \sum_{j'=1, j'\neq j}^{n} \theta_{jj'}/(n-1)$ is the average coancestry of individual j with every other member of the sample, using Equations 2.20 and 2.16,

$$
\begin{aligned}
\mathcal{E}(X_{jl}\tilde{p}_l) &= \frac{1}{2n}\left[\mathcal{E}(X_{jl}^2) + \sum_{\substack{j'=1\\ j'\neq j}}^{n} \mathcal{E}\Big((x_{j1l}+x_{j2l})(x_{j'1l}+x_{j'2l})\Big)\right] \\
&= \frac{1}{2n}\{\Big[2(\pi_l)(1+\pi_l) + 2(\pi_l)(1-\pi_l)F_j\Big] + \sum_{j'=1, j'\neq j}^{n} \Big[4\Big((\pi_l)^2 + (\pi_l)(1-\pi_l)\theta_{jj'}\Big)\Big]\} \\
&= \frac{1}{2n}\{\Big[2(\pi_l)(1+\pi_l) + 2(\pi_l)(1-\pi_l)F_j\Big] + 4(n-1)(\pi_l)^2 + 4(\pi_l)(1-\pi_l) \sum_{j'=1, j'\neq j}^{n} \theta_{jj'}\} \\
&= \frac{1}{2n}\{[2\pi_l + 2\pi_l^2 + 2\pi_l(1-\pi_l)F_j] + 4(n-1)[\pi_l^2 + \pi_l(1-\pi_l)\psi_j]\} \\
&= 2\pi_l^2 + 2\pi_l(1-\pi_l)\psi_j + \frac{1}{n}\pi_l(1-\pi_l)(1+F_j-2\psi_j) \qquad (2.22)
\end{aligned}
$$

2.6.3 Matching Proportions

Within individual j, the matching proportion at locus l, $\tilde{M}_{j_l} = (X_{jl}-1)^2$, has expectation

$$
\begin{aligned}
\mathcal{E}(\tilde{M}_{j_l}) &= \mathcal{E}((X_{jl}-1)^2) = \mathcal{E}(X_{jl}^2 - 2X_{jl} + 1) \\
&= \mathcal{E}(X_{jl}^2) - (2\mathcal{E}X_{jl}) + 1 \\
&= \Big[2(\pi_l)(1+\pi_l) + 2\pi_l(1-\pi_l)F_j\Big] - 2\Big[2\pi_l\Big] + 1 \\
&= 1 - 2(\pi_l)(1-\pi_l)(1-F_j) \\
\mathcal{E}(1-\tilde{M}_{j_l}) &= 2\pi_l(1-\pi_l)(1-F_j), \qquad (2.23)
\end{aligned}
$$

and, adding over loci:

$$
\mathcal{E}[\sum_l (1-\tilde{M}_{jl})] \quad = \quad H_e(1-F_j),
$$

where $H_e := \sum_l 2\pi_l(1-\pi_l)$.

Between pairs of individuals j, j' the matching proportion $\tilde{M}_{jj'}$ and its expectation are

$$\tilde{M}_{jj'_l} = \frac{1}{2}[1 + (X_{jl} - 1)(X_{j'l} - 1)]$$
$$= \frac{1}{2}[1 + X_{jl}X_{j'l} - X_{j'l} - X_{jl} + 1],$$

and, averaging over distinct pairs of individuals from the same population of n individuals,

$$\tilde{M}_{Sl} = \frac{1}{n(n-1)} \sum_{\substack{j=1 \\ j \neq j'}}^{n} \sum_{j'=1}^{n} \tilde{M}_{jj'_l}$$
$$\mathcal{E}(\tilde{M}_{Sl}) = \frac{1}{n(n-1)} \sum_{\substack{j=1 \\ j \neq j'}}^{n} \sum_{j'=1}^{n} \mathcal{E}((\tilde{M}_{jj'_l})$$
$$= \frac{1}{n(n-1)} \sum_{\substack{j=1 \\ j \neq j'}}^{n} \sum_{j'=1}^{n} \left[1 - 2\pi_l(1-\pi_l)(1-\theta_{jj'})\right]$$
$$= 1 - 2\pi_l(1-\pi_l)(1-\theta_S) \tag{2.24}$$
$$\mathcal{E}(1 - \tilde{M}_{Sl}) = 2\pi_l(1-\pi_l)(1-\theta_S). \tag{2.25}$$

Adding over loci and considering Equation 2.25, we have

$$\mathcal{E}[\sum_l (1 - \tilde{M}_{Sl})] = \sum_l (\mathcal{E}[1 - \tilde{M}_{Sl}])$$
$$= \sum_l (2\pi_l(1-\pi_l)(1-\theta_S))$$
$$= (H_e)(1-\theta_S).$$

2.6.4 Match Statistics in Terms of Heterozygosities

Sample match proportions can be expressed in terms of heterozygosities. For a heterozygous individual j at locus l, $(1 - \tilde{M}_{jl}) = 1$, so

$$\sum_{l=1}^{L} (1 - \tilde{M}_{jl}) = \tilde{H}_{jo}. \tag{2.26}$$

The match statistic $\tilde{M}_{Sl}$ considers pairs of individuals in the same population. Using Equation 2.8, noting that a sample reference allele frequency at locus l is $\tilde{p}_l = \sum_{j=1}^{n} X_{jl}/(2n)$, using the indicator of whether individual j is observed to be heterozygous at locus l $\tilde{H}_{jo_l} = (1-\tilde{M}_{jl}) = (1 - X_{jl}^2 + 2X_{jl} - 1)$, and recalling definitions from Section 2.3.4.1 that $\tilde{H}_{jo} = \sum_{l=1}^{L} \tilde{H}_{jo_l}$ and $\tilde{H}_o = \sum_{j=1}^{n} \tilde{H}_{jo}/n$, we can relate $\sum_{l=1}^{L}(1-\tilde{M}_{Sl})$ to heterozygosities:

$$\begin{aligned}
\sum_{l=1}^{L}(1-\tilde{M}_{Sl}) &= \sum_{l=1}^{L}\left[1 - \frac{1}{2n(n-1)}\sum_{\substack{j=1 \\ j\neq j'}}^{n}\sum_{j'=1}^{n}[1+(X_{jl}-1)(X_{j'l}-1)]\right] \\
&= \sum_{l=1}^{L}\left\{1 - \frac{1}{2n(n-1)}\left[(\sum_{j=1}^{n}X_{jl})^2 - \sum_{j=1}^{n}X_{jl}^2 - 2(n-1)\sum_{j=1}^{n}X_{jl} + 2n(n-1)\right]\right\} \\
&= \sum_{l=1}^{L}\left\{1 - \frac{1}{2n(n-1)}\left[4n^2\tilde{p}_l^2 - 4n\tilde{p}_l + \sum_{j=1}^{n}\tilde{H}_{jo_l} - 4n(n-1)\tilde{p}_l + 2n(n-1)\right]\right\} \\
&= \frac{n}{n-1}\left[\sum_{l=1}^{L}2\tilde{p}_l(1-\tilde{p}_l) - \frac{1}{2n^2}\sum_{j=1}^{n}\tilde{H}_{jo}\right] \\
&= \frac{n}{n-1}\left(\tilde{H}_e - \frac{1}{2n}\tilde{H}_o\right).
\end{aligned}$$

2.6.5 Difference Between Simple MME and Weir and Goudet MME Converges Almost Surely to 0 *Under Certain Conditions*

From relations and definitions mentioned in Section 2.3.4.1, such as Equation 2.4 giving $\hat{\beta}_{j_{\mathrm{Sim.}}}$ in terms of heterozygosities, and Equations 2.10 and 2.26 giving heterozygosities in terms of match statistics, the simple MME is

$$
\begin{aligned}
\hat{\beta}_{j_{\text{Sim.}}} &= 1 - \tilde{H}_{jo}/\tilde{H}_e \\
&= 1 - \sum_l (1 - \tilde{M}_{jl})/[\frac{n-1}{n}\sum_{l=1}^{L}(1 - \tilde{M}_{Sl}) + \frac{1}{2n}\sum_{j=1}^{n}\frac{\sum_{l=1}^{L}(1 - \tilde{M}_{jl})}{n}].
\end{aligned}
$$

The first equation is Equation 2.4, and the second equation follows from Equations 2.10 and 2.26. As a shorthand, we use $\sum_l$ to mean $\sum_{l=1}^{L}$. Now recall from Equation 2.9,

$$
\hat{\beta}_{jw} = 1 - \sum_{l=1}^{L}(1 - \tilde{M}_{jl})/\sum_{l=1}^{L}(1 - \tilde{M}_{Sl})
$$

The difference is

$$
\begin{aligned}
\hat{\beta}_{j_{\text{Sim.}}} - \hat{\beta}_{jw} &= \Big[1 - \sum_l (1 - \tilde{M}_{jl})/[\frac{n-1}{n}\sum_l (1 - \tilde{M}_{Sl}) + \frac{1}{2n}\sum_{j=1}^{n}\frac{\sum_l (1 - \tilde{M}_{jl})}{n}]\Big] \\
&\quad - \Big[1 - \sum_l (1 - \tilde{M}_{jl})/\sum_l (1 - \tilde{M}_{Sl})\Big] \\
&= \Big[\sum_l (1 - \tilde{M}_{jl})/\sum_l (1 - \tilde{M}_{Sl})\Big] \\
&\quad - \Big[\sum_l (1 - \tilde{M}_{jl})/[\frac{n-1}{n}\sum_l (1 - \tilde{M}_{Sl}) + \frac{1}{2n}\sum_{j=1}^{n}\frac{\sum_l (1 - \tilde{M}_{jl})}{n}]\Big]
\end{aligned} \tag{2.27}
$$

Introducing means by dividing by L, we have:

$$
\begin{aligned}
\hat{\beta}_{j_{\text{Sim.}}} - \hat{\beta}_{jw} &= \Big[[\sum_l (1 - \tilde{M}_{jl})/L]/[\sum_l (1 - \tilde{M}_{Sl})/L]\Big] \\
&\quad - \Big[[\sum_l (1 - \tilde{M}_{jl})/L]/[\frac{n-1}{n}[\sum_l (1 - \tilde{M}_{Sl})/L] + \frac{1}{2n}\sum_{j=1}^{n}\frac{[\sum_l (1 - \tilde{M}_{jl})/L]}{n}]\Big]
\end{aligned}
$$

If different loci's match statistics $[1-\tilde{M}_{jl}, 1-\tilde{M}_{Sl}]$ are independent and identically distributed, then by the Strong Law of Large Numbers (SLLN) (page 6 of Vaart (1998)), we conclude that $[\sum_l (1-\tilde{M}_{jl})/L, \sum_l (1-\tilde{M}_{Sl})/L]$ converges almost surely to $[\mathcal{E}(1-\tilde{M}_{jl}), \mathcal{E}(1-\tilde{M}_{Sl})]$ as $L \to \infty$, denoted $[\sum_l (1-\tilde{M}_{jl})/L, \sum_l (1-\tilde{M}_{Sl})/L] \to_{a.s.} [\mathcal{E}(1-\tilde{M}_{jl}), \mathcal{E}(1-\tilde{M}_{Sl})]$. Note that $[1-\tilde{M}_{jl}, 1-\tilde{M}_{Sl}]$ being identically distributed for all $l \in \{1, ..., L\}$ implies that π_l is the same quantity π for all l, $\mathcal{E}(1-\tilde{M}_{jl})$ is the same quantity $\mathcal{E}(1-\tilde{M}_j)$ for all l, and $\mathcal{E}(1-\tilde{M}_{Sl})$ is the same quantity $\mathcal{E}(1-\tilde{M}_S)$ for all l.

For ease of notation, let us define the following quantities,

$$X_L := \sum_l (1-\tilde{M}_{jl})/L$$
$$Y_L := \sum_l (1-\tilde{M}_{Sl})/L,$$

so that

$$\begin{aligned}\hat{\beta}_{j_{\text{Sim.}}} - \hat{\beta}_{jw} &= \frac{X_L}{Y_L} - \frac{X_L}{\frac{n-1}{n} Y_L + \frac{1}{2n}\sum_{j=1}^{n} \frac{X_L}{n}} \\ &= g(X_L, Y_L)\end{aligned}$$

where function $g(x,y) := x/y - x/\left[((n-1)/(n))y + 1/(2n)\sum_{j=1}^{n} x/n\right]$ is continuous at every point $[x, y] \in \mathbb{R}^2$ such that $x \neq 0$ and $y \neq 0$. By the Continuous Mapping Theorem (Theorem 2.3 of Vaart (1998)), we know that if a function $f : \mathbb{R}^k \mapsto \mathbb{R}^m$ is a continuous function at every point of a set C such that $P(X \in C) = 1$, and if a sequence of random variables $X_n \to_{a.s.} X$, then $f(X_n) \to_{a.s.} f(X)$. The n in this statement of the Continuous Mapping Theorem is like our L, and the function f is like our g. Because $[X_L, Y_L] \to_{a.s.} [\mathcal{E}(1-\tilde{M}_j), \mathcal{E}(1-\tilde{M}_S)]$ by SLLN, by the Continuous Mapping Theorem, if $[\mathcal{E}(1-\tilde{M}_j), \mathcal{E}(1-\tilde{M}_S)]$ does not have a 0 element (i.e. is not a point where function g is discontinuous), then $g(X_L, Y_L) \to_{a.s.} g(\mathcal{E}(1-\tilde{M}_j), \mathcal{E}(1-\tilde{M}_S))$. Expectations $\mathcal{E}(1-\tilde{M}_{jl})$ and $\mathcal{E}(1-\tilde{M}_{Sl})$ for any locus $l \in \{1, ..., L\}$ are derived in Section 2.6.3 as $2(\pi_l)(1-\pi_l)(1-F_j)$ and $2(\pi_l)(1-\pi_l)(1-\theta_S)$ respectively.

Hence, $[\mathcal{E}(1-\tilde{M}_j), \mathcal{E}(1-\tilde{M}_S)]$ not having a 0 element is equivalent to $\pi \neq 0$, $\pi \neq 1$, $F_j \neq 1$ and $\theta_S \neq 1$.

Explicitly writing out $g(\mathcal{E}(1-\tilde{M}_j), \mathcal{E}(1-\tilde{M}_S))$, we see that it can be seen as a function of n:

$$g(\mathcal{E}(1-\tilde{M}_j), \mathcal{E}(1-\tilde{M}_S)) = \frac{\mathcal{E}(1-\tilde{M}_j)}{\mathcal{E}(1-\tilde{M}_S)} - \frac{\mathcal{E}(1-\tilde{M}_j)}{\frac{n-1}{n}\mathcal{E}(1-\tilde{M}_S) + \frac{1}{2n}\sum_{j=1}^{n}\frac{\mathcal{E}(1-\tilde{M}_j)}{n}}.$$

Taking the limit as $n \to \infty$, $g(\mathcal{E}(1-\tilde{M}_j), \mathcal{E}(1-\tilde{M}_S)) \to 0$.

In summary, we have, if $[1-\tilde{M}_{jl}, 1-\tilde{M}_{Sl}]$ are independent and identically distributed for all loci $l \in \{1, ..., L\}$ (implying π_l is the same quantity π for all l, $\mathcal{E}(1-\tilde{M}_{jl})$ is the same quantity $\mathcal{E}(1-\tilde{M}_j)$ for all l, and $\mathcal{E}(1-\tilde{M}_{Sl})$ is the same quantity $\mathcal{E}(1-\tilde{M}_S)$ for all l) and $[\mathcal{E}(1-\tilde{M}_j), \mathcal{E}(1-\tilde{M}_S)] = [2(\pi)(1-\pi)(1-F_j), 2(\pi)(1-\pi)(1-\theta_S)]$ does not have a 0 element (i.e. $\pi \neq 0$, $\pi \neq 1$, $F_j \neq 1$, and $\theta_S \neq 1$), then the difference $\hat{\beta}_{j_{\text{Sim.}}} - \hat{\beta}_{jw}$ between the simple MME and Weir and Goudet estimators converges almost surely to 0 as $L \to \infty$ and $n \to \infty$.

Chapter 3

ESTIMATION OF INDIVIDUAL PAIR COANCESTRY COEFFICIENTS AND POPULATION-SPECIFIC F_{ST}

3.1 Introduction

In Chapter 2, we considered estimation of the inbreeding coefficient, a measure of the relationship between the two alleles an individual carries at each locus. We showed that estimators are affected by all the individuals in a study that are used to provide information about allele frequencies, but our focus was on estimating $F_j^i = \theta_{j1,j2}^{ii}$, the ibd probability between alleles 1 and 2 in the same individual j, so necessarily in the same population i.

We now extend our treatment to estimating $\theta_{jj'}^{ii'}$, the ibd probability between alleles in different individuals j and j' from populations i and i' respectively, where $i = i'$ (the individuals are from the same population) or $i \neq i'$ (the individuals are from different populations), and each individual may be inbred ($F_j^i \neq 0$ or $F_{j'}^{i'} \neq 0$) or not ($F_j^i = 0$ or $F_{j'}^{i'} = 0$), i.e. $F_j^i \geq 0$ and $F_{j'}^{i'} \geq 0$. If alleles taken from two individuals in the same population are identical by descent (ibd), the individuals are said to be *related* and the probability of this event is called the *coancestry* or *kinship coefficient* of those individuals. If two alleles taken from the same population have a higher probability of being ibd than alleles from two different populations, then there is said to be *population structure*.

Population structure can be measured by population-specific F_{ST}, defined as $\beta^i = (\theta^i - \theta_B)/(1 - \theta_B)$, where $\theta^i = \sum_{j=1}^{n} \sum_{j'=1, j \neq j'}^{n} \theta_{jj'}^{ii}/(n^2 - n)$ is the probability that two distinct alleles sampled without replacement from a sample of n individuals from population i are ibd, and $\theta_B := \sum_{i=1}^{r} \sum_{i'=1, i' \neq i}^{r} \theta^{ii'}/(r^2 - r)$ is the average of probabilities $\theta^{ii'}$'s for an allele from one population to be ibd to an allele from a different population, where the average is taken over all pairs of different populations in a collection of r populations. Intuitively, β^i

is a measure of the genetic relatedness θ^i of population i relative to the genetic relatedness θ_B of a collection of populations of which i is a member. When we consider the definition of β^i, the alleles belonging to population i are those in what we call a "target set" of alleles, whose genetic relatedness we are interested in measuring relative to the genetic relatedness of another collection of alleles.

Relatedness and population structure have a common basis and we will show that estimation of kinship and population structure parameters can have a single formal structure. We note that many published accounts of estimating population structure parameters assume that individuals from different populations are unrelated just as we showed in Chapter 2 that many published estimators of the inbreeding coefficient assume that all pairs of individuals in a study are unrelated.

Many population genetic activities, ranging from forensic identification to association mapping to evolutionary studies, rely on appropriate estimates of population structure or relatedness. In forensic science, the match probability, the chance that an untyped person (e.g. perpetrator) will have a certain genotype given that a typed person (e.g. suspect who is not the perpetrator) has that genotype, assuming that the two people belong to the same population i, depends on θ^i (Scientific Working Group on DNA Analysis Methods, 2014; Buckleton et al., 2016; Balding and Nichols, 1995) as Equation 1 of Buckleton et al. (2016) shows. Since ibd can be estimated only in a relative sense, not an absolute sense, in practice, θ^i is replaced by an estimate of β^i, defined as $(\theta^i - \theta_B)/(1 - \theta_B)$. In genetic association mapping, the association statistics are affected by population structure to a degree quantified by the appropriate θ parameters, estimated as β parameters (Gravel et al., 2013). Kinship coeffiicents $\{\theta^{ii'}_{jj'}\}$ also arise in a genetic relatedness matrix used to allow for relatedness and inbreeding in genome-wide association studies (Conomos, 2014), and $\beta^{ii'}_{jj'} := (\theta^{ii'}_{jj'} - \theta_S)/(1 - \theta_S)$ may be useful instead of $\{\theta^{ii'}_{jj'}\}$ (Goudet, personal communication). In evolutionary genetics, outlying estimates of $\beta^i = (\theta^i - \theta_B)/(1 - \theta_B)$ can be used to identify genomic regions that may have different allele frequencies in different populations because of natural selection (Beaumont and Balding, 2004). Population-specific β^is have an advantage over the standard F_{ST}, i.e. β_W, the average over populations of β^is, in showing regions of differentiated frequencies that can be hidden by standard F_{ST} (Weir et al., 2005).

All applications of population structure or relatedness require recognition that quantities with an underlying meaning of allelic dependence are not defined in an absolute sense, but instead are made "relative to" some set of alleles other than the target set. We take the reference set to be the population from which study individuals have been sampled. If A^i_{jk} is allele k at some locus in individual j in (a sample from) population i, we assign the allelic indicator x^i_{jk} the value 1 if it is the reference allele type. The reference allele dosage, X^i_j for individual j in population i is $X^i_j = x^i_{j1} + x^i_{j2}$. The expected value of x^i_{jk}, where expectation is over replicates of the evolutionary history of that allele, is π, the probability a random allele is the reference type, regardless of which individual carries that allele or which population contains that individual. The essence of our treatment rests on the expectation of the products of two x's, or the probabilities that pairs of alleles are both the reference type. For alleles A^i_{jk} and $A^{i'}_{j'k'}$,

$$\mathcal{E}(x^i_{jk}x^{i'}_{j'k'}) \quad = \quad \pi^2 + \pi(1-\pi)\theta^{ii'}_{jk,j'k'} \tag{3.1}$$

For a single individual j in population i, the inbreeding coefficient F^i_j is the coancestry coefficient of the two alleles k, k': $F^i_j = \theta^{ii}_{jk,jk'}, k \neq k'$. The coancestry, or kinship, coefficient $\theta^{ii'}_{jj'}$ for individuals j, j' is the average of the four allele-pair coancestry coefficients $\theta^{ii'}_{jk,j'k'}$ over $k = 1, 2$ and $k' = 1, 2$. When j and j' are members of the same population i, the superscript pair ii' may be written simply as i or even omitted, and similarly for genetic relatedness measures involving a single individual j, so for example, F^i_j is the same as F_j.

The coancestry for pairs of populations is the average coancestry for pairs of alleles, one in each population: for populations i, i' the quantity $\theta^{ii'}_B$ is the average over all j, j' and all k, k' of $\theta^{ii'}_{jk,j'k'}$, and when $i = i'$ we write the average as θ^i_S or θ^i. For randomly-mating populations, $F^i_j = \theta^i_{jj'}$ for all j, j', and we can drop the j subscripts in referring to coancestries, θ^i within population i and $\theta^{ii'}$ between populations i, i'.

There is a large literature giving moment estimators of individual- and population-level coancestries, although very little recognition that the same estimators apply to individuals and to populations, and that population-level estimators are averages of individual-level estimators. There is a smaller literature for likelihood-based estimators of individual-level and population-level estimators.

We review estimators of the coancestry coefficient $\theta_{jj'}$ between individuals j and j' when individuals come from the same population. These include the simple method of moments estimator (MME), Weir and Goudet $\hat{\beta}_{jj'}$ weighted and unweighted MMEs, GCTA $\hat{\theta}_{jj'}$ weighted and unweighted MMEs, and MLEs. We also consider a novel dosage variance-covariance estimator. We evaluate many of these estimators on simulated genetic and pedigree data for one population.

We review estimators of population-i specific F_{ST} $\beta^i := (\theta^i_S - \theta_B)/(1 - \theta_B)$. These include Weir and Goudet's $\hat{\beta}^i$ MME (which allows $\theta_B > 0$), and existing Bayesian methods by Foll and Gaggiotti (2006) (which assumes $\theta_B = 0$) and Gunther and Coop (2013) (which allows $\theta_B > 0$). We develop a novel Bayesian estimator of β^i that allows two different populations' allele frequencies to be dependent random variables (allowing $\theta_B = \theta^{12} > 0$), and assigns each allele frequency a Beta prior distribution, which has a population genetic justification (Wright, 1931). We evaluate these estimators on simulated genetic data for two populations, and demonstrate an application to real data from the HealthABC cohort study.

3.2 Methods: Two Individuals from the Same Population

3.2.1 Methods: Calculation of Pedigree-Determined Parameter Values, and Simulated Data

In this section we consider pairs of individuals in the same population i and, for convenience, omit the i superscript. The ibd probability for alleles taken randomly from individuals j, j' is written as $\theta_{jj'}$. These values can be determined from the pedigree of j and j': if diploid individual J is ancestral to both j and j', and if there are n individuals in the pedigree path joining j to j' through J, including j and j', then $\theta_{jj'} = \sum(0.5)^n(1 + F_J)$ where F_J is the inbreeding coefficient of J and the sum is over all ancestors J and all paths joining j to j' through J (Wright 1922). If ancestor J is further back in time than the reference time, then it does not contribute to the relatedness of individuals j and j'.

The *sim* package was used by J. Goudet (personal communication) to construct a base population with SNPs in linkage blocks of 20 Morgans. A set of 20 founders was selected for a defined pedigree, and 79,069 SNPs were chosen independently to provide genotypes for 115 descendants in this pedigree. Only data for the non-founders were used in the analyses reported here. There are 17 individuals with inbreeding coefficient of 0.125, and 98 with inbreeding coefficient of zero, so average inbreeding was $F_W = 0.0185$ for the sample of 115 individuals. For the 6,555 pairs of distinct individuals, the coancestries (and numbers of pairs with each value) were 0 (6340), 1/32 (1146), 1/16 (2176), 3/32 (786), 1/8 (1398), 5/32 (228), 3/16 (302), 1/4 (584) and 5/16 (150). The average coancestry $\theta_S = 0.0538$ was greater than the average inbreeding F_W.

We removed monomorphic markers, which left $L = 72,529$ SNPs, and then applied various estimators, which we will describe, to this simulated data set, and obtained approximate 95% confidence intervals with 1,000 bootstrap samples.

3.2.2 Methods: Method of Moments Estimation (MME)

The allele dosages X_{j_l} for reference allele A at locus l allow the sample allelic matching proportion $\tilde{M}_{jj'_l}$ for individuals j and j' to be calculated as

$$\tilde{M}_{jj'_l} = \frac{1}{2}[1 + (X_{j_l} - 1)(X_{j'_l} - 1)]. \tag{3.2}$$

This quantity has the value 1 for AA, AA and BB, BB pairs, where B is the non-reference allele; the value 0.5 for pairs $AA, AB; AB, AB; BB, AB; AB, AA; AB, BB$ and the value 0 for pairs $AA, BB; BB, AA$.

To clarify the parameters being estimated by coancestry estimators, we return to the discussion of within-population and total expectations in Chapter 2. Conditional on the actual allele proportions p_l for SNP l in a population, the probability $P(A,A)_{jj'_l}$ an allele from individual j and one from individual j' are both the reference type A at locus l introduces the within-population coancestry coefficient $\phi_{jj'}$:

$$P(A,A)_{jj'_l} \;=\; p_l^2 + p_l(1-p_l)\phi_{jj'}.$$

Taking expectations over evolutionary replicates of the population brings in the coancestry coefficient $\theta_{jj'}$ for individuals j, j', the probability an allele drawn from one is ibd to an allele from the other,

$$P(A,A)_{jj'_l} \;=\; \pi_l^2 + \pi_l(1-\pi_l)\theta_{jj'}.$$

With this framework, the within-population expectation of the matching statistic $\tilde{M}_{jj'_l}$ is

$$\begin{aligned}\mathcal{E}_W(\tilde{M}_{jj'_l}) &= 1 - 2p_l + P(A,A)_{jj'_l} \\ &= 1 - 2p_l(1-p_l)(1-\phi_{jj'}),\end{aligned}$$

and the total expectation is

$$\mathcal{E}_T(\tilde{M}_{jj'_l}) \;=\; 1 - 2\pi_l(1-\pi_l)(1-\theta_S)(1-\phi_{jj'}).$$

The within-population coancestry coefficient $\phi_{jj'}$ therefore satisfies

$$\phi_{jj'} \;=\; \frac{\theta_{jj'} - \theta_S}{1-\theta_S},$$

and it is this quantity, not $\theta_{jj'}$, that can be estimated with data from a single population. The ibd probabilities $\theta_{jj'}$ and θ_S can be calculated if the pedigrees of all individuals in the sample are known and these provide the pedigree-determined values of the target parameters of this section. We also write $\phi_{jj'}$ as $\beta_{jj'}$ in the same way that we wrote the within-population inbreeding coefficient f_j as β_j in Chapter 2.

3.2.2.1 Methods: Simple MME

If the allele probabilities π_l are known, then a simple method of moments estimator (MME) is

$$\hat{\theta}_{jj'_{\text{idealSim.}}} = 1 - \frac{\sum_l (1 - \tilde{M}_{jj'_l})}{\sum_l 2\pi_l(1 - \pi_l)},$$

where "idealSim." stands for the "ideal simple" estimator. We call the estimator ideal because its expected value $E(\hat{\theta}_{jj'_{\text{idealSim.}}}) = \theta_{jj'}$. To deal with unknown π_l, the substitution of sample allele frequencies $\{\tilde{p}_l\}$ for $\{\pi_l\}$ has been proposed. A formula analogous to $\hat{\theta}_{jj'_{\text{idealSim.}}}$ was given as Equation 2 by Ritland (1996).

When the allele probabilities π_l are known, $\hat{\theta}_{jj'_{\text{idealSim.}}}$ provides an unbiased estimate of the coancestry $\theta_{jj'}$. If the sample allele proportions $\tilde{p}_l$ are used in place of π_l, however, it is the within-population coancestry $\beta_{jj'}$ that can be considered. By analogy to the argument motivating the definition of the within-population inbreeding coefficient in Section 2.2.1, replacing Equation 2.1 with $\mathcal{E}_W(x_{jk}x_{j'k'}) = p^2 + p(1-p)\theta_{jj'}$, we define the within-population coancestry coefficient $\beta_{jj'}$ as $(\theta_{jj'} - \theta_S)/(1 - \theta_S)$. We write

$$\hat{\beta}_{jj'_{\text{Sim.}}} \quad = \quad 1 - \frac{\sum_l (1 - \tilde{M}_{jj'_l})}{\sum_l 2\tilde{p}_l(1 - \tilde{p}_l)},$$

where "Sim." stands for "simple." As the expectation of $2\tilde{p}_l(1 - \tilde{p}_l)$ is not $2\pi_l(1 - \pi_l)$, this estimator does not provide an unbiased estimate of $\theta_{jj'}$, although we show in Appendix 3.8.1 that the ratio of expected values for its numerator and denominator is

$$\frac{\mathcal{E}_T(\text{Numerator})}{\mathcal{E}_T(\text{Denominator})} = \frac{\theta_{jj'} - \theta_S - \frac{1}{2n}(1 + F_W - 2\theta_S)}{1 - \theta_S - \frac{1}{2n}(1 + F_W - 2\theta_S)}. \tag{3.3}$$

We define the simple estimator $\hat{\beta}_{jj'_{\text{Sim.}}}$'s "pedigree-determined parameter" to be the within-population coancestry coefficient $\beta_{jj'}$. The simple estimator $\hat{\beta}_{jj'_{\text{Sim.}}}$ might be regarded as an estimator of its pedigree-determined parameter, and we show some numerical support for this later. We note that when $n \to \infty$, the expression on the right-hand side of Equation 3.3 approaches $\hat{\beta}_{jj'_{\text{Sim.}}}$'s pedigree-determined parameter $\beta_{jj'}$.

We can obtain bootstrap confidence intervals using a procedure entirely analogous to those described in Sections 3.2.2.2 and 3.2.2.3.

3.2.2.2 Methods: Weir and Goudet $\hat{\beta}_{jj',w}$ and $\hat{\beta}_{jj',u}$ Estimators, and Bootstrap Confidence Intervals

The ratio of means (i.e. weighted) Weir and Goudet estimator is $\hat{\beta}_{jj',w}$, and the mean of ratios (i.e. unweighted) Weir and Goudet estimator is $\hat{\beta}_{jj',u}$:

$$\hat{\beta}_{jj',w} := \frac{\sum_l [\tilde{M}_{jj'_l} - \tilde{M}_{S_l}]}{\sum_l [1 - \tilde{M}_{S_l}]} \tag{3.4}$$

$$\hat{\beta}_{jj',u} := \frac{1}{L} \sum_l [\frac{\tilde{M}_{jj'_l} - \tilde{M}_{S_l}}{1 - \tilde{M}_{S_l}}], \tag{3.5}$$

where

$$\tilde{M}_{S_l} := \frac{1}{n(n-1)} \sum_{\substack{j=1 \\ j \neq j'}}^{n} \sum_{j'=1}^{n} \tilde{M}_{jj'_l}$$

and $\tilde{M}_{jj'_l}$ was defined in Equation 3.2. Note that the unweighted MME can suffer from divide by 0 error and thus be undefined when applied to monomorphic SNPs.

Weir and Goudet (2017) introduced the weighted and unweighted $\hat{\beta}_{jj'}$ quantities as estimators of $\beta_{jj'}$. The ratio of expectations of $(\tilde{M}_{jj_l} - \tilde{M}_{S_l})$ and $(1 - \tilde{M}_{S_l})$ is indeed $\beta_{jj'}$, as we show in Appendix 3.8.2. Later, we will refer to the weighted or unweighted $\hat{\beta}_{jj'}$ estimator's "pedigree-determined parameter," and by this, we mean $\beta_{jj'}$. The expectation of the estimator will differ from this ratio $\beta_{jj'}$ in a complex way that depends, among other factors, on the number of SNPs used. Weir and Goudet showed some numerical properties of their estimator, and here we assess quality with mean squared error and plots.

Notice that $\hat{\beta}_{jj',w}$ and $\hat{\beta}_{jj',u}$ both involve the same factors $\hat{\beta}_{Num,jj'_l}$ and $\hat{\beta}_{Denom,l}$, defined as follows:

$$\hat{\beta}_{Num,jj'_l} := \tilde{M}_{jj'_l} - \tilde{M}_{S_l} \text{ and}$$
$$\hat{\beta}_{Denom,l} := 1 - \tilde{M}_{S_l}.$$

$\hat{\beta}_{jj',w}$ can be written:

$$\hat{\beta}_{jj',w} = \sum_{l=1}^{L}(\hat{\beta}_{Num,jj'_l})/\sum_{l=1}^{L}(\hat{\beta}_{Denom,l}).$$

$\hat{\beta}_{jj',u}$ can be written:

$$\hat{\beta}_{jj',u} = (1/L)\sum_{l=1}^{L}(\hat{\beta}_{Num,jj'_l}/\hat{\beta}_{Denom,l}).$$

Assuming dosages at different loci are independent, we can compute quantities $\hat{\beta}_{Num,jj'_l}$ and $\hat{\beta}_{Denom,l}$ only once to save computation in the following bootstrap procedure. For each (j, j'), after computing $\tilde{M}_{jj',l}$s, we can compute the corresponding $\tilde{M}_{S_l}$, which is the mean of those $\tilde{M}_{jj',l}$s, and corresponding $\hat{\beta}_{Num,jj'_l}$s and $\hat{\beta}_{Denom,l}$s. We save the $\hat{\beta}_{Num,jj'_l}$s and $\hat{\beta}_{Denom,l}$s for re-use, saving computation in bootstrap calculations.

To obtain a bootstrap sample, we perform the following procedure, re-sampling loci's $\hat{\beta}_{Num,jj'_l}$s and $\hat{\beta}_{Denom,l}$s to mimic genetic sampling discussed in Chapter 2:

(1) Sample L 2-vectors from the set of 2-vectors

$\{[\hat{\beta}_{Num,jj'_{l=1}}, \hat{\beta}_{Denom,l=1}], [\hat{\beta}_{Num,jj'_{l=2}}, \hat{\beta}_{Denom,l=2}], ..., [\hat{\beta}_{Num,jj'_{l=L-1}}, \hat{\beta}_{Denom,l=L-1}], [\hat{\beta}_{Num,jj'_{l=L}}, \hat{\beta}_{Denom,l=L}]\}$

with replacement, obtaining boostrap sample

$\{[\hat{\beta}^*_{Num,jj'_{l=1}}, \hat{\beta}^*_{Denom,l=1}], [\hat{\beta}^*_{Num,jj'_{l=2}}, \hat{\beta}^*_{Denom,l=2}], ..., [\hat{\beta}^*_{Num,jj'_{l=L-1}}, \hat{\beta}^*_{Denom,l=L-1}], [\hat{\beta}^*_{Num,jj'_{l=L}}, \hat{\beta}^*_{Denom,l=L}]\}$

(2) For each MME (weighted or unweighted), calculate an estimate for each distinct pair (j, j') s.t. $j \neq j'$ of individuals' $\beta_{jj'}$ parameter on the boostrap sample.

For each distinct pair (j, j'), repeat Steps (1)-(2) B = 1,000 times to get 1,000 weighted $\hat{\beta}_{jj',w}$ estimates and 1,000 unweighted $\hat{\beta}_{jj',u}$ estimates. Take (0.025,0.975) quantiles of 1,000 $\hat{\beta}_{jj',w}$ estimates to form a 95% confidence interval, and take (0.025,0.975) quantiles of 1,000 $\hat{\beta}_{jj',u}$ estimates to form a 95% confidence interval. Coverage for a parameter can be calculated as the fraction of confidence intervals that cover the parameter.

When we applied this bootstrap procedure to simulated one-population data, our implementation of $\hat{\beta}_{jj',w}$ and $\hat{\beta}_{jj',u}$ estimates and bootstrap CIs was written with base R code, with $\tilde{M}_{jj'_l}$s calculated using Rcpp for speed-up.

3.2.2.3 Methods: GCTAw and GCTAu Estimators, and Bootstrap Confidence Intervals

Another moment estimator for the kinship coefficient $\theta_{jj'}$ between individuals j and j' is $\hat{\theta}_{jj'_{\mathrm{GCTA}}}$, and can be derived from Yang et al. (2011)'s moment estimator for the coefficient of relationship $2\theta_{jj'}$, which is twice the kinship coefficient. Ritland (1996) previously wrote of the same estimator, but used different notation. Equation 3 of Yang et al. (2011) gives the following moment estimator for the "genetic relationship between individuals j and k:"

$$\text{``}A_{jk} = \frac{1}{N} \sum_{i=1}^{N} \frac{(x_{ij} - 2p_i)(x_{ik} - 2p_i)}{2(p_i)(1 - p_i)}, \text{''} \tag{3.6}$$

where "x_{ij} is the number of copies of the reference allele for the i-th SNP of the j-th individual" and "p_i is the frequency of the reference allele" presumably in the sample, since there is no mention of an "ancestral" population, and since A_{jk}, being an estimator, cannot be defined in terms of an unknown parameter p_i. Note that Yang et al. (2011)'s A_{jk} is not the same as our A^i_{jk} mentioned earlier. In our notation, Yang et al. (2011)'s estimator of the genetic relationship between individuals j and j' would be:

$$A_{jj'} = \frac{1}{N}\sum_{l=1}^{N}\frac{(X_{jl}-2\tilde{p}_l)(X_{j'l}-2\tilde{p}_l)}{2(\tilde{p}_l)(1-\tilde{p}_l)}, \tag{3.7}$$

where l indexes the N loci, X_{jl} is individual j's dosage at locus l, and $\tilde{p}_l$ is the sample reference allele frequency at locus l.

While the words "kinship" and "coancestry" are absent from Yang et al. (2011)'s paper, we infer that their moment estimator $A_{jj'}$ estimates the coefficient of relationship, which is twice the kinship coefficient (i.e. $2\theta_{jj'}$), since if one were to treat sample frequencies $\tilde{p}_l$s as known ancestral population frequencies π_ls, one would obtain that $E(A_{jj'}) = 2\theta_{jj'}$. Hence, $E(A_{jj'}/2) = E\Big(\sum_{l=1}^{L}\Big[(X_{jl}-2\pi_l)(X_{j'l}-2\pi_l)\Big]/\Big[4(\pi_l)(1-\pi_l)L\Big]\Big) = \theta_{jj'}$. Hence, the GCTA estimator $A_{jj'}/2$ for $\theta_{jj'}$ involves a factor of 4 (not 2 as given in the GCTA paper) in its denominator, which was recognized by Weir and Goudet (2017) in Equation 7.

We define the GCTAw estimator of kinship coefficient $\theta_{jj'}$ between individuals j and j' as:

$$\hat{\theta}_{jj'_{\text{GCTAw}}} = \frac{\sum_l (X_{j_l}-2\tilde{p}_l)(X_{j'_l}-2\tilde{p}_l)}{\sum_l 4\tilde{p}_l(1-\tilde{p}_l)}. \tag{3.8}$$

Equation 3.8 gives the "ratio of means" (i.e. weighted) method for combining over loci, whereas Yang et al. (2011) used the "mean of ratios" (i.e. unweighted) method:

$$\hat{\theta}_{jj'_{\text{GCTAu}}} = \frac{1}{L}\sum_{l=1}^{L}\frac{(X_{j_l}-2\tilde{p}_l)(X_{j'_l}-2\tilde{p}_l)}{4\tilde{p}_l(1-\tilde{p}_l)}. \tag{3.9}$$

Note that the unweighted MME can suffer from divide by 0 error and thus be undefined when applied to monomorphic SNPs. In Appendix 3.8.3, we show that the ratio of the expected value of the numerator of $\hat{\theta}_{jj'_{\text{GCTAw}}}$ to the expected value of its denominator is the same as the mean over loci of the expectation of a single locus $\hat{\theta}_{jj'_{\text{GCTAu}}}$'s numerator to the expectation of a single locus $\hat{\theta}_{jj'_{\text{GCTAu}}}$'s denominator. In Equations 3.18 and 3.20, we show that both quantities are equal to $\Big[2(\theta_{jj'} + \big[\theta_S + \frac{1}{2n}(1 + F_W - 2\theta_S)\big]) - 2\psi_{j'} - \frac{1}{n}(1 +$

$F_{j'} - 2\psi_{j'}) - 2\psi_j - \frac{1}{n}(1 + F_j - 2\psi_j)\Big] / \Big[2\Big((1-\theta_S) - \frac{1}{2n}\left(1 + F_W - 2\theta_S\right)\Big)\Big]$. We define ψ_j as $\sum_{j'=1, j'\neq j}^{n} \theta_{jj'}/(n-1)$, the average coancestry of individual j with every other member of the sample of n individuals from the population to which j belongs, and similarly for $\psi_{j'}$. We define weighted and unweighted GCTA MMEs' pedigree-determined parameter to be $[2(\theta_{jj'} + \theta_S) - 2\psi_{j'} - 2\psi_j] / [2(1-\theta_S)]$, which equals the expression of Equations 3.18 and 3.20, as the number of individuals $n \to \infty$. Unlike the simple MME and Weir and Goudet weighted and unweighted MME estimators, GCTA estimators' pedigree-determined parameter is not a monotone-increasing function of $\theta_{jj'}$ for fixed θ_S because the pedigree-determined parameter is a function of $\psi_{j'}$ and ψ_j.

To obtain bootstrap confidence intervals in a computationally efficient way, we noted that estimators (Equations 3.8 and 3.9) could be written in terms of constituent components $\hat{a}_{jj',l} := (X_{j_l} - 2\tilde{p}_l)(X_{j'_l} - 2\tilde{p}_l)$ and $\hat{b}_l := 4(\tilde{p}_l)(1-\tilde{p}_l)$, as

$$\hat{\theta}_{jj'_{\text{GCTAw}}} = \frac{\sum_l \hat{a}_{jj',l}}{\sum_l \hat{b}_l}$$
$$\hat{\theta}_{jj'_{\text{GCTAu}}} = \frac{1}{L}\sum_{l=1}^{L} \frac{\hat{a}_{jj',l}}{\hat{b}_l}.$$

Instead of sampling with replacement L vectors of dosages from the set of dosages $\{\vec{X}_{.1}, ..., \vec{X}_{.L}\}$ in the simulated data, where each $\vec{X}_{.l} = [X_{1l}, ..., X_{jl}, ..., X_{nl}]$, calculating $\hat{\theta}_{jj',GCTAw}$ and $\hat{\theta}_{jj',GCTAu}$ on the bootstrap sample, and repeating that process for $B = 1,000$ bootstrap samples, we save computation with the following procedure that computes $\hat{a}_{jj',l}$s and $\hat{b}_l$s once and re-uses calculated values. For each pair of individuals (j, j') and locus l, we compute $\hat{a}_{jj',l}$s and $\hat{b}_l$s on the simulated data. We then sample L loci from $\{1, ..., L\}$ with replacement, select their corresponding $\hat{a}_{jj',l}$s and $\hat{b}_l$s, and calculate a GCTAw estimate and a GCTAu estimate on the bootstrap sample. We repeat the process of sampling and calculating GCTAw and GCTAu estimates $B = 1,000$ times, obtaining $B = 1,000$ bootstrap GCTAw estimates and $B = 1,000$ bootstrap GCTAu estimates for each $\theta_{jj'}$. We construct a 95% confidence interval for our GCTAw estimate computed on the original data set by taking the 2.5% and

97.5% quantiles of the $B = 1,000$ calculated GCTAw estimates, and similarly for GCTAu.

When we applied this bootstrap procedure to simulated one-population data, our implementation of GCTAw and GCTAu estimates and bootstrap CIs was written with base R code, with $\hat{a}_{jj',l}$s calculated using Rcpp for speed-up.

3.2.3 Methods: Estimator from Variances and Covariances over Loci

A quite different method of moments estimator is based on the covariance over loci of allele dosages for two individuals. Individuals j, j' have dosages $X_{j_l}, X_{j'_l}$ at locus l. Each dosage has expected value $2\pi_l$ and the covariance of dosages is

$$\mathrm{Cov}(X_{j_l}, X_{j'_l}) \quad = \quad 4\pi_l(1-\pi_l)\theta_{jj'}.$$

The sample covariance over L loci is

$$s_{jj'} \quad = \quad \frac{1}{L-1}\sum_{l=1}^{L}(X_{j_l} - \bar{X}_j)(X_{j'_l} - \bar{X}_{j'}),$$

where $\bar{X}_j = \sum_{l=1}^{L} X_{j_l}/L$. Under the assumption of different loci's dosages being independent:

$$\mathcal{E}(s_{jj'}) = 2\theta_{jj'}H_e + 4\sigma_L^2, \tag{3.10}$$

where

$$\begin{aligned} H_e &= \frac{1}{L}\sum_l 2\pi_l(1-\pi_l) \\ \sigma_L^2 &= \frac{1}{L-1}\sum_{l=1}^{L}(\pi_l - \bar{\pi})^2, \;\; \bar{\pi} = \frac{1}{L}\sum_{l=1}^{L}\pi_L. \end{aligned}$$

A possible advantage of this estimator is that its mean is not affected by inbreeding of target individuals j and j', or by inbreeding coefficients or relatedness (i.e. coancestry coefficients) of any other pair of individuals in the sample. Information from other individuals is not used.

3.2.4 Methods: Maximum Likelihood Estimation

One of the first maximum-likelihood approaches to relatedness estimation was by Thompson (1975). She restricted attention to non-inbred relatives and gave details for estimating the probabilities k_i of two individuals sharing $i = 0, 1, 2$ pairs of alleles ibd. The work of Yengo et al. (2017) has shown that it is not realistic to assume lack of inbreeding.

Milligan (2003) removed this restriction to non-inbred relatives and so had to work with all nine “Jacquard coefficients” Δ_i for two individuals sharing 2, 3, or 4 2-pairs of alleles ibd. Maximizing the likelihood for eight parameters (the nine Δs sum to one) is not computationally feasible for large numbers of SNPs. There is the additional problem of non-identifiability with SNP data pointed out by Csűrös (2014): different sets of the $\{\Delta_i\}$ can give the same set of genotype probabilities for two individuals. Csűrös (2014) showed that the inbreeding coefficients for each individual and the coancestry coefficient for both individuals were estimable, but he restricted attention to the simple estimator discussed here and in Chapter 2. Anderson and Weir (2007) considered maximum likelihood estimation of coancestry coefficients in the presence of background evolutionary relatedness.

The papers of Thompson (1975), Milligan (2003), and Anderson and Weir (2007) all substituted sample allele proportions for unknown allele probabilities. Based on our inbreeding work in Chapter 2, we expect that MLEs will not be satisfactory, because this substitution results in estimates that are affected by kinship coefficients between all pairs of individuals used to obtain sample allele proportions. Milligan said “In practice, realistic samples will often involve enough individuals that errors in the allele-frequency distribution will be quite small.” However, we have shown (e.g. in pedigree-determined parameter expressions for estimators $\hat{\beta}_{jj'_{\text{Sim.}}}$, $\hat{\beta}_{jj',w}$, $\hat{\theta}_{jj'_{GCTAw}}$, and $\hat{\beta}_{jj'_{GCTAu}}$) that sample size (the number of individuals n) has little effect on the bias introduced by relatedness among individuals (i.e. coancestry coefficients between pairs of individuals) providing data for the sample allele proportions.

Work with large-scale association studies, e.g. (Conomos et al., 2016), shows that is not realistic to assume the general absence of relatedness between pairs of n individuals other than the target pair j and j' for which one estimates $\theta_{jj'}$, so that the use of sample allele proportions may be misleading.

Other authors, e.g. (Balding, 2003), assumed allele probabilities are known: in practice estimated from data outside the present study. We prefer not to make that assumption.

In summary, for the above reasons, namely unrealistic assumptions (i.e. related individuals who are not inbred, absence of relatedness amongst pairs of n sample individuals other than the target pair j and j', allele probabilities estimated from a second data set), computational infeasibility, non-identifiability, and the introduction of bias through the use of sample allele frequencies in place of unknown allele probabilities, we do not pursue likelihood-based inference for relatedness within one population.

3.3 Methods: Two Alleles from Different Populations

We now consider pairs of alleles drawn from individuals in different populations. For convenience we will restrict attention to the case where allelic, rather than genotypic, data are considered. This will remove the need to consider inbreeding. Identity by descent for pairs of alleles, within or between individuals, is assumed present in each population. However, we depart from most previous analyses in allowing for relatedness between individuals in different populations.

Recall from Section 3.1 the definition of ibd probability θ_B. We explicitly recognize that θ_B is not zero, and Weir and Goudet (2017) showed many scenarios involving genetic drift, mutation and migration that led to non-zero values. The parameters of interest are θ_W^i, the average ibd probability for pairs of distinct alleles in population i, averaging over all such pairs, $\theta_W^{ii'}$, the average ibd probability for pairs of alleles, one from population i and one from a different population i', averaging over all such pairs, and population-specific F_{ST}, also called β^{ii} and β^i, defined as $(\theta_W^i - \theta_B)/(1-\theta_B)$ (Weir and Goudet, 2017). Intuitively, the parameter β^i measures the genetic relatedness of population i relative to the genetic relatedness of a set of populations to which population i belongs. When there is a non-zero difference $\theta_W^i - \theta_B$ of within- versus between-population coancestry, we say there is population structure.

Actual allele proportions $\{p_{il}\}$ for the reference allele at locus l in population i vary over populations. They are random due to genetic sampling discussed in Chapter 2, and we assign them the same expected values $\{\pi_l\}$. Considering a sample of alleles from each population, reference allele frequencies in the samples are $\{\tilde{p}_{il}\}$ for locus l in population i.

3.3.1 Methods: Method of Moments Estimation

3.3.1.1 Methods: Beta Estimator

Definitions in this section are the same as those given by Buckleton et al. (2016), except that we use ii instead of i, ii' instead of ij, and consider bi-allelic loci instead of multi-allelic loci. Consider a single locus l with two alleles, the reference allele and the alternative allele. In a reference e.g. ancestral population, these alleles have frequencies $\{\pi_l\}$. A set of populations $i \in \{1, 2, \ldots r\}$ have descended from this ancestral population, and the current allele frequencies in population i at locus l are $\{p_{il}\}$. We sample n_{il} alleles from population i at locus l. Of these sampled alleles, n_{ilu} are the reference allele u, and the sample reference allele frequencies are $\tilde{p}_{il} = n_{ilu}/n_{il}$.

The proportions of pairs of alleles that match within a sample from population i or between samples from populations $i, i' : i \neq i'$ are

$$
\begin{aligned}
\tilde{M}_{iil} &= \frac{(n_{ilu})(n_{ilu}-1)}{(n_{il})(n_{il}-1)} + \frac{(n_{il}-n_{ilu})((n_{il}-n_{ilu})-1)}{(n_{il})(n_{il}-1)} \\
&\approx (\tilde{p}_{ilu})^2 + (1-\tilde{p}_{ilu})^2 \\
\tilde{M}_{ii'l} &= \frac{n_{ilu}n_{i'lu}}{n_{il}n_{i'l}} + \frac{(n_{il}-n_{ilu})(n_{i'l}-n_{i'lu})}{n_{il}n_{i'l}} \\
&= \tilde{p}_{ilu}\tilde{p}_{i'lu} + (1-\tilde{p}_{ilu})(1-\tilde{p}_{i'lu}),
\end{aligned}
$$

where the approximations hold for large samples n_{il} and $n_{i'l}$ of alleles.

The β estimators are

$$\hat{\beta}_l^{ii} = \frac{\tilde{M}_{iil} - \tilde{M}_{Bl}}{1 - \tilde{M}_{Bl}} \quad \text{and} \quad \hat{\beta}_l^{ii'} = \frac{\tilde{M}_{ii'l} - \tilde{M}_{Bl}}{1 - \tilde{M}_{Bl}},$$

where we define $\tilde{M}_{Bl} := \sum_{i=1}^{r_l} \sum_{i'=1,i'\neq i}^{r_l} \tilde{M}_{ii'l}/[r_l(r_l - 1)]$, r_l being the number of populations that had locus l typed.

If the same values of θ hold for a series of loci l, then the ratio of means (i.e. weighted) multi-locus estimators are

$$\hat{\beta}^{ii} = \frac{\sum_l (\tilde{M}_{iil} - \tilde{M}_{Bl})}{\sum_l (1 - \tilde{M}_{Bl})} \text{ and} \tag{3.11}$$

$$\hat{\beta}^{ii'} = \frac{\sum_l (\tilde{M}_{ii'l} - \tilde{M}_{Bl})}{\sum_l (1 - \tilde{M}_{Bl})}, \tag{3.12}$$

and $\hat{\beta}_{ii}$ is an estimator for population-i-specific F_{ST} β^{ii}. Importantly, we note $\theta_B = 0$ is not assumed in the construction of these estimators. Note that the sum of $\hat{\beta}^{ii'}$ over all pairs (i, i') results in 0 by construction.

3.3.2 Methods: Bayesian Estimation

3.3.2.1 Methods: Existing Bayesian Methods for Estimating Population-Specific F_{ST}

Others have developed Bayesian approaches for estimating population-specific F_{ST} (Nicholson et al., 2002; Foll and Gaggiotti, 2006; Gunther and Coop, 2013). Figure 3.1 depicts the Bayesian hierarchical model of Nicholson et al. (2002). In this model, each population's allele frequency is modeled as following a truncated normal distribution, and populations have independent allele frequencies, implying $\theta_B = 0$. Figure 3.2 depicts the Bayesian hierarchical model of Gunther and Coop (2013), implemented in the software Bayenv2.0. The Bayenv model is an extension of the model of Nicholson et al. (2002), allowing $\theta_B > 0$ through an Ω matrix of $\theta^{ii'}$s, where the element in the i-th row and i'-th column of Ω is $\theta^{ii'}$. Figure 3.3 depicts the Bayesian hierarchical model of Foll and Gaggiotti (2006), in which populations'

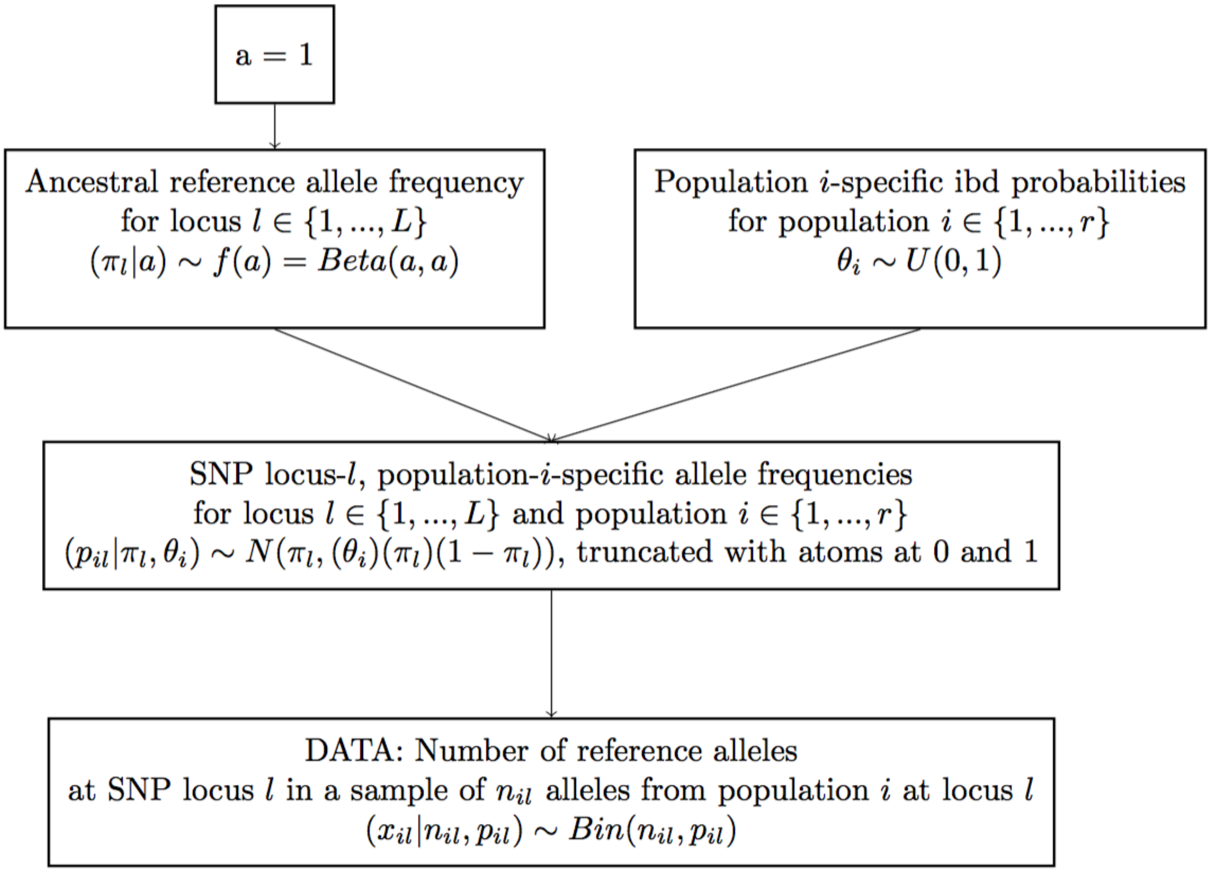

Figure 3.1: Bayesian hierarchical model of Nicholson et al. (2002).

allele frequencies are independent of each other, again implying $\theta_B = 0$. The model of Foll and Gaggiotti (2006) is implemented in the software BayeScan (Foll and Gaggiotti, 2008; Foll et al., 2010).

The BayeScan Bayesian hierarchical model models population-i and locus-l-specific reference allele frequencies $(p_{il}|\pi_l, \theta^i)$ as being distributed $Beta((1 - \theta^i)(p_{il})/(\theta^i), (1 - \theta^i)(1 - p_{il})/(\theta^i))$. Reasons for using a Beta distribution in the model are that Sewall Wright showed that under various population genetic models the allele frequency distribution is Beta (Wright, 1931), and the Beta distribution is conjugate to the binomial distribution for sample allele counts, which typically makes obtaining samples from the posterior distribution

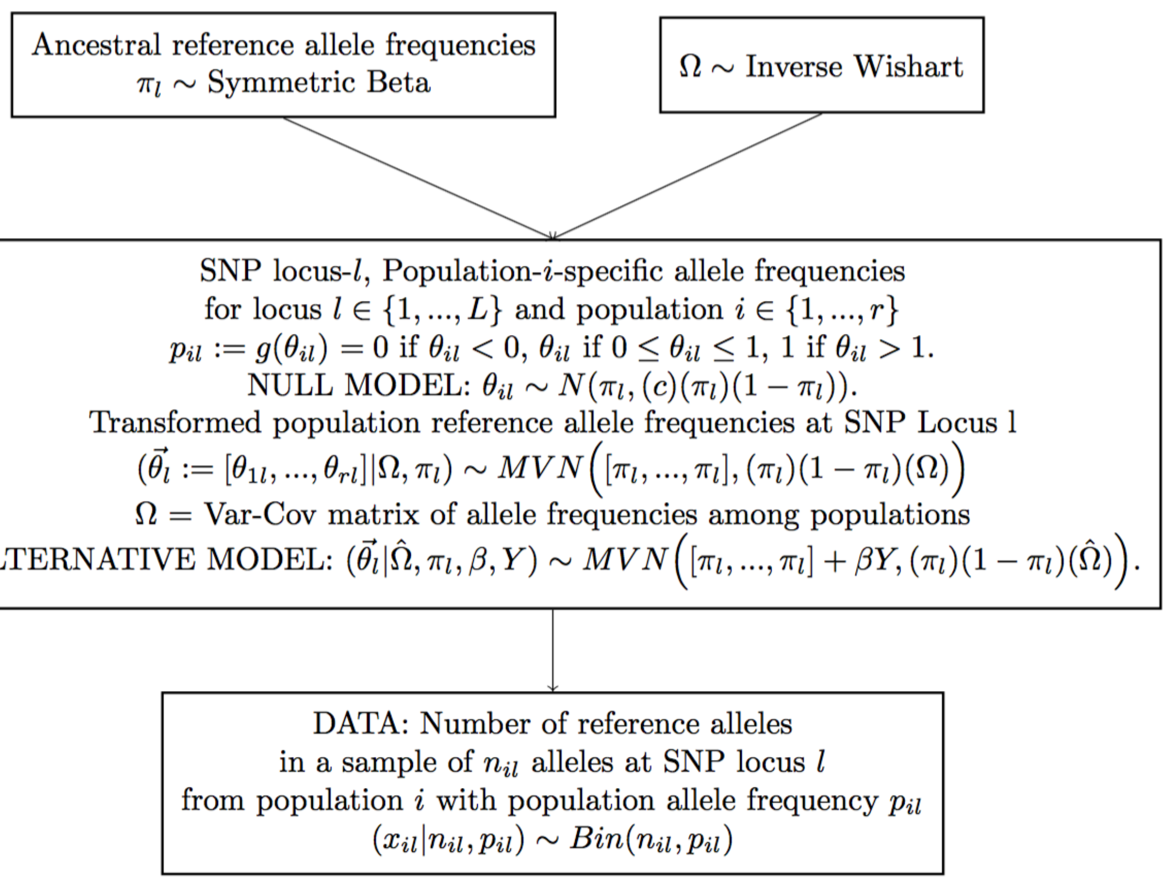

Figure 3.2: Bayesian hierarchical model of Gunther and Coop (2013).

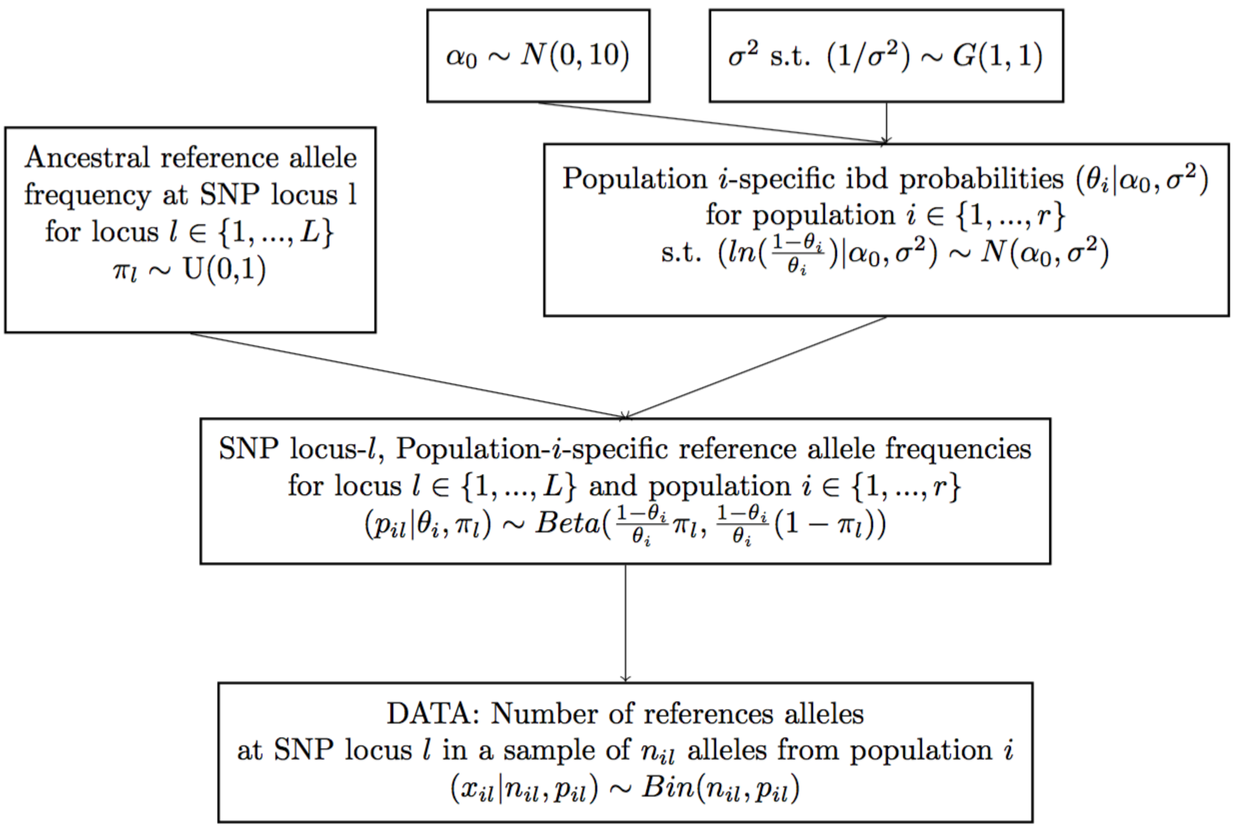

Figure 3.3: Bayesian hierarchical model of Foll and Gaggiotti (2006).

more tractable.

The assumption of populations' allele frequencies being independent implies that $\theta_B = 0$ and negative estimates of β^i are not possible. The model of Nicholson et al. (2002) and the model of Foll and Gaggiotti (2006) are not appropriate for general situations where migration among the studied populations is a possibility, so that $\theta_B > 0$ and β^i can be negative.

3.3.2.2 Methods: A New Bayesian Estimator of Population-Specific F_{ST}

We extend the method of Foll and Gaggiotti (2006) by allowing for dependent population allele frequencies, while preserving the well-justified Beta marginal distribution for each allele frequency. Figure 3.4 depicts our proposed novel hierarchical model for two populations that allows θ^{12}, which is θ_B for two populations, to be non-zero. It employs a Gaussian copula and inverse cumulative distribution function (cdf) F_{il}^{-1} of a $Beta((1-\theta^i)(\pi_l)/(\theta^i),(1-\theta^i)(1-\pi_l)/(\theta^i))$ distribution to achieve $\mathcal{E}(p_{il}) = \pi_l$, $\mathrm{Var}(p_{il}) = (\pi_l)(1-\pi_l)(\theta^i)$, covariance $\mathrm{Cov}(p_{1l}, p_{2l}) \approx (\pi_l)(1-\pi_l)(\theta^{12})$, and Beta marginal distributions for p_{1l} and p_{2l}, consistent with the justification of Wright (1931) for a Beta allele frequency distribution. Pitt et al. (2006) described using the Gaussian copula and inverse cdf to obtain a multivariate distribution with specified marginal distributions.

This model only applies to two populations. Extending this model to more than two populations, say a third, by adding a Beta-distributed $Corr(p_{1l}, p_{3l})$ for the correlation between the first and the third populations' allele frequencies would not work, because such a model would not account for the fact that between-population correlations are dependent: Knowing $Corr(p_{1l}, p_{2l})$ and $Corr(p_{2l}, p_{3l})$ bounds $Corr(p_{1l}, p_{3l})$. For example, $\mathrm{Corr(X,Y)} = a > 0$, $\mathrm{Corr(Y, Z)} = b > 0 \implies \mathrm{Corr(X,Z)} \geq ab - \sqrt{(1-a^2)(1-b^2)}$. This bound on $Corr(p_{1l}, p_{3l})$ would be inconsistent with a Beta prior distribution, which allows $Corr(p_{1l}, p_{3l}) \in [0, 1]$.

We previously investigated modeling $([p_{1l}, p_{2l}] \mid \theta^1, \theta^2, \theta^{12}, \pi_l) \sim$ Lee Sarmanov Bivariate Beta $(a_1 = (1-\theta^1)(\pi_l)/(\theta^1),\ b_1 = (1-\theta^1)(1-\pi_l)/(\theta^1),\ a_2 = (1-\theta^2)(\pi_l)/(\theta^2),\ b_2 = (1-\theta^2)(1-\pi_l)/(\theta^2),\ \omega = (\theta^{12})/(\theta^1\theta^2(\pi_l)(1-\pi_l)))$ (Ting Lee, 1996), but found that it

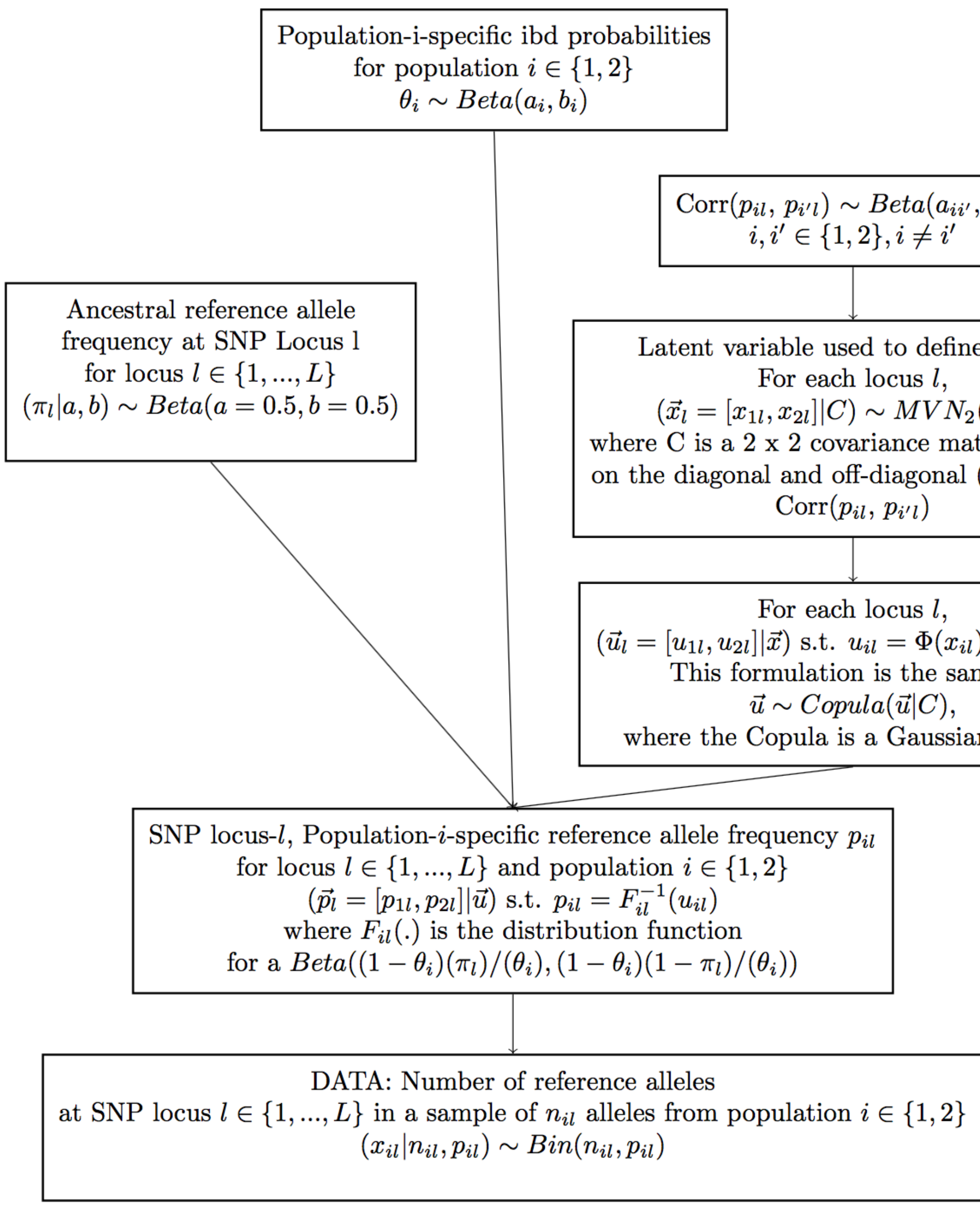

Figure 3.4: Our Two-Population Bayesian Hierarchical Copula Model. We employ a copula to model populations' dependent allele frequencies.

imposed a constraint $\theta^{12} < \sqrt{\theta^1\theta^2}$ that does not hold for a two-island population genetic model, so we did not pursue this modeling approach.

We also investigated modeling

$$([logit(p_{1l}), logit(p_{2l})]|\theta^1, \theta^2, \theta^{12}, \pi_l) \sim N\left([logit(\pi_l), logit(\pi_l)], \left[\begin{smallmatrix} \theta^1 & \theta^{12} \\ \theta^{12} & \theta^2 \end{smallmatrix}\right] \tfrac{1}{(\pi_l)^2(1-\pi_l)^2}\right),$$

as well as modeling $([p_{1l}, p_{2l}]|\theta^1, \theta^2, \theta^{12}, \pi_l) \sim$ bivariate truncated normal. However, we decided to pursue the copula model instead of these other modeling approaches, because we found, by obtaining Monte Carlo draws from each model, that the copula model allowed us to generate allele frequencies with sample mean closest to $\mathcal{E}(p_{il}) = \pi_l$, sample variance closest to $\text{Var}(p_{il}) = (\pi_l)(1-\pi_l)(\theta^i)$, and sample covariance closest to $\text{Cov}(p_{1l}, p_{2l}) = (\pi_l)(1-\pi_l)(\theta^{12})$. We also realized that with the bivariate truncated normal distribution, each population i's allele frequency p_{il} at locus l has much mass near the middle of [0,1], so cannot be U-shaped, or in other words, bimodal with modes near 0 and 1.

3.3.3 Methods: Simulation of Two-Population Data According to the Two-Island Model

Here, we describe the simulation of data from two populations. Note that the simulation procedure described here differs from the simulation procedure of Section 3.2.1, which was used to generate data for individuals from the same population.

We simulated data from a two-island model depicted in Figure 3.5. Two diploid populations start randomly mating at some generation time 0 for t generations. At each subsequent generation time, individuals can migrate between the two populations with rates $m_i, i \in \{1, 2\}$ where m_i is the probability an individual in population i migrated from the other population, and each allele at a locus can mutate to a new allele with probability μ. Population $i \in \{1, 2\}$ has diploid population size N_i at every generation. From population 1 at generation time $t = 1,000$, we sample $n = 100$ individuals, and similarly for population 2. Our data consists of counts of reference alleles at $L = 100$ unlinked bi-allelic loci for n individuals in each sample.

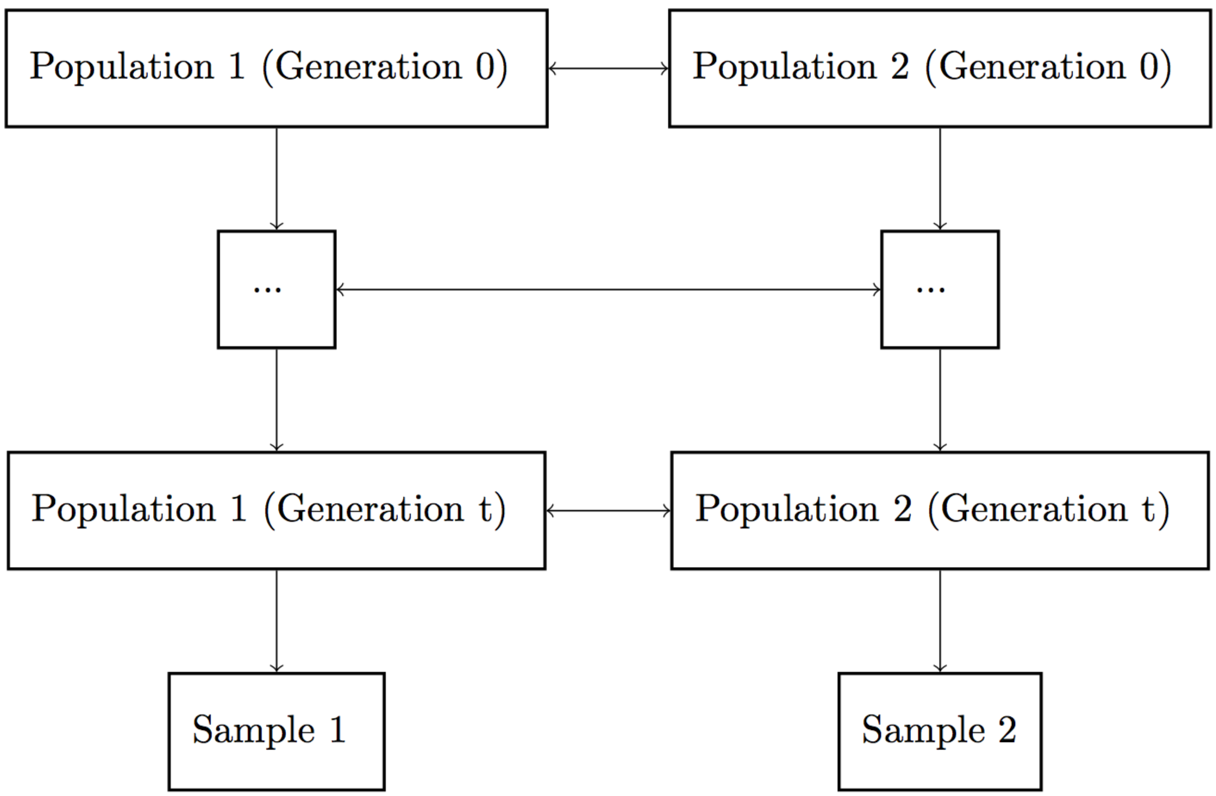

Figure 3.5: Two-Island Model. Population genetic parameters: Mutation rate μ, migration rates m_1 and m_2, and number of diploids N_i, $i \in \{1, 2\}$ per population. Boxes with "..." indicate generations 1 to $(t-1)$ of 2 populations.

For the two-island model, the following transition equations hold:

$$\begin{aligned}\theta^{ii}(t+1) &= (1-\mu)^2\left[(1-m_i)^2\theta^{ii*}(t)+2m_i(1-m_i)\theta^{12}(t)+m_i^2\theta^{i'i'*}(t)\right], \quad i=1,2; i'\neq i \\ \theta^{12}(t+1) &= (1-\mu)^2\left[(1-m_1)m_2\theta^{11*}(t)+[(1-m_1)(1-m_2)+m_1m_2]\theta^{12}(t)+m_1(1-m_2)\theta^{22*}(t)\right],\end{aligned} \tag{3.13}$$

where $\theta^{ii}(t+1), i \in \{1,2\}$ is the population-i ibd probability at generation time $t+1$, $\theta^{12}(t+1)$ is the between-population 1 and 2 ibd probability $\theta^{12}(t+1)$ at generation time $t+1$, $\theta^{ii*}(t) = 1/(2N_i) + (2N_i - 1)\theta^{ii}(t)/(2N_i)$, the mutation rate is μ, and population $i : i = 1, 2$ receives a fraction m_i of its alleles each generation from population $i' : i' \neq i$. These simultaneous recursive transition equations are similar to those of Maruyama (1970), but different, in that Maruyama considered a general, finite number of populations, while we consider two populations, and Maruyama (1970) does not have a between-population ibd probability, while we have θ^{12}. We treat the true population-i-specific F_{ST} as $\beta^i(t) = (\theta^i(t) - \theta^{12}(t))/(1-\theta^{12}(t)), i \in \{1,2\}$. As t grows, $\theta^{ii}(t)$ and $\theta^{12}(t)$ approach constant values, i.e. equilibrium values.

A consequence of these equations is that $\theta^{11}(t) + \theta^{22}(t) \geq 2\theta^{12}(t)$, or that $\theta_W := (\theta^{11} + \theta^{22})/2 \geq \theta_B := \theta^{12}$, so that standard F_{ST} $\beta_W := (\theta_W - \theta_B)/(1-\theta_B)$ is nonnegative. However, it is not necessary that either θ^{11} or θ^{22} exceed θ^{12}. For example, if $N_1 = 100, N_2 = 1,000$ and $\mu = m_1 = m_2 = 0.01$, then the equilibrium θ values are $\theta^{11} = 0.1264, \theta^{12} = 0.0403$, and $\theta^{22} = 0.0320$ so that $\beta^{11} = 0.0897$ and $\beta^{22} = -0.0087$. The negative population-2-specific F_{ST} β^{22} value shows that unlike β_W, population-specific F_{ST} can be negative. The correlation between the two populations' allele frequencies is 0.63.

We simulated six parameter settings depicted in Table 3.1, with each setting defined by a combination of 5 population genetic parameter values (μ, m_1, m_2, N_1, N_2). For each simulation setting, we simulated 100 data sets with $L = 100$ loci per data set. Simulation settings Ind A, Ind B, and Ind C have $\theta^{12} = 0$ ("Ind" for "Independence"), while simulation settings Dep A, Dep B, and Dep C have $\theta^{12} \neq 0$ ("Dep" for "Dependence").

Setting	N_1	N_2	μ	m_1	m_2	θ^1	θ^2	θ^{12}	$F_{ST,1}$	$F_{ST,2}$	$F_{ST,W}$
Dep A	100	1000	0.01	0.01	0.01	0.13	0.03	0.04	0.09	-0.01	0.04
Dep B	1000	1000	0.01	0.01	0.01	0.02	0.02	0.01	0.01	0.01	0.01
Dep C	100	1000	0.01	0.00	0.01	0.20	0.04	0.07	0.14	-0.02	0.06
Ind A	100	1000	0.01	0.00	0.00	0.20	0.02	0.00	0.20	0.02	0.11
Ind B	1000	100	0.01	0.00	0.00	0.02	0.20	0.00	0.02	0.20	0.11
Ind C	1000	1000	0.01	0.00	0.00	0.02	0.02	0.00	0.02	0.02	0.02

Table 3.1: Six Simulation Settings Defined by a Combination of Population Genetic Parameter Values (Mutation rate μ, migration rates m_1 and m_2, number of diploids N_i, $i \in \{1, 2\}$). From population genetic parameter values, two-island model transition equations yielded θ^1, θ^2, and θ^{12} values for generation time $t = 1,000$. From definitions, θ^1, θ^2, and θ^{12} values yielded $\beta^1 := (\theta^1 - \theta^{12})/(1 - \theta^{12})$ (also known as population-1-specific F_{ST} $F_{ST,1}$), $\beta^2 := (\theta^2 - \theta^{12})/(1 - \theta^{12})$ (also known as population-2-specific F_{ST} $F_{ST,2}$), and $\beta_W := ((\theta^1 + \theta^2)/2 - \theta^{12})/(1 - \theta^{12})$ (also known as standard F_{ST} $F_{ST,W}$) values.

Using J. Goudet's hierfstat R package, downloaded from GitHub at https://github.com/jgx65/hierfstat, we simulated data with the function sim.genot.metapop.t, specifying arguments size (the number of sampled individuals per population), mut (the mutation rate of each locus), nbpop (the number of populations to simulate - two in our case), N (the effective population size of each population), nbloc (the number of loci to simulate), nbal (the number of alleles per locus - two in our case), mig (a matrix of migration rates), and t (the number of generations since the two populations were created) (Goudet, 2005).

3.3.4 Methods: Simulation of Two-Population ibd Probabilities According to the Two-Island Model

Rather than use the data simulated in Section 3.3.3 to both inform our Adapted BayeScan copula model priors and assess our copula model ("use the data twice"), we used ibd proportions simulated as described in this Section to inform our model's priors when assessing our model on data simulated according to Section 3.3.3, thereby avoiding giving our model an unrealistic advantage in simulation study results.

SNP identifiers (described later in this Section) were simulated for two islands. Random mating within populations was assumed. Random migration from each island to the other and infinite-alleles mutation were allowed. Infinite-alleles mutation means that any locus in any individual could mutate at each generation, and any mutation led to a different allelic type. Generations were discrete and non-overlapping.

3.3.4.1 Methods: Parameters and Initial Conditions

The numbers of individuals were fixed at N_1 and N_2 diploid individuals in each generation. Constant migration rates were set at m_1 from island 2 to island 1, and m_2 from island 1 to island 2. A constant mutation rate of μ for all allele types was set. All $2N_1 + 2N_2$ alleles in the initial populations were given unique identifiers. None of these alleles was identical by descent to another allele in the initial populations. The three forces of drift, migration and mutation were all so weak that they could be treated additively, and in any order. For convenience, the order was drift-migration-mutation.

3.3.4.2 Methods: Drift

The $2N_i$ alleles in island $i, i = 1, 2$ at generation $(t+1)$ were drawn independently. Each had a probability $1/(2N_i)$ of being a copy of one of the $2N_i$ alleles in island i at generation t.

3.3.4.3 Methods: Migration

After the drift selection process, each allele in island i at generation $(t + 1)$ was re-assigned to an allele from island $(3-i)$ with probability m_i. This migrant allele had equal probability $1/(2N_{3-i})$ of being any one of the $2N_{3-i}$ alleles in island $(3 - i)$ at generation t.

3.3.4.4 Methods: Mutation

After the drift and migration process, each allele in generation $(t+1)$ was allowed to mutate to a previously non-existing allelic type, with probability μ. Identity by descent was preserved only if there was no mutation.

3.3.4.5 Methods: Identity-by-descent Probabilities

At each generation, the ibd proportions within and between islands were the proportions of pairs of distinct alleles (i.e. allele copies), within or between islands, that had the same identifier. The sample means, variances, and covariances of these realized ibd proportions was calculated across replicates of the simulation for $t = 1,000$ generations.

3.3.5 Methods: Evaluation of Methods on Simulated Two-Population Data

Here, we describe the evaluation of methods on data simulated according to Section 3.3.3. For each parameter $\theta \in \{\theta^1, \theta^2, \theta^{12}, \beta^1, \text{and } \beta^2\}$, we found a method's estimate $\hat{\theta}_d$ of θ on the d-th data set, mean estimate $= \sum_{d=1}^{100}(\hat{\theta}_d)/100$, Bias$^2 = \left[\sum_{d=1}^{100}(\hat{\theta}_d)/100 - \theta\right]^2$, Variance $= \sum_{d=1}^{100}(\hat{\theta}_d - \sum_{d=1}^{100}\hat{\theta}_d/100)^2/100$, and Mean squared error (MSE) $= \sum_{d=1}^{100}(\hat{\theta}_d - \theta)^2/100$.

We also performed a sensitivity analysis of our Adapted BayeScan copula model priors, to see if which prior we used in a class of reasonable priors would change simulated data results much. Informative priors were obtained by simulating θ^1, θ^2, and $\theta^{12}/\sqrt{\theta^1\theta^2}$ values according to Section 3.3.4 to inform priors $\theta^i \sim Beta(a_i, b_i), i \in \{1, 2\}$ and $Corr(p_{1l}, p_{2l}) \sim Beta(a_{12}, b_{12})$. More specifically, we calculated a_i and b_i such that the mean and variance of a $Beta(a_i, b_i)$ distribution matched the sample mean and sample variance of simulated θ^i values, and we calculated a_{12} and b_{12} so that the mean and variance of a $Beta(a_{12}, b_{12})$ distribution matched the sample mean and sample variance of simulated $\theta^{12}/\sqrt{\theta^1\theta^2}$ values. Uniform priors involved modeling $\theta^1 \sim U(0, 1)$, $\theta^2 \sim U(0, 1)$, and $Corr(p_{l1}, p_{l2}) \sim U(0, 1)$. For a dependence simulation setting, we put a prior on $Corr(p_{l1}, p_{l2})$. For an independence simulation setting, we set $Corr(p_{l1}, p_{l2})$ to 0.

For every simulated data set, we applied methods as follows. To implement the MME, we used hierfstat R package's betas() function. To obtain MCMC draws from the posterior of Foll et al.'s BayeScan model, we used the BayeScan software. To obtain MCMC draws for our model with informative priors and for our model with U(0,1) priors, we used Stan version 2.14. To obtain MCMC draws from the Bayenv model, we used the Bayenv2.0 software. Note that whenever we used the Bayenv2.0 software, we did not fit the null model or alternative model described in the Bayenv paper (Gunther and Coop, 2013), but rather only used the software to estimate the Ω matrix of $\theta^{ii'}$ parameters. For every Bayesian method on a simulated data set, we checked for convergence of MCMC chains by checking for $\hat{R} < 1.1$ for every ouputted parameter (Brooks and Gelman, 1998).

3.3.6 Methods: Demonstration of Methods on Real HealthABC Data

We obtained data from the Health Aging and Body Composition (Health ABC) Study, a cohort study of 70 to 79-year-old adults recruited from Memphis, Tennessee and Pittsburgh, Pennsylvania between April 15, 1997 and June 5, 1998 (Raphael et al., 2016). The study's original mission was to study factors contributing to incident disability in healthier older persons. The HealthABC study's dbgap accession number is phs000169.v1.p1.

For populations "Black" and "White" in the HealthABC data, we imputed Chr 22 SNPs. If an individual was missing a genotype at a SNP, we replaced the missing genotype with the most common genotype in that population at that SNP. If multiple genotypes were most common, we randomly picked one of the most common genotypes to do the replacement. We then excluded monomorphic SNPs, those with allele type 1 frequency of 0 or 1. We then kept every (number of SNPs)/100-th SNP of the Chr 22 vcf file, with the goal of obtaining roughly independent SNPs.

We applied the Weir and Goudet $\hat{\beta}^i, i \in \{1, 2\}$ MME, our Adapted BayeScan copula method (Figure 3.4) with priors $\theta^i \sim Beta(a_i = 36.14143, b_i = 211.9552), i \in \{1, 2\}$ and $Corr(p_{1l}, p_{2l}) \sim Beta(a_{12} = 1, b_{12} = 1)$, Foll and Gaggiotti (2006)'s BayeScan method (Figure 3.3), and Bayenv (Figure 3.2) to estimate population-specific F_{ST} β_1 and population-specific F_{ST} β_2, treating Population 1 as Black and Population 2 as White. For MMEs, 95% confidence intervals were obtained using the hierfstat R package's betas function by bootstrapping over loci. For Bayesian methods, 95% credible intervals were obtained using the 2.5% and 97.5% quantiles of MCMC samples.

We calculated MME's using the hierfstat R package. We obtained MCMC samples for our Adapted BayeScan copula model with Stan, BayeScan samples using the BayeScan software, and Bayenv samples using the Bayenv2.0 software. Adapted BayeScan copula model values $a_i = 36.14143$ and $b_i = 211.9552$ were determined as follows: We bootstrapped over loci's genotypes, calculated $1,000$ Weir and Goudet $\hat{\beta}_1$ MMEs, $1,000$ Weir and Goudet $\hat{\beta}_2$ MMEs, and averaged their values $\hat{\beta}_{avg} = (\hat{\beta}_1 + \hat{\beta}_2)/2$. We then calculated a_i and b_i such that Beta(a_i, b_i) would have a mean corresponding to the sample mean of $1,000$ $\hat{\beta}_{avg}$ values and a variance corresponding to the sample variance of $1,000$ $\hat{\beta}_{avg}$ values. For all three Bayesian methods, we used the rstan R package to compute $\hat{R}$, and checked for $\hat{R} < 1.1$ for every outputted parameter (Brooks and Gelman, 1998), before using MCMC samples to obtain estimates.

3.4 Results: Two Individuals From the Same Population: Simulated Data

3.4.1 Results: Method of Moments Estimation

3.4.1.1 Results: Simple Estimator

In Figure 3.6, we plot simple MME estimates against pedigree-determined $\theta_{jj'}$ values, the simple MME's pedigree-determined parameter $\beta_{jj'}$, and Weir and Goudet weighted $\hat{\beta}_{jj',w}$ estimates, for one-population simulated data. For each $\theta_{jj'}$ value, the mean of simple MME estimates for pairs (j, j') of individuals with that $\theta_{jj'}$ value is lower than the $\theta_{jj'}$ value, consistent with the simple MME being biased for $\theta_{jj'}$. As seen in Chapter 2, this is a consequence of using sample allele frequencies $\tilde{p}_l$ in place of allele probabilities π_l. Simple MME estimates plotted versus their pedigree-determined parameter $\beta_{jj'}$ fall near the $y = x$ line, illustrating that the simple MME is a better estimator of its pedigree-determined parameter than of $\theta_{jj'}$. Simple MME estimates plotted versus Weir and Goudet's weighted

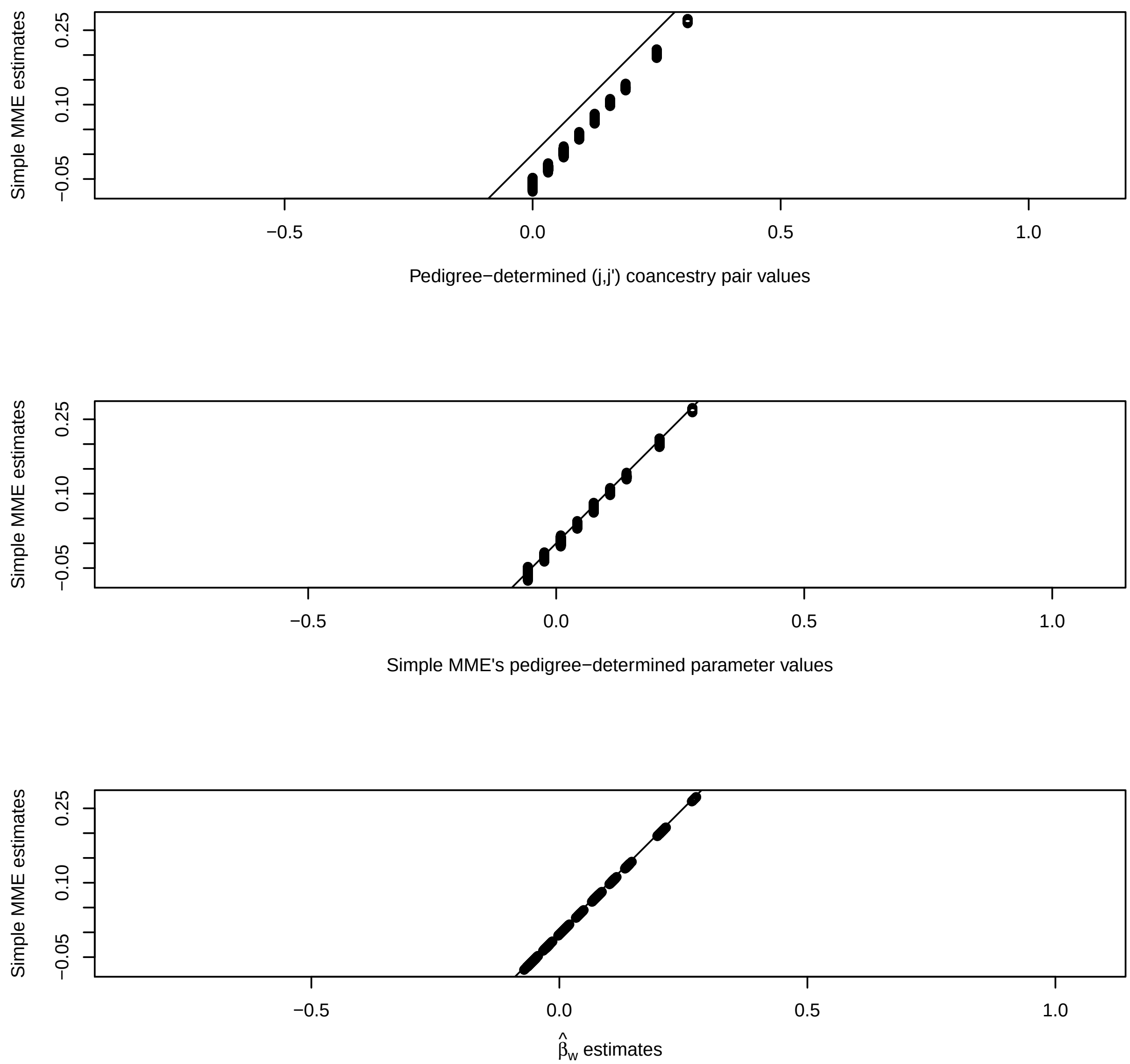

Figure 3.6: Simple MME estimates of within-population coancestry $\beta_{jj'}$, evaluated on simulated one-population data ($n = 115$ individuals' dosages at $L \approx 73$K loci). **Top**: Simple MME estimates vs pedigree-determined coancestry pair $\theta_{jj'}$ values. **Middle**: Simple MME estimates vs Simple MME's pedigree-determined parameter $\beta_{jj'}$. **Bottom**: Simple MME estimates vs Weir and Goudet $\hat{\beta}_{jj',w}$ estimates. Line: y = x.

$\hat{\beta}_{jj',w}$ estimates fall near the $y = x$ line, illustrating that the estimates are close to each other.

3.4.1.2 Results: $\hat{\beta}_{jj',w}$ and $\hat{\beta}_{jj',u}$ Estimators, and Bootstrap Confidence Intervals

Weir and Goudet weighted $\hat{\beta}_{jj',w}$ and unweighted $\hat{\beta}_{jj',u}$ estimates applied to simulated one-population data are plotted against their pedigree-determined parameter values and against pedigree-determined $\theta_{jj'}$ values in Figure 3.7. For each $\theta_{jj'}$ value, the mean of $\hat{\beta}_{jj',w}$ estimates for pairs (j, j') of individuals with that $\theta_{jj'}$ value is lower than the $\theta_{jj'}$ value, consistent with the $\hat{\beta}_{jj',w}$ estimator being biased for $\theta_{jj'}$. Weir and Goudet weighted $\hat{\beta}_{jj',w}$ estimates plotted versus their pedigree-determined parameter $\beta_{jj'}$ values fall near the y = x line, illustrating that they are better estimators of their pedigree-determined parameter $\beta_{jj'}$ than of $\theta_{jj'}$. This observation is consistent with Table 3.2, from which we see that the weighted estimator's MSE for its pedigree-determined parameter is lower than its MSE for $\theta_{jj'}$.

Figure 3.7 also plots Weir and Goudet unweighted $\hat{\beta}_{jj',u}$ estimates versus $\theta_{jj'}$ and Weir and Goudet unweighted $\hat{\beta}_{jj',u}$ estimates versus its pedigree-determined parameter $\beta_{jj'}$. The contrast between the plots is not as obvious as what we saw with the $\hat{\beta}_{jj',w}$ plots. However, Table 3.2 illustrates that the MSE of $\hat{\beta}_{jj',u}$ for its pedigree-determined parameter $\beta_{jj'}$ is lower than the MSE of $\hat{\beta}_{jj',u}$ for $\theta_{jj'}$.

For each of 6,555 individual pairs (j,j')'s, we calculated a 95% confidence interval corresponding to point estimate $\hat{\beta}_{jj',w}$. We calculated coverage as the fraction of 95% CIs that covered the parameter, either $\theta_{jj'}$ or $\hat{\beta}_{jj',w}$'s pedigree-determined parameter $\beta_{jj'}$. Both $\hat{\beta}_{jj',w}$ and $\hat{\beta}_{jj',u}$ bootstrap confidence intervals (CIs) had closer to 95% coverage for their pedigree-determined parameter $\beta_{jj'}$ than for $\theta_{jj'}$. Coverages of $\hat{\beta}_{jj',w}$ nominal 95% bootstrap CIs were 0% for the pedigree-determined $\theta_{jj'}$ and 92.5% for $\hat{\beta}_{jj',w}$ estimator's pedigree-determined parameter $\beta_{jj'}$. Coverages of $\hat{\beta}_{jj',u}$ nominal 95% bootstrap confidence intervals were 1.2% for the pedigree-determined $\theta_{jj'}$ and 11.9% for $\hat{\beta}_{jj',u}$ estimator's pedigree-determined parameter $\beta_{jj'}$.

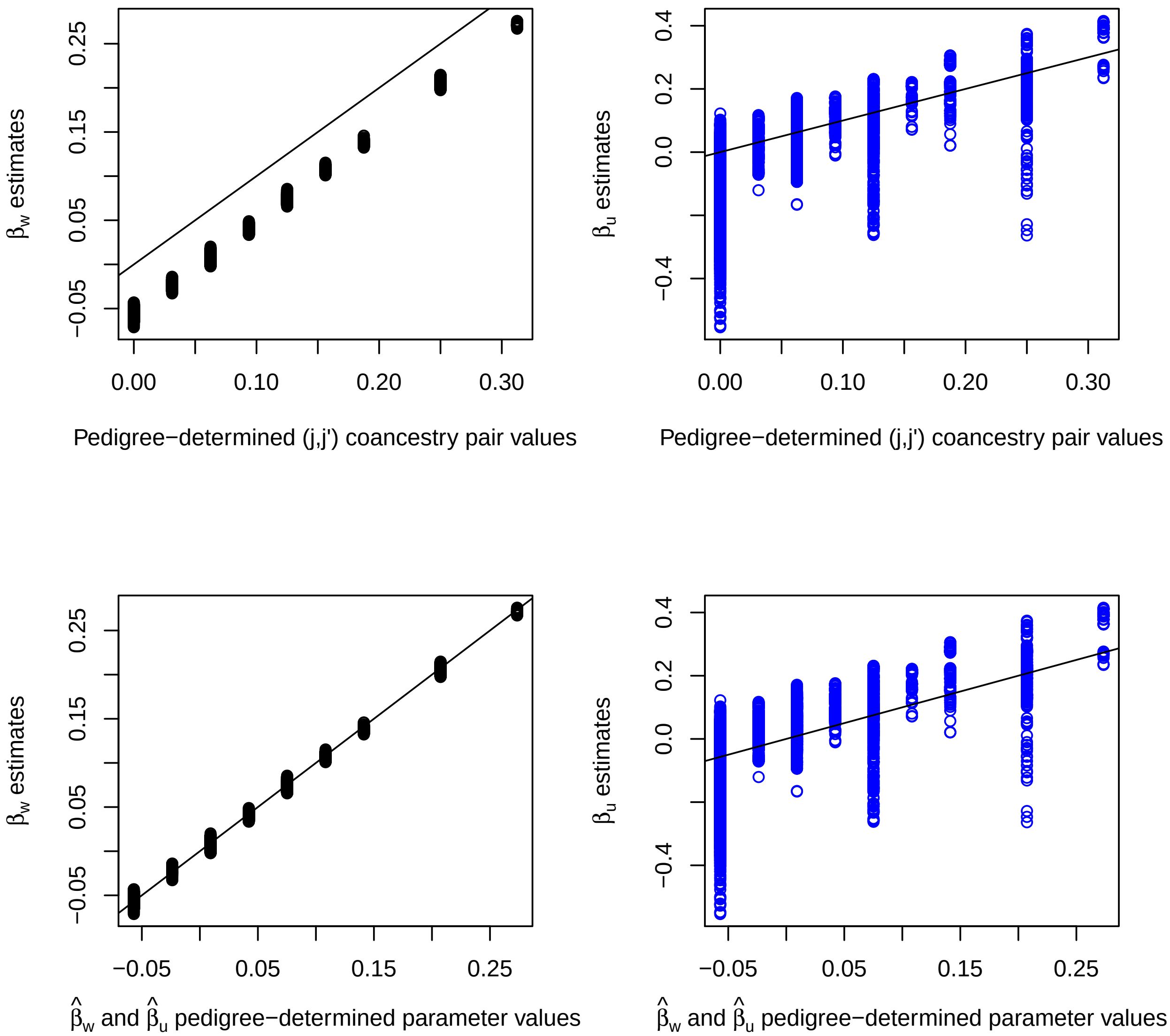

Figure 3.7: Top: Estimates vs Pedigree-determined coancestry pair $\theta_{jj'}$ values. Bottom: Estimates vs $\hat{\beta}_{jj',w}$ and $\hat{\beta}_{jj',u}$'s pedigree-determined parameter $\beta_{jj'}$ values. Left: $\hat{\beta}_{jj',w}$ Estimates. Right: $\hat{\beta}_{jj',u}$ Estimates. Black line in all plots: y = x. Estimates were computed on a simulated one-population data set with $n = 115$ individuals and $\approx$ 73K SNPs. Equation 3.4 gives the $\hat{\beta}_{jj',w}$ formula and Equation 3.5 gives the $\hat{\beta}_{jj',u}$ formula.

Estimator	MSE $\theta_{jj'}$	MSE Pedigree-Determined Parameter
Weir and Goudet $\hat{\beta}_{jj',w}$ MME	2.92×10^{-3}	1.40×10^{-5}
Weir and Goudet $\hat{\beta}_{jj',u}$ MME	1.35×10^{-2}	1.04×10^{-2}
GCTA $\hat{\theta}_{jj',w}$ MME	3.95×10^{-3}	6.39×10^{-6}
GCTA $\hat{\theta}_{jj',u}$ MME	3.96×10^{-3}	9.42×10^{-5}

Table 3.2: MSEs for Weir and Goudet $\hat{\beta}_{jj',w}$ and $\hat{\beta}_{jj',u}$ MMEs and GCTA $\hat{\theta}_{jj',w}$ and $\hat{\theta}_{jj',u}$ MMEs in Simulated One-Population Data. The Weir and Goudet MMEs' pedigree-determined parameter is $\beta_{jj'} := (\theta_{jj'} - \theta_S)/(1 - \theta_S)$. GCTA MMEs' pedigree-determined parameter is given by Equation 3.20.

3.4.1.3 Results: GCTAw and GCTAu Estimators, and Bootstrap Confidence Intervals

GCTAw and GCTAu estimates applied to simulated one-population data are plotted against their pedigree-determined parameter values and against pedigree-determined $\theta_{jj'}$ values in Figure 3.8. For each $\theta_{jj'}$ value, the mean of GCTAw estimates and the mean of GCTAu estimates for pairs (j, j') of individuals with that $\theta_{jj'}$ value are both lower than the $\theta_{jj'}$ value, consistent with the GCTAw and GCTAu estimators being biased for $\theta_{jj'}$. GCTAw and GCTAu estimates plotted versus their pedigree-determined parameter values fall near the y = x line, illustrating that they are better estimators of their pedigree-determined parameter than of $\theta_{jj'}$. GCTAw estimates vs pedigree-determined parameter values fall near the y = x line more so than GCTAu estimates vs pedigree-determined parameter values, illustrating that GCTAw is better than GCTAu as an estimator of GCTAw and GCTAu's pedigree-determined parameter.

We calculated a 95% confidence interval corresponding to point estimate $\hat{\theta}_{jj',GCTAw}$ and a 95% confidence interval corresponding to point estimate $\hat{\theta}_{jj',GCTAu}$. We calculated coverage as the fraction of 6,555 95% CIs that covered the parameter, either $\theta_{jj'}$ or the estimator's pedigree-determined parameter. Both $\hat{\theta}_{jj',w}$ and $\hat{\theta}_{jj',u}$ bootstrap CIs had closer to 95% cov-

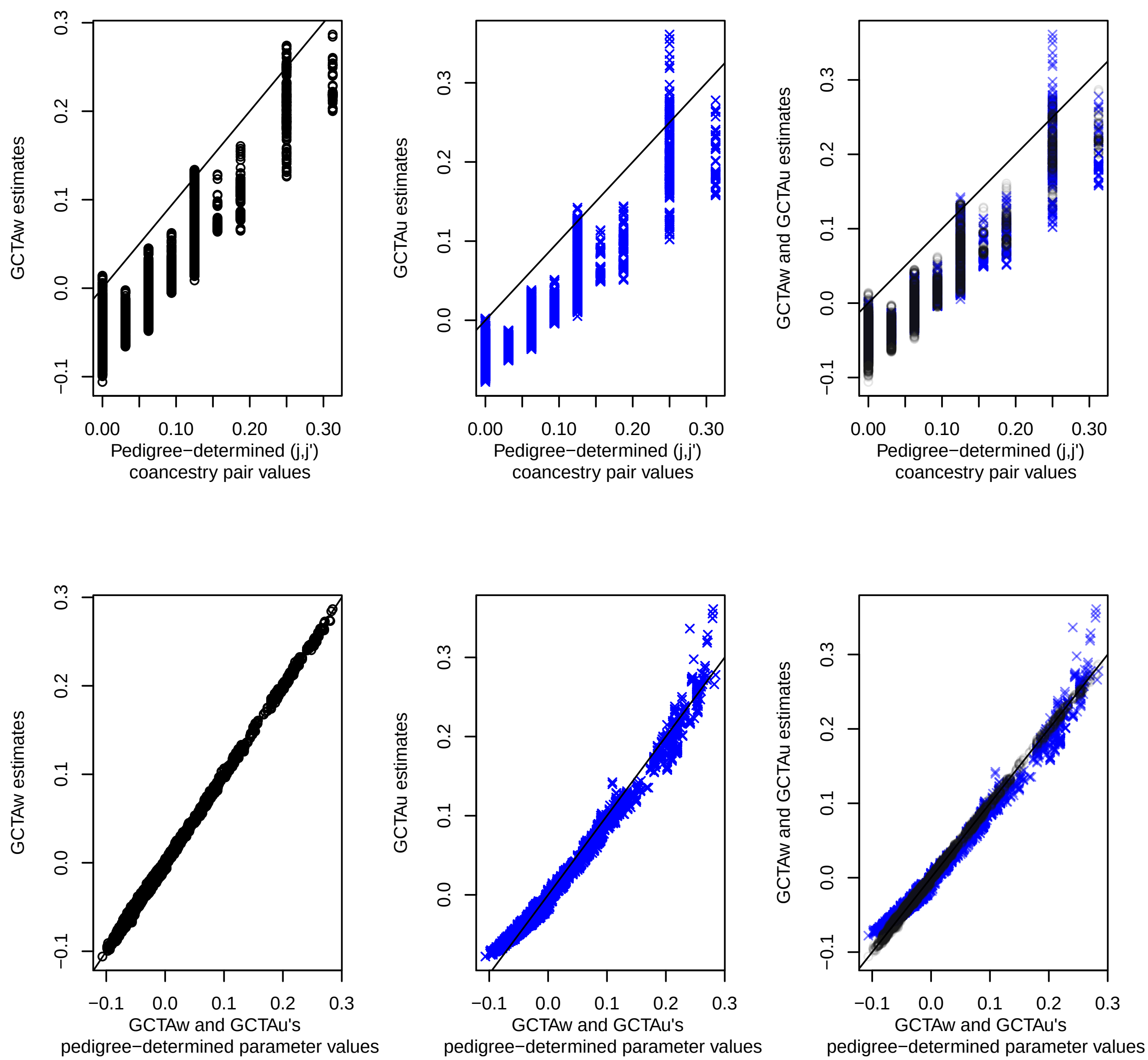

Figure 3.8: **Top:** Estimates vs Pedigree-determined coancestry pair $\theta_{jj'}$ values. **Bottom:** Estimates vs $\hat{\theta}_{jj',GCTAw}$ and $\hat{\theta}_{jj',GCTAu}$'s pedigree-determined parameter values. **Column 1:** GCTA weighted $\hat{\theta}_{jj',GCTAw}$ estimates (black circle). **Column 2:** GCTA unweighted $\hat{\theta}_{jj',GCTAu}$ estimates (blue cross). **Column 3:** GCTA weighted $\hat{\theta}_{jj',GCTAw}$ estimates (black circle) and unweighted $\hat{\theta}_{jj',GCTAu}$ estimates (blue cross). **Black line in all plots:** y = x. Estimates were computed on a simulated one-population data set with $n = 115$ individuals and ≈ 73K SNPs. Equation 3.8 gives the $\hat{\theta}_{jj',GCTAw}$ formula. Equation 3.9 gives the $\hat{\theta}_{jj',GCTAu}$ formula.

erage for their pedigree-determined parameter than for $\theta_{jj'}$. Coverages of $\hat{\theta}_{jj',GCTAw}$ nominal 95% bootstrap confidence intervals were 1.2% for the pedigree-determined $\theta_{jj'}$ and 88.7% for $\hat{\theta}_{jj',GCTAw}$ estimator's pedigree-determined parameter. Coverages of $\hat{\theta}_{jj',GCTAu}$ nominal 95% bootstrap confidence intervals were 1.1% for the pedigree-determined $\theta_{jj'}$ and 28.8% for $\hat{\theta}_{jj',GCTAu}$ estimator's pedigree-determined parameter.

3.5 Results: Two Alleles From Different Populations: Simulated Data

	$\theta^1 = 0.198$				$\theta^2 = 0.024$				$\theta^{12} = 0$			
	Mean	Var	Bias^2	MSE	Mean	Var	Bias^2	MSE	Mean	Var	Bias^2	MSE
ABayeScanInf	1.92×10^{-1}	4.55×10^{-4}	3.35×10^{-5}	4.89×10^{-4}	2.31×10^{-2}	2.66×10^{-8}	9.19×10^{-7}	9.46×10^{-7}	NA	NA	NA	NA
ABayeScanU01	1.94×10^{-1}	5.29×10^{-4}	1.06×10^{-5}	5.40×10^{-4}	2.17×10^{-2}	7.53×10^{-6}	5.27×10^{-6}	1.28×10^{-5}	NA	NA	NA	NA
Bayenv	4.81×10^{-2}	1.13×10^{-5}	2.23×10^{-2}	2.24×10^{-2}	3.33×10^{-2}	3.40×10^{-6}	8.59×10^{-5}	8.93×10^{-5}	-2.47×10^{-2}	5.67×10^{-6}	6.09×10^{-4}	6.15×10^{-4}

	$\beta^1 = \theta^1 = 0.198$				$\beta^2 = \theta^2 = 0.024$			
	Mean	Var	Bias^2	MSE	Mean	Var	Bias^2	MSE
MME	2.01×10^{-1}	2.53×10^{-4}	9.46×10^{-6}	2.63×10^{-4}	2.48×10^{-2}	3.56×10^{-5}	5.84×10^{-7}	3.62×10^{-5}
BayeScanFoll	1.86×10^{-1}	5.07×10^{-4}	1.27×10^{-4}	6.35×10^{-4}	3.11×10^{-2}	5.21×10^{-6}	5.04×10^{-5}	5.56×10^{-5}
ABayeScanInf	1.92×10^{-1}	4.55×10^{-4}	3.35×10^{-5}	4.89×10^{-4}	2.31×10^{-2}	2.66×10^{-8}	9.19×10^{-7}	9.46×10^{-7}
ABayeScanU01	1.94×10^{-1}	5.29×10^{-4}	1.06×10^{-5}	5.40×10^{-4}	2.17×10^{-2}	7.53×10^{-6}	5.27×10^{-6}	1.28×10^{-5}
Bayenv	7.10×10^{-2}	2.91×10^{-5}	1.60×10^{-2}	1.60×10^{-2}	5.66×10^{-2}	1.57×10^{-5}	1.06×10^{-3}	1.07×10^{-3}

Table 3.3: **Two-Population Simulation Study: Assessment of Methods on** 100 **Data Sets Simulated Under Simulation Setting Ind A** $(N_1 = 100, N_2 = 1000, \mu = 0.01, m_1 = 0, m_2 = 0)$**.** **MME**: Weir and Goudet method of moments estimator. **BayeScanFoll**: MCMC samples obtained from the BayeScan implementation of Foll and Gaggiotti (2006)'s model. **ABayeScanInf**: Our Bayesian model with informative priors implemented in Stan version 2.14. **ABayeScanU01**: Our Bayesian model with U(0,1) priors implemented in Stan version 2.14. **Bayenv**: Bayenv implementation of Gunther and Coop (2013)'s model. True parameter values are given in column headers. They were the generation time t = 1,000 values for $\theta^1(t)$, $\theta^2(t)$, $\theta^{12}(t)$, population-1-specific F_{ST} $\beta^1(t)$, and population-2-specific F_{ST} $\beta^2(t)$, determined by iterating transition equations for the two-island model.

Tables 3.3, 3.4, 3.5, 3.6, 3.7, and 3.8 give results for simulation settings Ind A, Ind B,

	$\theta^1 = 0.024$				$\theta^2 = 0.198$				$\theta^{12} = 0$			
	Mean	Var	Bias^2	MSE	Mean	Var	Bias^2	MSE	Mean	Var	Bias^2	MSE
ABayeScanInf	2.31×10^{-2}	2.92×10^{-8}	9.58×10^{-7}	9.87×10^{-7}	1.95×10^{-1}	5.44×10^{-4}	7.23×10^{-6}	5.51×10^{-4}	NA	NA	NA	NA
ABayeScanU01	2.13×10^{-2}	8.46×10^{-6}	7.44×10^{-6}	1.59×10^{-5}	1.98×10^{-1}	6.26×10^{-4}	1.55×10^{-7}	6.26×10^{-4}	NA	NA	NA	NA
Bayenv	3.35×10^{-2}	4.36×10^{-6}	8.94×10^{-5}	9.37×10^{-5}	4.79×10^{-2}	1.51×10^{-5}	2.24×10^{-2}	2.24×10^{-2}	-2.47×10^{-2}	7.74×10^{-6}	6.11×10^{-4}	6.19×10^{-4}

	$\beta^1 = \theta^1 = 0.024$				$\beta^2 = \theta^2 = 0.198$			
	Mean	Var	Bias^2	MSE	Mean	Var	Bias^2	MSE
MME	2.36×10^{-2}	4.57×10^{-5}	1.83×10^{-7}	4.58×10^{-5}	2.02×10^{-1}	3.07×10^{-4}	1.85×10^{-5}	3.25×10^{-4}
BayeScanFoll	3.09×10^{-2}	5.58×10^{-6}	4.75×10^{-5}	5.31×10^{-5}	1.90×10^{-1}	5.90×10^{-4}	6.14×10^{-5}	6.52×10^{-4}
ABayeScanInf	2.31×10^{-2}	2.92×10^{-8}	9.58×10^{-7}	9.87×10^{-7}	1.95×10^{-1}	5.44×10^{-4}	7.23×10^{-6}	5.51×10^{-4}
ABayeScanU01	2.13×10^{-2}	8.46×10^{-6}	7.44×10^{-6}	1.59×10^{-5}	1.98×10^{-1}	6.26×10^{-4}	1.55×10^{-7}	6.26×10^{-4}
Bayenv	5.68×10^{-2}	2.09×10^{-5}	1.07×10^{-3}	1.09×10^{-3}	7.09×10^{-2}	3.94×10^{-5}	1.61×10^{-2}	1.61×10^{-2}

Table 3.4: **Two-Population Simulation Study: Assessment of Methods on** 100 **Data Sets Simulated Under Simulation Setting Ind B** ($N_1 = 1,000, N_2 = 100, \mu = 0.01, m_1 = 0, m_2 = 0$)**.** **MME**: Weir and Goudet method of moments estimator. **BayeScanFoll**: MCMC samples obtained from the BayeScan implementation of Foll and Gaggiotti (2006)'s model. **ABayeScanInf**: Our Bayesian model with informative priors implemented in Stan version 2.14. **ABayeScanU01**: Our Bayesian model with U(0,1) priors implemented in Stan version 2.14. **Bayenv**: Bayenv implementation of Gunther and Coop (2013)'s model. True parameter values are given in column headers. They were the generation time t $= 1,000$ values for $\theta^1(t)$, $\theta^2(t)$, $\theta^{12}(t)$, population-1-specific F_{ST} $\beta^1(t)$, and population-2-specific F_{ST} $\beta^2(t)$, determined by iterating transition equations for the two-island model.

Ind C, Dep A, Dep B, and Dep C respectively. For methods that provide no inference on a parameter, the table entries are listed as NA. Under independence simulation settings for example, it is expected that our Bayesian model with Informative or U(0,1) priors would not yield a θ^{12} estimate, because in those settings, θ^{12} was set to 0. When a method does not estimate or its software implementation does not output estimates for any parameter in a table, the method is excluded from the table. For example, population-specific Weir and Goudet MMEs $\hat{\beta}^1$ and $\hat{\beta}^2$ estimate β^1 and β^2, but not θ^1, θ^2, or θ^{12}, and Foll's BayeScan software outputs MCMC draws for only parameters β^1 and β^2, so is excluded from tables of

	$\theta^1 = 0.024$				$\theta^2 = 0.024$				$\theta^{12} = 0$			
	Mean	Var	Bias^2	MSE	Mean	Var	Bias^2	MSE	Mean	Var	Bias^2	MSE
ABayeScanInf	2.38×10^{-2}	3.18×10^{-6}	5.07×10^{-8}	3.23×10^{-6}	2.40×10^{-2}	3.28×10^{-6}	5.34×10^{-10}	3.28×10^{-6}	NA	NA	NA	NA
ABayeScanU01	2.51×10^{-2}	1.78×10^{-5}	1.22×10^{-6}	1.90×10^{-5}	2.53×10^{-2}	1.78×10^{-5}	1.55×10^{-6}	1.94×10^{-5}	NA	NA	NA	NA
Bayenv	1.69×10^{-2}	3.59×10^{-7}	5.14×10^{-5}	5.18×10^{-5}	1.68×10^{-2}	3.56×10^{-7}	5.19×10^{-5}	5.22×10^{-5}	-2.03×10^{-3}	3.35×10^{-7}	4.11×10^{-6}	4.44×10^{-6}

	$\beta^1 = \theta^1 = 0.024$				$\beta^2 = \theta^2 = 0.024$			
	Mean	Var	Bias^2	MSE	Mean	Var	Bias^2	MSE
MME	2.43×10^{-2}	1.33×10^{-5}	6.56×10^{-8}	1.34×10^{-5}	2.42×10^{-2}	9.45×10^{-6}	4.61×10^{-8}	9.50×10^{-6}
BayeScanFoll	2.72×10^{-2}	1.66×10^{-5}	9.97×10^{-6}	2.66×10^{-5}	2.72×10^{-2}	1.55×10^{-5}	1.00×10^{-5}	2.56×10^{-5}
ABayeScanInf	2.38×10^{-2}	3.18×10^{-6}	5.07×10^{-8}	3.23×10^{-6}	2.40×10^{-2}	3.28×10^{-6}	5.34×10^{-10}	3.28×10^{-6}
ABayeScanU01	2.51×10^{-2}	1.78×10^{-5}	1.22×10^{-6}	1.90×10^{-5}	2.53×10^{-2}	1.78×10^{-5}	1.55×10^{-6}	1.94×10^{-5}
Bayenv	1.89×10^{-2}	1.35×10^{-6}	2.69×10^{-5}	2.82×10^{-5}	1.88×10^{-2}	1.34×10^{-6}	2.72×10^{-5}	2.86×10^{-5}

Table 3.5: **Two-Population Simulation Study: Assessment of Methods on** 100 **Data Sets Simulated Under Simulation Setting Ind C (**$N_1 = 1,000, N_2 = 1,000, \mu = 0.01, m_1 = 0, m_2 = 0$**). MME**: Weir and Goudet method of moments estimator. **BayeScanFoll**: MCMC samples obtained from the BayeScan implementation of Foll and Gaggiotti (2006)'s model. **ABayeScanInf**: Our Bayesian model with informative priors implemented in Stan version 2.14. **ABayeScanU01**: Our Bayesian model with U(0,1) priors implemented in Stan version 2.14. **Bayenv**: Bayenv implementation of Gunther and Coop (2013)'s model. True parameter values are given in column headers. They were the generation time t $= 1,000$ values for $\theta^1(t)$, $\theta^2(t)$, $\theta^{12}(t)$, population-1-specific F_{ST} $\beta^1(t)$, and population-2-specific F_{ST} $\beta^2(t)$, determined by iterating transition equations for the two-island model.

θ^1, θ^2, and θ^{12} results.

In simulation settings Ind A-C and Dep A-C, the Weir and Goudet $\hat{\beta}^i$ MME, our Adapted BayeScan copula model with Informative priors, and our Adapted BayeScan copula model with U(0,1) priors all yield lower β^1 MSE and lower β^2 MSE than BayeScan with Foll's implementation and Bayenv. In settings Dep A and Dep C, the true β^2 is negative. In simulation settings Dep A and Dep C, the MME and our Adapted BayeScan copula model (with either Informative or U(0,1) priors) yield negative mean $\hat{\beta}^2$ estimates, but the BayeScan and Bayenv models yield positive mean estimates, resulting in BayeScan and Bayenv each

	$\theta^1 = 0.126$				$\theta^2 = 0.032$				$\theta^{12} = 0.0403$			
	Mean	Var	Bias^2	MSE	Mean	Var	Bias^2	MSE	Mean	Var	Bias^2	MSE
ABayeScanInf	1.25×10^{-1}	2.23×10^{-4}	9.61×10^{-7}	2.24×10^{-4}	2.62×10^{-2}	1.32×10^{-7}	3.29×10^{-5}	3.31×10^{-5}	3.53×10^{-2}	1.93×10^{-6}	2.52×10^{-5}	2.71×10^{-5}
ABayeScanU01	1.11×10^{-1}	3.04×10^{-4}	2.32×10^{-4}	5.36×10^{-4}	1.83×10^{-2}	3.03×10^{-6}	1.87×10^{-4}	1.90×10^{-4}	2.25×10^{-2}	8.74×10^{-6}	3.16×10^{-4}	3.25×10^{-4}
Bayenv	2.35×10^{-2}	1.85×10^{-6}	1.06×10^{-2}	1.06×10^{-2}	1.97×10^{-2}	4.30×10^{-7}	1.50×10^{-4}	1.50×10^{-4}	-6.52×10^{-3}	8.40×10^{-7}	2.19×10^{-3}	2.19×10^{-3}

	$\beta^1 = 0.0897$				$\beta^2 =$ -0.00867			
	Mean	Var	Bias^2	MSE	Mean	Var	Bias^2	MSE
MME	9.35×10^{-2}	6.23×10^{-5}	1.42×10^{-5}	7.65×10^{-5}	-8.00×10^{-3}	1.15×10^{-5}	4.51×10^{-7}	1.20×10^{-5}
BayeScanFoll	7.07×10^{-2}	1.94×10^{-4}	3.61×10^{-4}	5.54×10^{-4}	2.26×10^{-2}	4.12×10^{-6}	9.79×10^{-4}	9.83×10^{-4}
ABayeScanInf	9.35×10^{-2}	2.06×10^{-4}	1.45×10^{-5}	2.21×10^{-4}	-9.46×10^{-3}	3.12×10^{-6}	6.19×10^{-7}	3.74×10^{-6}
ABayeScanU01	9.09×10^{-2}	2.59×10^{-4}	1.49×10^{-6}	2.61×10^{-4}	-4.59×10^{-3}	1.70×10^{-5}	1.67×10^{-5}	3.37×10^{-5}
Bayenv	2.98×10^{-2}	4.92×10^{-6}	3.59×10^{-3}	3.60×10^{-3}	2.61×10^{-2}	2.30×10^{-6}	1.21×10^{-3}	1.21×10^{-3}

Table 3.6: **Two-Population Simulation Study: Assessment of Methods on** 100 **Data Sets Simulated Under Simulation Setting Dep A** ($N_1 = 100, N_2 = 1,000, \mu = 0.01, m_1 = 0.01, m_2 = 0.01$)**.** **MME**: Weir and Goudet method of moments estimator. **BayeScanFoll**: MCMC samples obtained from the BayeScan implementation of Foll and Gaggiotti (2006)'s model. **ABayeScanInf**: Our Bayesian model with informative priors implemented in Stan version 2.14. **ABayeScanU01**: Our Bayesian model with U(0,1) priors implemented in Stan version 2.14. **Bayenv**: Bayenv implementation of Gunther and Coop (2013)'s model. True parameter values for this setting are given in column headers. They were the generation time t = $1,000$ values for $\theta^1(t)$, $\theta^2(t)$, $\theta^{12}(t)$, population-1-specific F_{ST} $\beta^1(t)$, and population-2-specific F_{ST} $\beta^2(t)$, determined by iterating transition equations for the two-island model.

having higher squared bias. BayeScan and Bayenv's higher squared bias contributes to their higher MSE, since MSE = Variance + Bias^2, for β^2 than MME and Adapted BayeScan models.

Figures 3.9, 3.10, 3.11, 3.12, 3.13, and 3.14 give violin plots for Weir and Goudet β^1 and β^2 MMEs evaluated on 100 simulated two-island model data sets for the six simulation settings of Table 3.1. Estimates of β^i appear roughly centered around generation $t = 1,000$ $\beta^i(t)$ values obtained from two-island model transition equations, which are given in Table

	$\theta^1 = 0.0161$				$\theta^2 = 0.0161$				$\theta^{12} = 0.00817$			
	Mean	Var	Bias^2	MSE	Mean	Var	Bias^2	MSE	Mean	Var	Bias^2	MSE
ABayeScanInf	1.62×10^{-2}	3.73×10^{-7}	9.06×10^{-9}	3.82×10^{-7}	1.60×10^{-2}	3.31×10^{-7}	6.53×10^{-10}	3.31×10^{-7}	7.93×10^{-3}	1.46×10^{-8}	5.68×10^{-8}	7.14×10^{-8}
ABayeScanU01	2.10×10^{-2}	5.34×10^{-6}	2.47×10^{-5}	3.01×10^{-5}	2.12×10^{-2}	6.12×10^{-6}	2.64×10^{-5}	3.25×10^{-5}	1.22×10^{-2}	2.83×10^{-7}	1.62×10^{-5}	1.65×10^{-5}
Bayenv	1.39×10^{-2}	5.80×10^{-8}	4.61×10^{-6}	4.66×10^{-6}	1.39×10^{-2}	5.24×10^{-8}	4.69×10^{-6}	4.75×10^{-6}	9.25×10^{-4}	5.66×10^{-8}	5.25×10^{-5}	5.26×10^{-5}

	$\beta^1 = 0.00795$				$\beta^2 = 0.00795$			
	Mean	Var	Bias^2	MSE	Mean	Var	Bias^2	MSE
MME	8.82×10^{-3}	4.07×10^{-6}	7.44×10^{-7}	4.81×10^{-6}	8.57×10^{-3}	4.28×10^{-6}	3.81×10^{-7}	4.66×10^{-6}
BayeScanFoll	1.86×10^{-1}	5.06×10^{-4}	1.27×10^{-4}	6.33×10^{-4}	3.12×10^{-2}	5.02×10^{-6}	5.14×10^{-5}	5.64×10^{-5}
ABayeScanInf	8.29×10^{-3}	5.41×10^{-7}	1.12×10^{-7}	6.52×10^{-7}	8.16×10^{-3}	4.89×10^{-7}	4.50×10^{-8}	5.34×10^{-7}
ABayeScanU01	8.94×10^{-3}	4.49×10^{-6}	9.78×10^{-7}	5.47×10^{-6}	9.11×10^{-3}	4.62×10^{-6}	1.35×10^{-6}	5.97×10^{-6}
Bayenv	1.30×10^{-2}	2.21×10^{-7}	2.55×10^{-5}	2.57×10^{-5}	1.30×10^{-2}	2.11×10^{-7}	2.53×10^{-5}	2.55×10^{-5}

Table 3.7: **Two-Population Simulation Study: Assessment of Methods on** 100 **Data Sets Simulated Under Simulation Setting Dep B (**$N_1 = 1,000, N_2 = 1,000, \mu = 0.01, m_1 = 0.01, m_2 = 0.01$**). MME**: Weir and Goudet method of moments estimator. **BayeScanFoll**: MCMC samples obtained from the BayeScan implementation of Foll and Gaggiotti (2006)'s model. **ABayeScanInf**: Our Bayesian model with informative priors implemented in Stan version 2.14. **ABayeScanU01**: Our Bayesian model with U(0,1) priors implemented in Stan version 2.14. **Bayenv**: Bayenv implementation of Gunther and Coop (2013)'s model. True parameter values for this setting are given in column headers. They were the generation time t = 1,000 values for $\theta^1(t)$, $\theta^2(t)$, $\theta^{12}(t)$, population-1-specific F_{ST} $\beta^1(t)$, and population-2-specific F_{ST} $\beta^2(t)$, determined by iterating transition equations for the two-island model.

3.1 per simulation setting. This observation is consistent with Tables 3.3, 3.4, 3.5, 3.6, 3.7, and 3.8 in which we see that the mean Weir and Goudet $\hat{\beta}^i$ MME estimate is close to the true parameter value β^i determined by transition equations. The mean would perhaps be closer to the true parameter value if the number of simulated data sets per simulation setting were greater than 100 or the number of loci L per simulated data set were greater than 100.

Note that simulation settings Ind A and Ind B are symmetric in that they have the same 5 two-island model population genetic parameter values, except that population sizes N_1 and

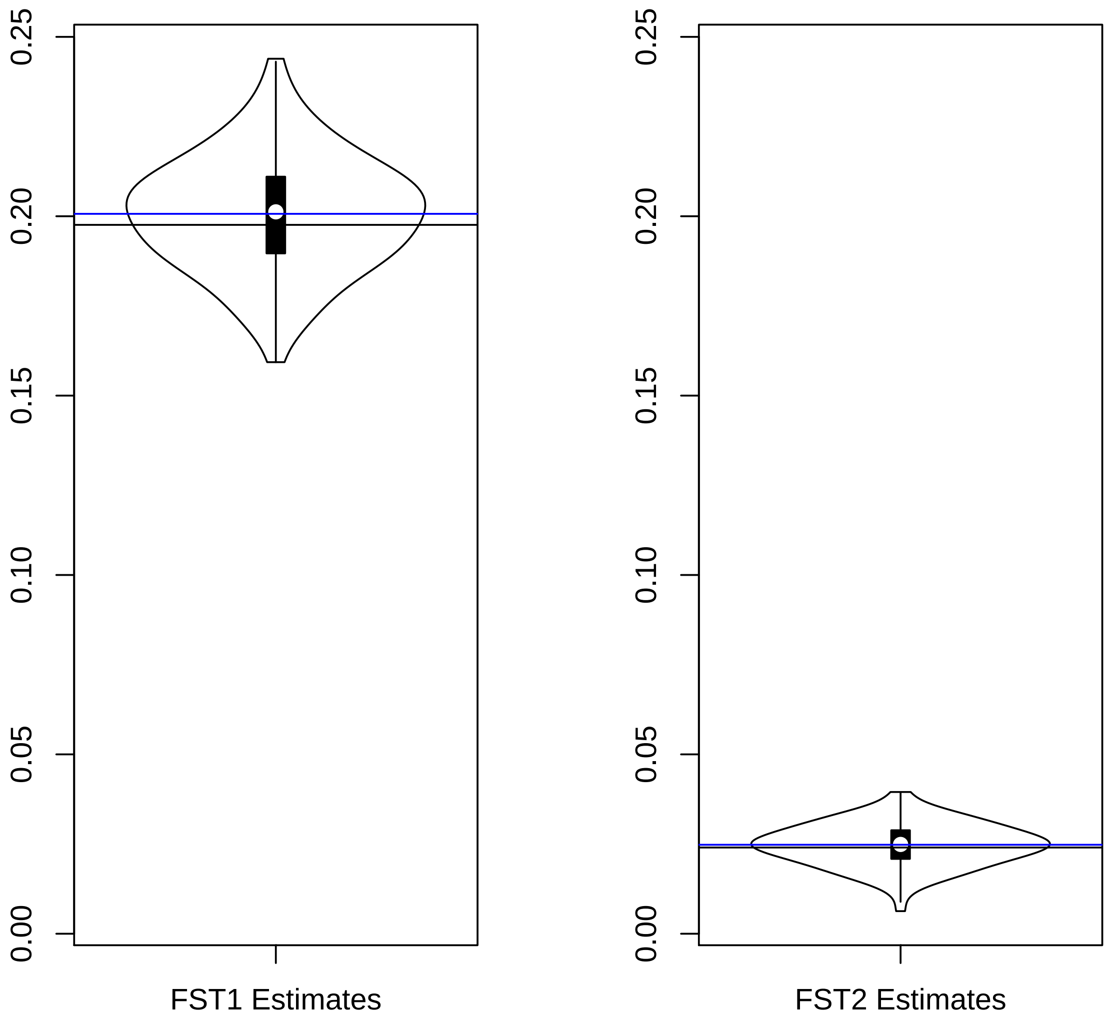

Figure 3.9: Violin plots of 100 Weir and Goudet (2017) population-1-specific F_{ST} β^1 MME estimates (left) and 100 Weir and Goudet (2017) population-2-specific F_{ST} β^2 MME estimates (right) for simulation setting Ind A. Simulation setting Ind A is defined by 5 population genetic parameter values resulting in true θ^1, θ^2, θ^{12}, β^1, and β^2 values given in Table 3.1. Black line: β^i value determined by two-island model's transition equations. Blue line: Mean of 100 β^i estimates, with one estimate obtained per data set simulated under this setting.

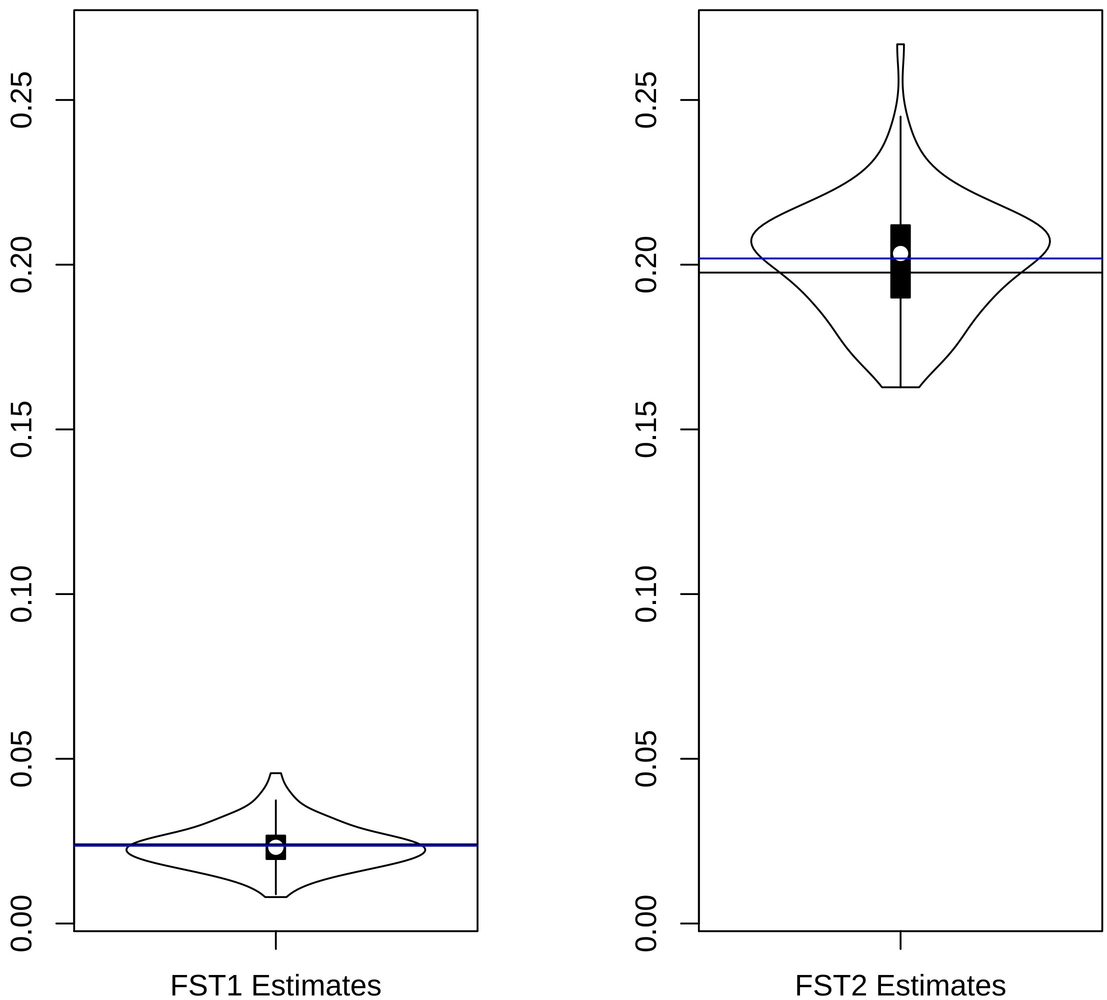

Figure 3.10: Violin plots of 100 Weir and Goudet (2017) population-1-specific F_{ST} β^1 MME estimates (left) and 100 Weir and Goudet (2017) population-2-specific F_{ST} β^2 MME estimates (right) for simulation setting Ind B. Simulation setting Ind B is defined by 5 population genetic parameter values resulting in true θ^1, θ^2, θ^{12}, β^1, and β^2 values given in Table 3.1. Black line: β^i value determined by two-island model's transition equations. Blue line: Mean of 100 β^i estimates, with one estimate obtained per data set simulated under this setting.

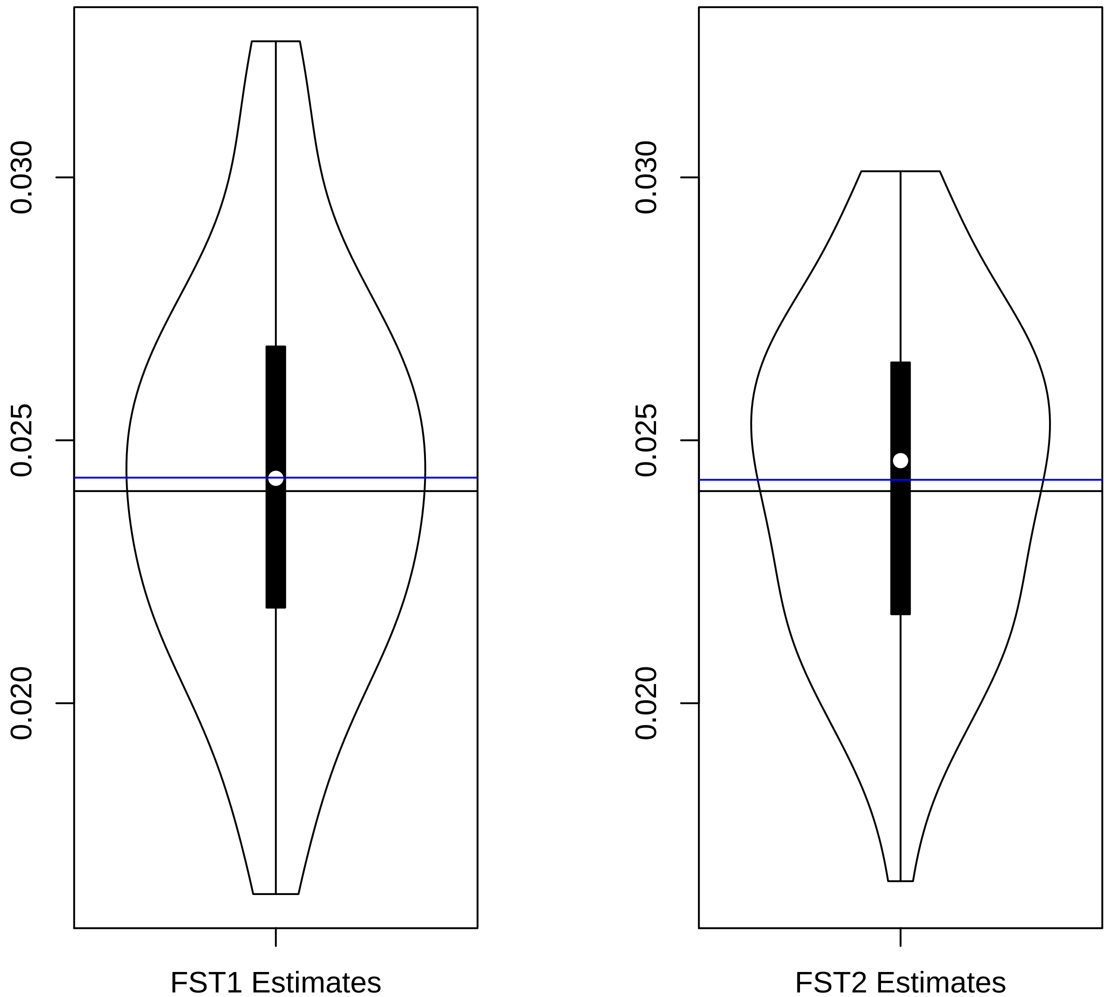

Figure 3.11: Violin plots of 100 Weir and Goudet (2017) population-1-specific F_{ST} β^1 MME estimates (left) and 100 Weir and Goudet (2017) population-2-specific F_{ST} β^2 MME estimates (right) for simulation setting Ind C. Simulation setting Ind C is defined by 5 population genetic parameter values resulting in true θ^1, θ^2, θ^{12}, β^1, and β^2 values given in Table 3.1. Black line: β^i value determined by two-island model's transition equations. Blue line: Mean of 100 β^i estimates, with one estimate obtained per data set simulated under this setting.

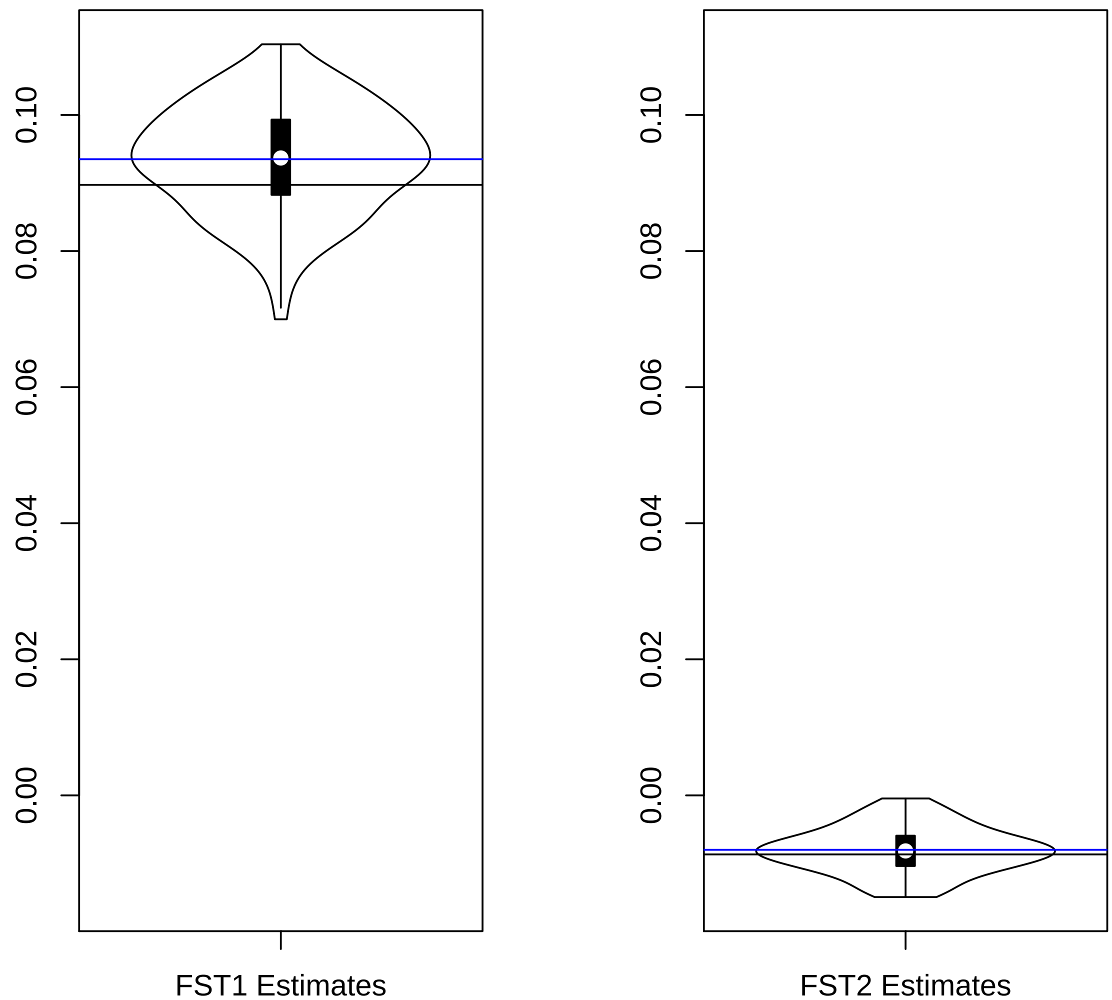

Figure 3.12: Violin plots of 100 Weir and Goudet (2017) population-1-specific F_{ST} β^1 MME estimates (left) and 100 Weir and Goudet (2017) population-2-specific F_{ST} β^2 MME estimates (right) for simulation setting Dep A. Simulation setting Dep A is defined by 5 population genetic parameter values resulting in true θ^1, θ^2, θ^{12}, β^1, and β^2 values given in Table 3.1. Black line: β^i value determined by two-island model's transition equations. Blue line: Mean of 100 β^i estimates, with one estimate obtained per data set simulated under this setting.

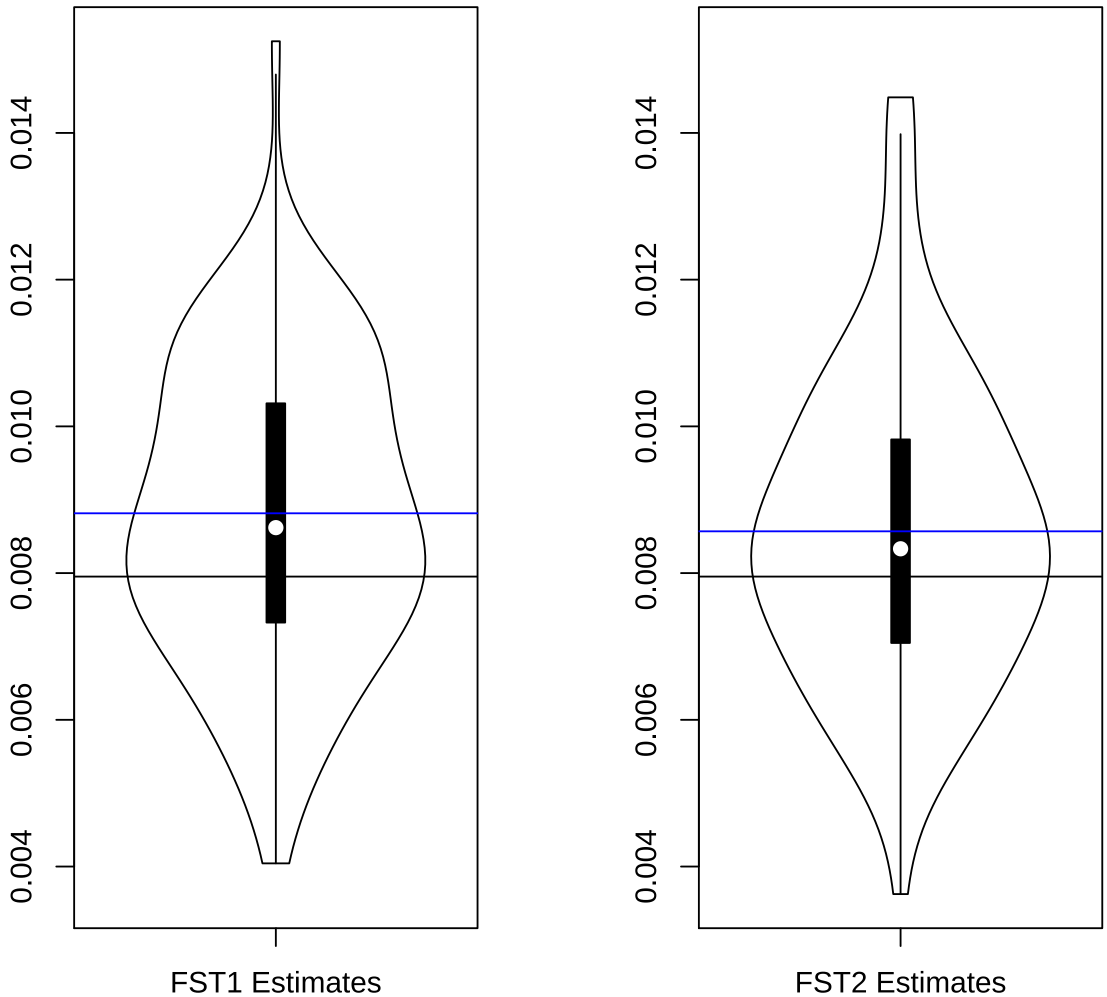

Figure 3.13: Violin plots of 100 Weir and Goudet (2017) population-1-specific F_{ST} β^1 MME estimates (left) and 100 Weir and Goudet (2017) population-2-specific F_{ST} β^2 MME estimates (right) for simulation setting Dep B. Simulation setting Dep B is defined by 5 population genetic parameter values resulting in true θ^1, θ^2, θ^{12}, β^1, and β^2 values given in Table 3.1. Black line: β^i value determined by two-island model's transition equations. Blue line: Mean of 100 β^i estimates, with one estimate obtained per data set simulated under this setting.

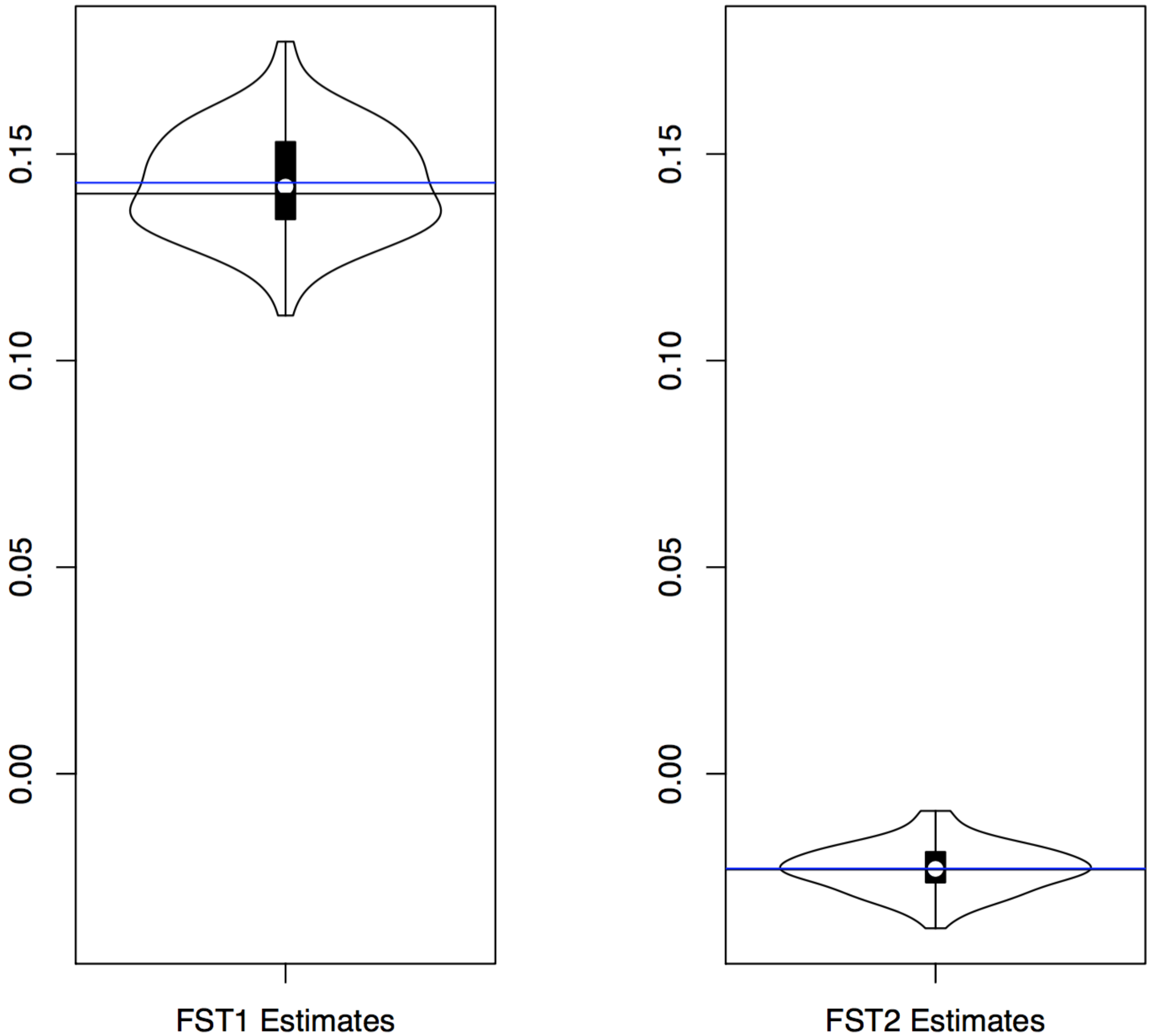

Figure 3.14: Violin plots of 100 Weir and Goudet (2017) population-1-specific F_{ST} β^1 MME estimates (left) and 100 Weir and Goudet (2017) population-2-specific F_{ST} β^2 MME estimates (right) for simulation setting Dep C. Simulation setting Dep C is defined by 5 population genetic parameter values resulting in true θ^1, θ^2, θ^{12}, β^1, and β^2 values given in Table 3.1. Black line: β^i value determined by two-island model's transition equations. Blue line: Mean of 100 β^i estimates, with one estimate obtained per data set simulated under this setting.

	$\theta^1 = 0.198$				$\theta^2 = 0.0449$				$\theta^{12} = 0.0665$			
	Mean	Var	Bias^2	MSE	Mean	Var	Bias^2	MSE	Mean	Var	Bias^2	MSE
ABayeScanInf	1.95×10^{-1}	3.27×10^{-4}	6.10×10^{-6}	3.33×10^{-4}	3.06×10^{-2}	2.05×10^{-7}	2.05×10^{-4}	2.06×10^{-4}	5.44×10^{-2}	1.11×10^{-5}	1.47×10^{-4}	1.58×10^{-4}
ABayeScanU01	1.70×10^{-1}	4.48×10^{-4}	7.71×10^{-4}	1.22×10^{-3}	1.85×10^{-2}	2.57×10^{-6}	6.98×10^{-4}	7.01×10^{-4}	3.22×10^{-2}	4.32×10^{-5}	1.18×10^{-3}	1.22×10^{-3}
Bayenv	3.08×10^{-2}	4.43×10^{-6}	2.78×10^{-2}	2.78×10^{-2}	2.24×10^{-2}	8.00×10^{-7}	5.09×10^{-4}	5.09×10^{-4}	-1.10×10^{-2}	1.72×10^{-6}	6.01×10^{-3}	6.01×10^{-3}

	$\beta^1 = 0.14$				$\beta^2 =$ -0.0232			
	Mean	Var	Bias^2	MSE	Mean	Var	Bias^2	MSE
MME	1.43×10^{-1}	1.55×10^{-4}	6.94×10^{-6}	1.62×10^{-4}	-2.29×10^{-2}	3.21×10^{-5}	4.62×10^{-8}	3.22×10^{-5}
BayeScanFoll	1.17×10^{-1}	2.67×10^{-4}	5.71×10^{-4}	8.38×10^{-4}	2.13×10^{-2}	3.16×10^{-6}	1.98×10^{-3}	1.98×10^{-3}
ABayeScanInf	1.49×10^{-1}	2.85×10^{-4}	7.46×10^{-5}	3.59×10^{-4}	-2.54×10^{-2}	1.24×10^{-5}	5.13×10^{-6}	1.76×10^{-5}
ABayeScanU01	1.43×10^{-1}	3.29×10^{-4}	4.88×10^{-6}	3.34×10^{-4}	-1.47×10^{-2}	3.58×10^{-5}	7.09×10^{-5}	1.07×10^{-4}
Bayenv	4.14×10^{-2}	1.09×10^{-5}	9.81×10^{-3}	9.82×10^{-3}	3.30×10^{-2}	4.46×10^{-6}	3.15×10^{-3}	3.16×10^{-3}

Table 3.8: **Two-Population Simulation Study: Assessment of Methods on** 100 **Data Sets Simulated Under Simulation Setting Dep C** ($N_1 = 100, N_2 = 1,000, \mu = 0.01, m_1 = 0, m_2 = 0.01$)**.** **MME**: Weir and Goudet method of moments estimator. **BayeScanFoll**: MCMC samples obtained from the BayeScan implementation of Foll and Gaggiotti (2006)'s model. **ABayeScanInf**: Our Bayesian model with informative priors implemented in Stan version 2.14. **ABayeScanU01**: Our Bayesian model with U(0,1) priors implemented in Stan version 2.14. **Bayenv**: Bayenv implementation of Gunther and Coop (2013)'s model. True parameter values are given in column headers. They were the generation time t $= 1,000$ values for $\theta^1(t)$, $\theta^2(t)$, $\theta^{12}(t)$, population-1-specific F_{ST} $\beta^1(t)$, and population-2-specific F_{ST} $\beta^2(t)$, determined by iterating transition equations for the two-island model.

N_2 are switched, so that the transition equation-determined β^1 and β^2 values are switched (Table 3.1). As expected, comparing Tables 3.3 and 3.4, we see that each method's two orders of magnitude, one for the β^1 MSE and one for the β^2 MSE, are switched between the simulation settings. For example, in Table 3.3, the MME has a β^1 MSE on the order of 10^{-4} and the β^2 MSE on the order of 10^{-5}, while in Table 3.4, the MME has a β^1 MSE on the order of 10^{-5} and the β^2 MSE on the order of 10^{-4}. This pattern holds for the other methods and acts as a check of methods' implementations.

3.6 Results: Two Alleles From Different Populations: HealthABC Cohort Study

Results of the analysis treating African Americans as one population and Caucasian Americans as another population and applying the Weir and Goudet $\hat{\beta}^{ii}$ MME, our Adapted BayeScan Bayesian hierarchical copula model, Foll et al.'s BayeScan model, and Bayenv are given in Table 3.9. Figure 3.6 plots values given in Table 3.9.

From Figure 3.6 and Table 3.9, we see that the Bayenv method can give a $\hat{\theta}^{12}$ estimate that is negative, and so out of range for an ibd probability. Comparing methods, we see that Bayenv gives the lowest $F_{ST,1}$ estimate and the lowest $F_{ST,2}$ estimate. We see that our Adapted BayeScan copula model, BayeScan, and the MME have overlapping $F_{ST,1}$ intervals and overlapping $F_{ST,2}$ intervals, but Bayenv's intervals do not overlap with all of the other methods' corresponding intervals. For example, for $F_{ST,1}$, Bayenv's interval barely overlaps with Bayescan and MME intervals, and does not overlap with our Adapted Bayescan copula interval. For $F_{ST,2}$, Bayenv's interval overlaps with Bayescan and MME intervals, and barely overlaps with our Adapted Bayescan copula interval. For θ^1, θ^2, and θ^{12}, our Adapted Bayescan intervals do not overlap with Bayescan intervals.

3.7 Discussion

We have described existing estimators of $\theta_{jj'}$ or $\beta_{jj'}$, and a novel estimator of the coancestry coefficient $\theta_{jj'}$ between individuals j and j', described bootstrap procedures for obtaining confidence intervals for method of moments estimators, evaluated the simple MME, weighted and unweighted Weir and Goudet MMEs $\hat{\beta}_{jj',w}$ and $\hat{\beta}_{jj',u}$, and weighted and unweighted GCTA MMEs $\hat{\theta}_{jj'_{\text{GCTAw}}}$ and $\hat{\theta}_{jj'_{\text{GCTAu}}}$ on simulated one-population data consisting of $n = 115$ individuals' dosages at $L \approx 73$K bi-allelic SNPs. We find that Weir and Goudet MMEs and GCTA MMEs are better at estimating their pedigree-determined parameters than at estimating $\theta_{jj'}$, as determined by MSE. Furthermore, plots of the simple MME, Weir and Goudet weighted and unweighted MME, and GCTA weighted and unweighted MMEs clearly

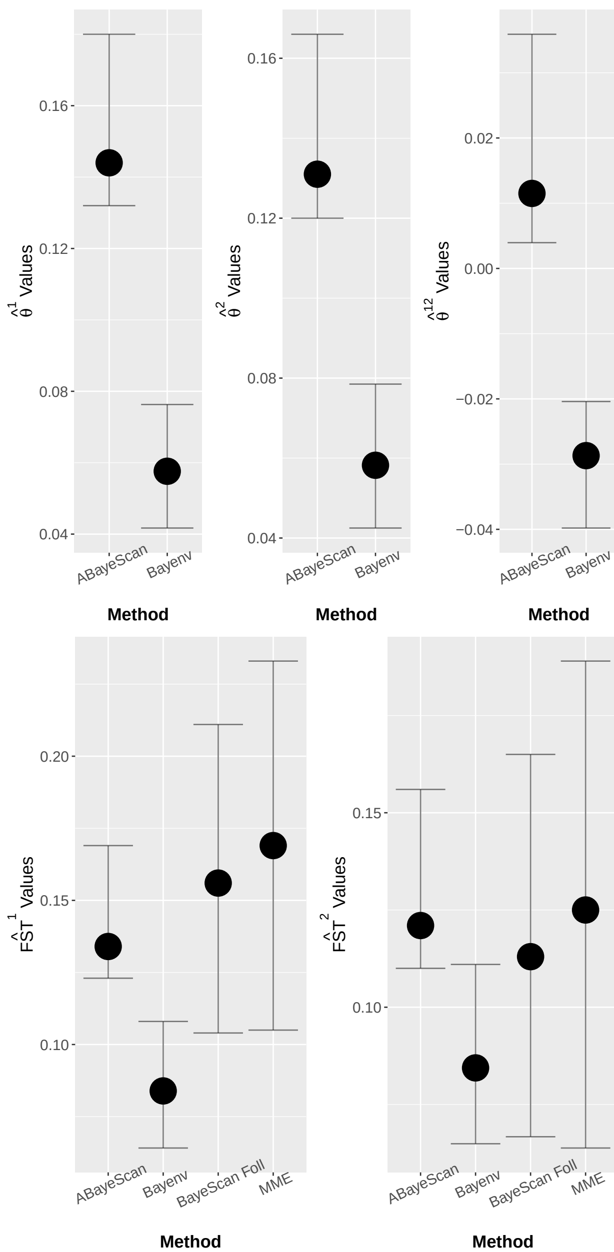

Figure 3.15: Weir and Goudet MME, our Adapted BayeScan Copula Model, BayeScan, and Bayenv Applied to HealthABC Data. **MME**: Weir and Goudet MME implemented with hierfstat. **BayeScan Foll**: Foll and Gaggiotti (2006)'s model implemented with BayeScan. **ABayeScan**: Our Adapted BayeScan copula model implemented with Stan. **Bayenv**: Gunther and Coop (2013)'s model implemented with Bayenv. $\beta^{\mathbf{i}}, \mathbf{i} \in \{1, 2\}$: Population i-specific F_{ST} $\beta^i := (\theta^i - \theta^{12})/(1 - \theta^{12})$. $\theta^1, \theta^2, \theta^{12}$: Population-1, Population-2, and Population-12 ibd probabilities respectively. **Dot**: Point estimate. **Bars**: 2.5% and 97.5% endpoints of confidence interval (for MME) or credible interval (for a Bayesian method).

illustrate that these estimators are biased for $\theta_{jj'}$.

Some estimators tell us more about $\theta_{jj'}$ than other estimators, because their pedigree-determined parameters tells us more about $\theta_{jj'}$ than other pedigree-determined parameters. The pedigree-determined parameters were $\beta_{jj'}$ for the weighted and unweighted Weir and Goudet MMEs and for the simple MME, and $(2(\theta_{jj'} + \theta_S) - 2\psi_{j'} - 2\psi_j)/(2(1 - \theta_S))$ for the weighted and unweighted GCTA MMEs. Weir and Goudet MMEs' pedigree-determined parameter $\beta_{jj'} = (\theta_{jj'} - \theta_S)/(1 - \theta_S)$ is a monotone-increasing function of $\theta_{jj'}$ for fixed θ_S, so ranks in the same order as $\theta_{jj'}$, and its sign indicates whether $\theta_{jj'}$ is less than, equal to, or greater than θ_S. The weighted and unweighted GCTA MMEs' pedigree-determined parameter is a function of $\theta_{jj'}$, θ_S, $\psi_{j'}$, and ψ_j, not just $\theta_{jj'}$.

We reviewed estimators of population-i-specific F_{ST} $\beta^i = (\theta^i - \theta_B)/(1 - \theta_B)$: the Weir and Goudet $\hat{\beta}^i$ MME (which allows $\theta_B > 0$), and existing Bayesian methods by Foll and Gaggiotti (2006) (which assumes $\theta_B = 0$) and Gunther and Coop (2013) (which allows $\theta_B > 0$). Using a copula, we developed a novel two-population Bayesian estimator of β^i that accounts for dependent population allele frequencies (allowing $\theta_B > 0$), while maintaining population genetic-justified Beta distributions for allele frequencies' marginal distributions, unlike Gunther and Coop (2013)'s model.

We evaluated these β^i estimators' under six simulation settings, with 100 data sets per setting simulated according to a two-island model. Each data set consisted of two populations' data, with $n = 100$ individuals' genotypes at $L = 100$ bi-allelic loci. We found that Weir and Goudet $\hat{\beta}^i$ MMEs were approximately unbiased for β^i in all six settings. We also found that in the two simulation settings with $\theta^i < \theta_B$ (so a negative population-i specific F_{ST} β^i), the MME and our Bayesian copula model yielded mean estimates that were negative, while the other Bayesian methods did not yield negative mean estimates, resulting in the MME and copula model having lower MSEs for the negative β^i than the other methods.

We evaluated these β^i estimators on real HealthABC data and found that the choice of estimator made a difference. Gunther and Coop (2013)'s model implemented in their Bayenv software gave a β^1 interval that barely overlapped with BayeScan and MME intervals, and did not overlap with our Adapted Bayescan copula interval. For θ^1, θ^2, and θ^{12}, our Adapted Bayescan copula model intervals did not overlap with Bayenv intervals.

Weir and Goudet's method of moments estimator $\hat{\beta}^i$ has a number of advantages. It has low bias for β^i. It is faster to compute than Bayesian methods, which require converged MCMC chains that could be infeasible to obtain on large-scale data. By bootstrapping over independent loci's dosages, one can obtain a confidence interval for β^i.

A Bayesian approach has a number of advantages. Due to the incorporation of a likelihood and priors, a Bayesian approach could have lower MSE than an MME, and one can obtain estimates of $\theta^i, i \in \{1, 2\}$ and θ_B, rather than the compound parameter $\beta^i = (\theta^i - \theta_B)/(1 - \theta_B)$. With posterior samples, one can also obtain a credible interval for any parameter of interest.

In many situations, it is plausible that θ_B could be greater than 0 due to migration between populations, so we suggest the use of an estimator that allows for this, which leaves the Weir and Goudet MME, Gunther and Coop (2013)'s model implemented in Bayenv, and our two-population copula model. If computation time rules out Bayesian methods, then we suggest the Weir and Goudet MME. If two populations are being analyzed, then we suggest our two-population copula model, since unlike Bayenv, it maintains Beta marginal allele frequency distributions, and had lower MSE for both population-specific F_{ST} parameters β^1 and β^2 under a variety of simulation settings.

Method	Parameter	Point Estimate	2.5% Quantile	CI 97.5% Quantile
MME	β^1	0.169	0.105	0.233
MME	β^2	0.125	0.0638	0.189
ABayeScan	θ^1	0.144	0.132	0.18
ABayeScan	θ^2	0.131	0.12	0.166
ABayeScan	θ^{12}	0.0115	0.00395	0.0359
ABayeScan	β^1	0.134	0.123	0.169
ABayeScan	β^2	0.121	0.11	0.156
BayeScan Foll	β^1	0.156	0.104	0.211
BayeScan Foll	β^2	0.113	0.0667	0.165
Bayenv	θ^1	0.0576	0.0417	0.0763
Bayenv	θ^2	0.0582	0.0425	0.0785
Bayenv	θ^{12}	-0.0287	-0.0398	-0.0204
Bayenv	β^1	0.0839	0.0641	0.108
Bayenv	β^2	0.0844	0.0649	0.111

Table 3.9: Application of Weir and Goudet MME, our Adapted BayeScan copula model, BayeScan, and Bayenv to HealthABC Cohort Study Data, $L = 100$ bi-allelic Chr 22 SNP's typed in Black and White individuals, treating Population 1 as Black and Population 2 as White. **MME**: Weir and Goudet method of moments estimator implemented with hierfstat R package. **BayeScan Foll**: Foll and Gaggiotti (2006)'s model implemented with BayeScan software. **ABayeScan**: Our Adapted BayeScan copula model implemented with Stan. **Bayenv**: Gunther and Coop (2013)'s model implemented with Bayenv software. **2.5% Quantile, 97.5% Quantile**: Endpoints of 95% confidence interval (for MME) or credible interval (for a Bayesian method). $\beta^{\mathbf{i}}, \mathbf{i} \in \{1, 2\}$: Population i-specific F_{ST} $\beta^i := (\theta^i - \theta^{12})/(1 - \theta^{12})$. $\theta^1, \theta^2, \theta^{12}$: Population-1, Population-2, and Population-12 ibd probabilities respectively.

3.8 Appendix:

3.8.1 $\hat{\beta}_{jj'_{\text{Sim.}}}$'s Pedigree-Determined Parameter

The match statistic between individuals j and j' at locus l is defined

$$\tilde{M}_{jj'_l} \;:=\; \frac{1}{2}[1 + (X_{j_l} - 1)(X_{j'_l} - 1)].$$

The simple MME of pedigree-determined parameter $\beta_{jj'}$ is

$$\begin{aligned}
\hat{\beta}_{jj'_{\text{Sim.}}} &= 1 - \frac{\sum_l (1 - \tilde{M}_{jj'_l})}{\sum_l 2\tilde{p}_l(1 - \tilde{p}_l)} \\
&= \frac{\sum_l 2\tilde{p}_l(1 - \tilde{p}_l)}{\sum_l 2\tilde{p}_l(1 - \tilde{p}_l)} - \frac{\sum_l (1 - \tilde{M}_{jj'_l})}{\sum_l 2\tilde{p}_l(1 - \tilde{p}_l)} \\
&= \frac{\sum_l [(2\tilde{p}_l(1 - \tilde{p}_l)) - (1 - \tilde{M}_{jj'_l})]}{\sum_l 2\tilde{p}_l(1 - \tilde{p}_l)} \\
&= \frac{\sum_l [2\tilde{p}_l - 2\tilde{p}_l^2 - 1 + \tilde{M}_{jj'_l}]}{\sum_l 2\tilde{p}_l(1 - \tilde{p}_l)} \\
&= \frac{\sum_l [\tilde{M}_{jj'_l} - \tilde{p}_l^2 - (-2\tilde{p}_l + \tilde{p}_l^2 + 1)]}{\sum_l 2\tilde{p}_l(1 - \tilde{p}_l)} \\
&= \frac{\sum_l [\tilde{M}_{jj'_l} - \tilde{p}_l^2 - (1 - \tilde{p}_l)^2]}{\sum_l 2\tilde{p}_l(1 - \tilde{p}_l)}.
\end{aligned}$$

The ratio of the numerator to the denominator of the simple MME is

$$\frac{\sum_l [E\Big(\tilde{M}_{jj',l} - \tilde{p}_l^2 - (1-\tilde{p}_l)^2\Big)]}{\sum_l [E\Big(2\tilde{p}_l(1-\tilde{p}_l)\Big)]} = \frac{\sum_l [E\Big(\frac{1}{2}[1 + (X_{j_l}-1)(X_{j'_l}-1)] - \tilde{p}_l^2 - (1-\tilde{p}_l)^2\Big)]}{\sum_l [E\Big(2\tilde{p}_l(1-\tilde{p}_l)\Big)]}$$
$$= \frac{\sum_l [E\Big(\frac{1}{2}[1 + (X_{j_l}X_{j'_l}) - X_{j'_l} - X_{j_l} + 1] - \tilde{p}_l^2 - (1 - 2\tilde{p}_l + \tilde{p}_l^2)\Big)]}{\sum_l [E\Big(2\tilde{p}_l(1-\tilde{p}_l)\Big)]}$$
$$= \frac{\sum_l [E\Big(\frac{1}{2}[2 + (X_{j_l}X_{j'_l}) - X_{j'_l} - X_{j_l}] - 1 + 2\tilde{p}_l - 2\tilde{p}_l^2\Big)]}{\sum_l [E\Big(2\tilde{p}_l(1-\tilde{p}_l)\Big)]}$$
$$= \frac{\sum_l [\frac{1}{2}[2 + E(X_{j_l}X_{j'_l}) - E(X_{j'_l}) - E(X_{j_l})] - 1 + 2E(\tilde{p}_l) - 2E(\tilde{p}_l^2)]}{\sum_l [2E\Big(\tilde{p}_l(1-\tilde{p}_l)\Big)]}. \tag{3.14}$$

Denoting the numerator of the fraction in Equation 3.14 $\mathcal{E}(\hat{\beta}_{jj'_{\text{Sim.}}})_{Num}$, and substituting expressions for $\mathcal{E}(X_{jl})$ of Equation 2.19, $\mathcal{E}(X_{jl}X_{j'l})$ of Equation 2.21, $\mathcal{E}(\tilde{p}_l^2)$ of Equation 2.18, and $\mathcal{E}(\tilde{p}_l)$ of Equation 2.17, we simplify $\mathcal{E}(\hat{\beta}_{jj'_{\text{Sim.}}})_{Num}$:

$$\mathcal{E}(\hat{\beta}_{jj'_{\text{Sim.}}})_{Num} = \sum_l \left[\frac{1}{2}[2 + E(X_{j_l}X_{j'_l}) - E(X_{j'_l}) - E(X_{j_l})] - 1 + 2E(\tilde{p}_l) - 2E(\tilde{p}_l^2)\right]$$
$$= \sum_l \Bigg[\frac{1}{2}[2 + \Big(4(\pi_l^2 + \pi_l(1-\pi_l)\theta_{jj'})\Big) - \Big(2\pi_l\Big) - \Big(2\pi_l\Big)]$$
$$- 1 + 2\Big(\pi_l\Big) - 2\Big(\pi_l^2 + \pi_l(1-\pi_l)\left[\theta_S + \frac{1}{2n}(1 + F_W - 2\theta_S)\right]\Big)\Bigg]$$
$$= \sum_l \Bigg[[1 + \Big(2(\pi_l^2 + \pi_l(1-\pi_l)\theta_{jj'})\Big) - 2\Big(\pi_l\Big)]$$
$$- 1 + 2\pi_l - 2\pi_l^2 - 2\pi_l(1-\pi_l)\left[\theta_S + \frac{1}{2n}(1 + F_W - 2\theta_S)\right]\Bigg]$$
$$= \sum_l \left[\Big(2\pi_l^2 + 2\pi_l(1-\pi_l)\theta_{jj'}\Big) - 2\pi_l^2 - 2\pi_l(1-\pi_l)\left[\theta_S + \frac{1}{2n}(1 + F_W - 2\theta_S)\right]\right]$$
$$= \sum_l \left[2\pi_l(1-\pi_l)\theta_{jj'} - 2\pi_l(1-\pi_l)\left[\theta_S + \frac{1}{2n}(1 + F_W - 2\theta_S)\right]\right]$$
$$= \sum_l \left[\Big(2\pi_l(1-\pi_l)\Big)\Big(\theta_{jj'} - \left[\theta_S + \frac{1}{2n}(1 + F_W - 2\theta_S)\right]\Big)\right].$$

Denoting the denominator of the fraction in Equation 3.14 by $\mathcal{E}(\hat{\beta}_{jj'_{\text{Sim.}}})_{Denom}$, and substituting in the expression for $\mathcal{E}[\tilde{p}_l(1-\tilde{p}_l)]$ (obtained from Equations 2.17 and 2.18) into the denominator, we obtain:

$$\mathcal{E}(\hat{\beta}_{jj'_{\text{Sim.}}})_{Denom} = \sum_l [2\Big(\pi_l(1-\pi_l)\left[(1-\theta_S) - \frac{1}{2n}\left(1+F_W-2\theta_S\right)\right]\Big)].$$

We then obtain

$$\begin{aligned}
\mathcal{E}(\hat{\beta}_{jj'_{\text{Sim.}}})_{Num}/\mathcal{E}(\hat{\beta}_{jj'_{\text{Sim.}}})_{Denom} &= \sum_l \left[\Big(2\pi_l(1-\pi_l)\Big)\Big(\theta_{jj'} - \left[\theta_S + \frac{1}{2n}\left(1+F_W-2\theta_S\right)\right]\Big)\right] \\
&\quad / \sum_l [2\Big(\pi_l(1-\pi_l)\left[(1-\theta_S) - \frac{1}{2n}\left(1+F_W-2\theta_S\right)\right]\Big)] \\
&= \left[\theta_{jj'} - \left[\theta_S + \frac{1}{2n}\left(1+F_W-2\theta_S\right)\right]\right] \\
&\quad / \left[(1-\theta_S) - \frac{1}{2n}\left(1+F_W-2\theta_S\right)\right] \\
&= \frac{(\theta_{jj'}-\theta_S) - \frac{1}{2n}\left(1+F_W-2\theta_S\right)}{(1-\theta_S) - \frac{1}{2n}\left(1+F_W-2\theta_S\right)}.
\end{aligned} \tag{3.15}$$

When $n \to \infty$, the last expression on the right-hand side of Equations 3.15 approaches $\hat{\beta}_{jj'_{\text{Sim.}}}$'s pedigree-determined parameter $\beta_{jj'} = (\theta_{jj'}-\theta_S)/(1-\theta_S)$.

3.8.2 $\hat{\beta}_{jj',w}$'s Pedigree-Determined Parameter

The ratio of the expected value of $\hat{\beta}_{jj',w}$'s numerator to the expected value of its denominator is $\sum_l [E(\tilde{M}_{jj'_l}) - E(\tilde{M}_{S_l})]/\sum_l [1 - E(\tilde{M}_{S_l})]$.

Using expectations for match statistics, e.g. Equation 2.24,

$$\sum_l [E(\tilde{M}_{jj'_l}) - E(\tilde{M}_{S_l})] / \sum_l [1 - E(\tilde{M}_{S_l})] = \frac{\sum_l [\Big(1 - (1 - \theta^{ii'}_{jj'}) 2\pi_l (1 - \pi_l)\Big) - \Big(1 - (1 - \theta_S) 2\pi_l (1 - \pi_l)\Big)]}{\sum_l [1 - \Big(1 - (1 - \theta_S) 2\pi_l (1 - \pi_l)\Big)]}$$

$$= \frac{\sum_l [(-1 + \theta^{ii'}_{jj'}) 2\pi_l (1 - \pi_l) + (1 - \theta_S) 2\pi_l (1 - \pi_l)]}{\sum_l [(1 - \theta_S) 2\pi_l (1 - \pi_l)]}$$

$$= \frac{\sum_l [(\theta^{ii'}_{jj'} - \theta_S) 2\pi_l (1 - \pi_l)]}{\sum_l [(1 - \theta_S) 2\pi_l (1 - \pi_l)]}$$

$$= \frac{\theta^{ii'}_{jj'} - \theta_S}{1 - \theta_S}.$$

Since j and j' are from the same population, $i = i'$, and we drop the ii' superscript. Hence,

$$\sum_l [E(\tilde{M}_{jj'_l}) - E(\tilde{M}_{S_l})] / \sum_l [1 - E(\tilde{M}_{S_l})] = \frac{\theta_{jj'} - \theta_S}{1 - \theta_S}, \tag{3.16}$$

which is $\hat{\beta}_{jj',w}$'s pedigree-determined parameter $\beta_{jj'}$.

3.8.3 $\hat{\theta}_{jj',GCTA_w}$ and $\hat{\theta}_{jj',GCTA_u}$'s Pedigree-Determined Parameter

The ratio of the expected value of the numerator of $\hat{\theta}_{jj'_{\text{GCTAw}}}$ to the expected value of its denominator is $[\sum_l \mathcal{E}((X_{j_l} - 2\tilde{p}_l)(X_{j'_l} - 2\tilde{p}_l))] / [\sum_l \mathcal{E}(4\tilde{p}_l(1 - \tilde{p}_l))]$.

The expectation in the numerator is

$$\begin{aligned} \mathcal{E}((X_{j_l} - 2\tilde{p}_l)(X_{j'_l} - 2\tilde{p}_l)) &= \mathcal{E}(X_{j_l} X_{j'_l} - 2\tilde{p}_l X_{j'_l} - 2X_{j_l}\tilde{p}_l + 4\tilde{p}_l^2) \\ &= \mathcal{E}(X_{j_l} X_{j'_l}) - 2\mathcal{E}(\tilde{p}_l X_{j'_l}) - 2\mathcal{E}(X_{j_l}\tilde{p}_l) + 4\mathcal{E}(\tilde{p}_l^2). \end{aligned}$$

Substituting in expressions for $\mathcal{E}(X_{j_l} X_{j'_l})$ from Equation 2.21, $\mathcal{E}(\tilde{p}_l X_{j'_l})$ and $\mathcal{E}(X_{j_l}\tilde{p}_l)$ from Equation 2.22, and $\mathcal{E}(\tilde{p}_l^2)$ from Equation 2.18, we obtain that the expectation in the numerator is

$$
\begin{aligned}
\mathcal{E}((X_{j_l} - 2\tilde{p}_l)(X_{j'_l} - 2\tilde{p}_l)) &= \mathcal{E}(X_{j_l} X_{j'_l}) - 2\mathcal{E}(X_{j'_l}\tilde{p}_l) - 2\mathcal{E}(X_{j_l}\tilde{p}_l) + 4\mathcal{E}(\tilde{p}_l^2) \\
&= \Big[4\Big(\pi_l^2 + \pi_l(1-\pi_l)\theta_{jj'}\Big)\Big] \\
&- 2\Big[2\pi_l^2 + 2\pi_l(1-\pi_l)\psi_{j'} + \frac{1}{n}\pi_l(1-\pi_l)(1 + F_{j'} - 2\psi_{j'})\Big] \\
&- 2\Big[2\pi_l^2 + 2\pi_l(1-\pi_l)\psi_{j} + \frac{1}{n}\pi_l(1-\pi_l)(1 + F_{j} - 2\psi_{j})\Big] \\
&+ 4\Big[\pi_l^2 + \pi_l(1-\pi_l)\left[\theta_S + \frac{1}{2n}\left(1 + F_W - 2\theta_S\right)\right]\Big] \\
&= \Big(2\pi_l(1-\pi_l)\Big)\Big(\Big[2\theta_{jj'}\Big] \\
&+ \Big[-2\psi_{j'} - \frac{1}{n}(1 + F_{j'} - 2\psi_{j'})\Big] \\
&+ \Big[-2\psi_{j} - \frac{1}{n}(1 + F_{j} - 2\psi_{j})\Big] \\
&+ \Big[2\left[\theta_S + \frac{1}{2n}\left(1 + F_W - 2\theta_S\right)\right]\Big]\Big). \qquad (3.17)
\end{aligned}
$$

Substituting into the denominator the expression for $\mathcal{E}(\tilde{p}_l(1-\tilde{p}_l))$, we have:

$$
\begin{aligned}
\mathcal{E}(\hat{\theta}_{jj'_{\text{GCTAw}}}) = \Big[\sum_l \Big(& 2\pi_l(1-\pi_l)\Big[2\theta_{jj'} - 2\psi_{j'} - \frac{1}{n}(1 + F_{j'} - 2\psi_{j'}) \\
& - 2\psi_j - \frac{1}{n}(1 + F_j - 2\psi_j) \\
& + 2\left[\theta_S + \frac{1}{2n}\left(1 + F_W - 2\theta_S\right)\right]\Big]\Big)\Big] \\
& / \Big[\sum_l 4\pi_l(1-\pi_l)\left[(1-\theta_S) - \frac{1}{2n}\left(1 + F_W - 2\theta_S\right)\right]\Big] \\
& = \Big[2\theta_{jj'} + -2\psi_{j'} - \frac{1}{n}(1 + F_{j'} - 2\psi_{j'}) - 2\psi_j - \frac{1}{n}(1 + F_j - 2\psi_j) \\
& + 2\left[\theta_S + \frac{1}{2n}\left(1 + F_W - 2\theta_S\right)\right]\Big] \\
& / \left[2\Big((1-\theta_S) - \frac{1}{2n}\left(1 + F_W - 2\theta_S\right)\Big)\right] \\
& = \Big[2(\theta_{jj'} + \left[\theta_S + \frac{1}{2n}\left(1 + F_W - 2\theta_S\right)\right]) \\
& - 2\psi_{j'} - \frac{1}{n}(1 + F_{j'} - 2\psi_{j'}) \\
& - 2\psi_j - \frac{1}{n}(1 + F_j - 2\psi_j)\Big] \\
& / \left[2\Big((1-\theta_S) - \frac{1}{2n}\left(1 + F_W - 2\theta_S\right)\Big)\right].
\end{aligned} \tag{3.18}
$$

This expression on the right-hand-side of Equation 3.18 with $n \to \infty$ gives what we call $\hat{\theta}_{jj'_{\text{GCTAw}}}$'s pedigree-determined parameter.

Isolating factors of $1/n$, we obtain:

$$
\begin{aligned}
\mathcal{E}(\hat{\theta}_{jj'_{\text{GCTAw}}}) = &\Big[2\theta_{jj'} + 2\theta_S + \frac{1}{n}\left(1 + F_W - 2\theta_S\right) \\
&- 2\psi_{j'} - \frac{1}{n}(1 + F_{j'} - 2\psi_{j'}) \\
&- 2\psi_j - \frac{1}{n}(1 + F_j - 2\psi_j)\Big] \\
&/ \left[2\Big((1-\theta_S) - \frac{1}{2n}\left(1 + F_W - 2\theta_S\right)\Big)\right] \\
&= \Big[2\theta_{jj'} + 2\theta_S - 2\psi_{j'} - 2\psi_j \\
&+ \frac{1}{n}\left(1 + F_W - 2\theta_S\right) \\
&- \frac{1}{n}(1 + F_{j'} - 2\psi_{j'}) - \frac{1}{n}(1 + F_j - 2\psi_j)\Big] \\
&/ \left[2\Big((1-\theta_S) - \frac{1}{2n}\left(1 + F_W - 2\theta_S\right)\Big)\right] \\
&= \Big[2(\theta_{jj'} + \theta_S - \psi_{j'} - \psi_j) \\
&+ \frac{1}{n}\Big(\left(1 + F_W - 2\theta_S\right) - (1 + F_{j'} - 2\psi_{j'}) - (1 + F_j - 2\psi_j)\Big)\Big] \\
&/ \left[2\Big((1-\theta_S) - \frac{1}{2n}\left(1 + F_W - 2\theta_S\right)\Big)\right]. \qquad (3.19)
\end{aligned}
$$

We can show that this expression is equivalent to Equation (13) of the Weir and Goudet paper:

$$
\begin{aligned}
\mathcal{E}(\hat{\theta}_{jj'_{\mathrm{GCTAw}}}) = &\Big[(\theta_{jj'} + \theta_S - \psi_{j'} - \psi_j) \\
&+ \frac{1}{2n}\Big(\,(1 + F_W - 2\theta_S) - (1 + F_{j'} - 2\psi_{j'}) - (1 + F_j - 2\psi_j)\Big)\Big] \\
&/\Big((1 - \theta_S) - \frac{1}{2n}\,(1 + F_W - 2\theta_S)\,\Big) \\
= &\Big[(\theta_{jj'} + \theta_S - \psi_{j'} - \psi_j) \\
&+ \frac{1}{2n}\Big(F_W - 2\theta_S - F_{j'} + 2\psi_{j'} - 1 - F_j + 2\psi_j\Big)\Big] \\
&/\Big((1 - \theta_S) - \frac{1}{2n}\,(1 + F_W - 2\theta_S)\,\Big) \\
= &\Big[(\theta_{jj'} + \theta_S - \psi_{j'} - \psi_j) \\
&- \frac{1}{2n}\Big(- F_W + 2\theta_S + F_{j'} - 2\psi_{j'} + 1 + F_j - 2\psi_j\Big)\Big] \\
&/\Big((1 - \theta_S) - \frac{1}{2n}\,(1 + F_W - 2\theta_S)\,\Big) \\
= &\Big[(\theta_{jj'} + \theta_S - \psi_{j'} - \psi_j) \\
&- \frac{1}{n}\Big(- F_W/2 + \theta_S + F_{j'}/2 - \psi_{j'} + 1/2 + F_j/2 - \psi_j\Big)\Big] \\
&/\Big((1 - \theta_S) - \frac{1}{2n}\,(1 + F_W - 2\theta_S)\,\Big) \\
= &\Big[(\theta_{jj'} + \theta_S - \psi_{j'} - \psi_j) \\
&- \frac{1}{n}\Big(- 1/2 - F_W/2 + \theta_S + F_{j'}/2 - \psi_{j'} + 1 + F_j/2 - \psi_j\Big)\Big] \\
&/\Big((1 - \theta_S) - \frac{1}{2n}\,(1 + F_W - 2\theta_S)\,\Big) \\
= &\Big[(\theta_{jj'} + \theta_S - \psi_{j'} - \psi_j) \\
&- \frac{1}{n}\Big(- F_S + \theta_S + F_{j'}/2 - \psi_{j'} + 1 + F_j/2 - \psi_j\Big)\Big] \\
&/\Big((1 - \theta_S) - \frac{1}{n}\,(1/2 + F_W/2 - \theta_S)\,\Big),
\end{aligned}
$$

where $F_S = (1 + F_W)/2$. Considering further that $\theta_{jj} = 1/2 + F_j/2$ and $\theta_{j'j'} = 1/2 + F_{j'}/2$, we have that

$$
\begin{aligned}
\mathcal{E}(\hat{\theta}_{jj'_{\mathrm{GCTAw}}}) = &\Big[(\theta_{jj'} + \theta_S - \psi_{j'} - \psi_j) \\
&- \frac{1}{n}\Big(1/2 + F_j/2 + 1/2 + F_{j'}/2 - \psi_j - \psi_{j'} - F_S + \theta_S\Big)\Big] \\
&/\Big((1-\theta_S) - \frac{1}{n}\left(F_S - \theta_S\right)\Big) \\
= &\Big[(\theta_{jj'} - \psi_j - \psi_{j'} + \theta_S) \\
&- \frac{1}{n}\Big(\theta_{jj} + \theta_{j'j'} - \psi_j - \psi_{j'} - F_S + \theta_S\Big)\Big] \\
&/\Big((1-\theta_S) - \frac{1}{n}\left(F_S - \theta_S\right)\Big),
\end{aligned}
$$

and this last expression is equivalent to Equation 13 of Weir and Goudet (2017).

Consider the ratio of the expectation of a single locus $\hat{\theta}_{jj'_{\mathrm{GCTAu}}}$'s numerator to the expectation of a single locus $\hat{\theta}_{jj'_{\mathrm{GCTAu}}}$'s denominator. Take the mean over loci of this ratio. The resultant quantity is $\frac{1}{L}\sum_{l=1}^{L}[\mathcal{E}((X_{j_l} - 2\tilde{p}_l)(X_{j'_l} - 2\tilde{p}_l))]/[\mathcal{E}(4\tilde{p}_l(1-\tilde{p}_l))]$.

Recalling Equation 3.17, we observe that $E((X_{jl} - 2\tilde{p}_l)(X_{j'l} - 2\tilde{p}_l))/(2\pi_l(1-\pi_l))$ does not depend on l:

$$
\begin{aligned}
\mathcal{E}((X_{j_l} - 2\tilde{p}_l)(X_{j'_l} - 2\tilde{p}_l))/\Big(2\pi_l(1-\pi_l)\Big) = &\Big[2\theta_{jj'}\Big] \\
&+ \Big[-2\psi_{j'} - \frac{1}{n}(1 + F_{j'} - 2\psi_{j'})\Big] \\
&+ \Big[-2\psi_j - \frac{1}{n}(1 + F_j - 2\psi_j)\Big] \\
&+ \Big[2\left[\theta_S + \frac{1}{2n}\left(1 + F_W - 2\theta_S\right)\right]\Big].
\end{aligned}
$$

We observe that $\mathcal{E}(\tilde{p}_l(1-\tilde{p}_l))/\Big(\pi_l(1-\pi_l)\Big)$ does not depend on l:

$$
\mathcal{E}(\tilde{p}_l(1-\tilde{p}_l))/\Big(\pi_l(1-\pi_l)\Big) = \left[(1-\theta_S) - \frac{1}{2n}\left(1 + F_W - 2\theta_S\right)\right].
$$

Hence, we have:

$$\begin{aligned}
\frac{1}{L}\sum_{l=1}^{L}\frac{\mathcal{E}((X_{j_l}-2\tilde{p}_l)(X_{j'_l}-2\tilde{p}_l))}{\mathcal{E}(4\tilde{p}_l(1-\tilde{p}_l))} &= \frac{1}{L}\sum_{l=1}^{L}\frac{\Big(2\pi_l(1-\pi_l)\Big)}{\Big(2\pi_l(1-\pi_l)\Big)}\frac{\mathcal{E}((X_{j_l}-2\tilde{p}_l)(X_{j'_l}-2\tilde{p}_l))/\Big(2\pi_l(1-\pi_l)\Big)}{\mathcal{E}(4\tilde{p}_l(1-\tilde{p}_l))/\Big(2\pi_l(1-\pi_l)\Big)} \\
&= \frac{\mathcal{E}((X_{j_l}-2\tilde{p}_l)(X_{j'_l}-2\tilde{p}_l))/\Big(2\pi_l(1-\pi_l)\Big)}{\mathcal{E}(4\tilde{p}_l(1-\tilde{p}_l))/\Big(2\pi_l(1-\pi_l)\Big)}\frac{1}{L}\sum_{l=1}^{L}\frac{\Big(2\pi_l(1-\pi_l)\Big)}{\Big(2\pi_l(1-\pi_l)\Big)} \\
&= \Big[2(\theta_{jj'}+\Big[\theta_S+\frac{1}{2n}\left(1+F_W-2\theta_S\right)\Big]) \\
&- 2\psi_{j'}-\frac{1}{n}(1+F_{j'}-2\psi_{j'}) \\
&- 2\psi_j-\frac{1}{n}(1+F_j-2\psi_j)\Big] \\
&/\left[2\Big((1-\theta_S)-\frac{1}{2n}\left(1+F_W-2\theta_S\right)\Big)\right]. \qquad (3.20)
\end{aligned}$$

This expression on the right-hand-side of Equation 3.20 with $n \to \infty$ gives what we call $\hat{\theta}_{jj'_{\mathrm{GCTAu}}}$'s pedigree-determined parameter. We observe that the expression on the right-hand-side of Equation 3.20 is the same expression as that on the right-hand-side of Equation 3.18. Hence, $\hat{\theta}_{jj'_{\mathrm{GCTAu}}}$'s pedigree-determined parameter is the same as $\hat{\theta}_{jj'_{\mathrm{GCTAw}}}$'s pedigree-determined parameter.

BIBLIOGRAPHY

David H Alexander, John Novembre, and Kenneth Lange. Fast model-based estimation of ancestry in unrelated individuals. *Genome Research*, 19(9):1655–1664, 2009.

Amy D Anderson and Bruce S Weir. A maximum-likelihood method for the estimation of pairwise relatedness in structured populations. *Genetics*, 176(1):421—440, May 2007.

David J Balding. Likelihood-based inference for genetic correlation coefficients. *Theoretical population biology*, 63(3):221 –230, May 2003.

David J Balding and Richard A Nichols. A method for quantifying differentiation between populations at multi-allelic loci and its implications for investigating identity and paternity. In *Human Identification: The Use of DNA Markers*, pages 3–12. Springer, 1995.

Mark A Beaumont and David J Balding. Identifying adaptive genetic divergence among populations from genome scans. *Molecular ecology*, 13(4):969–980, 2004.

Stephen Boyd and Lieven Vandenberghe. *Convex optimization.* Cambridge University Press, Cambridge, UK ; New York, 2004. ISBN 9780511804441.

Stephen P Brooks and Andrew Gelman. General methods for monitoring convergence of iterative simulations. *Journal of computational and graphical statistics*, 7(4):434–455, 1998.

Sharon R. Browning, Kelsey Grinde, Anna Plantinga, Stephanie M. Gogarten, Adrienne M. Stilp, Robert C. Kaplan, M. Larissa Avilés-Santa, Brian L. Browning, and Cathy C. Laurie. Local ancestry inference in a large US-based Hispanic/Latino study: Hispanic Community Health Study/Study of Latinos (HCHS/SOL). *G3 Genes Genomes Genetics*, 6(6):1525—1534, 2016.

John Buckleton, James Curran, Jérôme Goudet, Duncan Taylor, Alexandre Thiery, and BS Weir. Population-specific F_{ST} values for forensic STR markers: A worldwide survey. *Forensic Science International: Genetics*, 23:91–100, 2016.

Kristin M Burkart, Tamar Sofer, Stephanie J London, Ani Manichaikul, Fernando P Hartwig, Qi Yan, Marí-a Soler Artigas, Lydiana Avila, Wei Chen, Sonia Davis Thomas, Alejandro A Diaz, Ian P Hall, Bernardo L Horta, Robert C Kaplan, Cathy C Laurie, Ana M Menezes, Jean V Morrison, Elizabeth C Oelsner, Deepa Rastogi, Stephen S Rich, Manuel Soto-Quiros, Adrienne M Stilp, Martin D Tobin, Louise V Wain, Juan C CeledÃn, and R Graham Barr. A genome-wide association study in Hispanics/Latinos identifies novel signals for lung function. The Hispanic Community Health Study/Study of Latinos. *American journal of respiratory and critical care medicine*, 198(2):208–219, February 2018.

George Casella and Roger L Berger. *Statistical Inference*, volume 2. Duxbury Pacific Grove, CA, 2002.

C. C. Cockerham and B. S. Weir. Estimation of inbreeding parameters in stratified populations. *Annals of Human Genetics*, 50(3):271–281, July 1986.

C Clark Cockerham. Analyses of gene frequencies. *Genetics*, 74(4):679–700, 1973.

Susan Jane Colley. Vector calculus, 2012.

Matt Conomos. *Inferring, Estimating, and Accounting for Population and Pedigree Structure in Genetics Analyses*. PhD thesis, University of Washington, 2014.

Matthew P Conomos, Alexander P Reiner, Bruce S Weir, and Timothy A Thornton. Model-free estimation of recent genetic relatedness. *The American Journal of Human Genetics*, 98(1):127–148, 2016.

Miklós Csűrös. Non-identifiability of identity coefficients at biallelic loci. *Theoretical population biology*, 92:22–29, 2014.

Kenneth Dodds. Resampling methods in genetics and the effect of family structure in genetic data, January 1986. URL `http://search.proquest.com/docview/303426395/`.

Bradley Efron and David V Hinkley. Assessing the accuracy of the maximum likelihood estimator: Observed versus expected Fisher information. *Biometrika*, 65(3):457–483, 1978.

D. S. (Douglas Scott) Falconer. Introduction to quantitative genetics, 1996.

Daniel Falush, Matthew Stephens, and Jonathan K Pritchard. Inference of population structure using multilocus genotype data: linked loci and correlated allele frequencies. *Genetics*, 164(4):1567–1587, 2003.

Ziding Feng and Charles E. McCulloch. Statistical inference using maximum likelihood estimation and the generalized likelihood ratio when the true parameter is on the boundary of the parameter space. *Statistics Probability Letters*, 13:325–332, 1992.

Matthieu Foll and Oscar Gaggiotti. Identifying the environmental factors that determine the genetic structure of populations. *Genetics*, 174(2):875–891, 2006.

Matthieu Foll and Oscar Gaggiotti. A genome-scan method to identify selected loci appropriate for both dominant and codominant markers: a Bayesian perspective. *Genetics*, 180 (2):977–993, 2008.

Matthieu Foll, Martin C Fischer, Gerald Heckel, and Laurent Excoffier. Estimating population structure from AFLP amplification intensity. *Molecular ecology*, 19(21):4638–4647, 2010.

Tom R Gaunt, Santiago Rodríguez, and Ian NM Day. Cubic exact solutions for the estimation of pairwise haplotype frequencies: implications for linkage disequilibrium analyses and a web tool 'CubeX'. *BMC Bioinformatics*, 8(1):428–428, 2007.

Jérôme Goudet. Hierfstat, a package for r to compute and test hierarchical F-statistics. *Molecular ecology resources*, 5(1):184–186, 2005.

Simon Gravel, Fouad Zakharia, Andres Moreno-Estrada, Jake K Byrnes, Marina Muzzio, Juan L Rodriguez-Flores, Eimear E Kenny, Christopher R Gignoux, Brian K Maples, Wilfried Guiblet, et al. Reconstructing native american migrations from whole-genome and whole-exome data. *PLoS Genet*, 9(12):e1004023, 2013.

Torsten Gunther and Graham Coop. Robust identification of local adaptation from allele frequencies. *Genetics*, 195(1):205–220, September 2013.

Nathan Hall, Laina Mercer, Daisy Phillips, Jonathan Shaw, and Amy D. Anderson. Maximum likelihood estimation of individual inbreeding coefficients and null allele frequencies. *Genetics Research*, 94(3):151–161, 2012.

Eimeer Kenny. Local ancestry inference for 1000 genomes project phase I admixed samples. `ftp://ftp-trace.ncbi.nih.gov/1000genomes/ftp/phase1/analysis_results/ancestry_deconvolution/README_20120604_phase1_ancestry_deconvolution`, August 2017.

Kathleen F. Kerr, Christy L. Avery, Henry J. Lin, Laura M. Raffield, Qian S. Zhang, Brian L. Browning, Sharon R. Browning, Matthew P. Conomos, Stephanie M. Gogarten, Cathy C. Laurie, Tamar Sofer, Timothy A. Thornton, Chancellor Hohensee, Rebecca D. Jackson, Charles Kooperberg, Yun Li, Raúl Méndez-Giráldez, Marco V. Perez, Ulrike Peters, Alexander P. Reiner, Zhu-Ming Zhang, Jie Yao, Nona Sotoodehnia, Kent D. Taylor, Xiuqing Guo, Leslie A. Lange, Elsayed Z. Soliman, James G. Wilson, Jerome I. Rotter, Susan R. Heckbert, Deepti Jain, and Eric A. Whitsel. Genome-wide association study of heart rate and its variability in Hispanic/Latino cohorts. *Heart Rhythm*, 14(11):1675–1684, November 2017.

Xianyun Mao, Abigail W. Bigham, Rui Mei, Gerardo Gutierrez, Ken M. Weiss, Tom D. Brutsaert, Fabiola Leon-Velarde, Lorna G. Moore, Enrique Vargas, Paul M. McKeigue, Mark D. Shriver, and Esteban J. Parra. A genomewide admixture mapping panel for

Hispanic/Latino populations. *American Journal of Human Genetics*, 80(6):1171–1178, June 2007.

Brian K Maples, Simon Gravel, Eimear E Kenny, and Carlos D Bustamante. RFMix: a discriminative modeling approach for rapid and robust local-ancestry inference. *The American Journal of Human Genetics*, 93(2):278–288, 2013.

Alicia R. Martin, Christopher R. Gignoux, Raymond K. Walters, Genevieve L. Wojcik, Benjamin M. Neale, Simon Gravel, Mark J. Daly, Carlos D. Bustamante, and Eimear E. Kenny. Human demographic history impacts genetic risk prediction across diverse populations. *The American Journal of Human Genetics*, 100(4):635–649, April 2017.

Takeo Maruyama. Effective number of alleles in a subdivided population. *Theoretical population biology*, 1(3):273–306, 1970.

Brook G Milligan. Maximum-likelihood estimation of relatedness. *Genetics*, 163(3):1153–1167, 2003.

George Nicholson, Albert V Smith, Frosti Jónsson, Ómar Gústafsson, Kári Stefánsson, and Peter Donnelly. Assessing population differentiation and isolation from single-nucleotide polymorphism data. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 64(4):695–715, 2002.

Michael Pitt, David Chan, and Robert Kohn. Efficient bayesian inference for gaussian copula regression models. *Biometrika*, 93(3):537–554, 2006.

Alkes L Price, Arti Tandon, Nick Patterson, Kathleen C Barnes, Nicholas Rafaels, Ingo Ruczinski, Terri H Beaty, Rasika Mathias, David Reich, and Simon Myers. Sensitive detection of chromosomal segments of distinct ancestry in admixed populations (sensitive ancestry segment detection). *PLoS Genetics*, 5(6):e1000519, June 2009.

Jonathan K Pritchard, Matthew Stephens, and Peter Donnelly. Inference of population structure using multilocus genotype data. *Genetics*, 155(2):945–959, 2000.

Shaun Purcell, Benjamin Neale, Kathe Todd-Brown, Lori Thomas, Manuel AR Ferreira, David Bender, Julian Maller, Pamela Sklar, Paul IW De Bakker, Mark J Daly, et al. Plink: a tool set for whole-genome association and population-based linkage analyses. *The American Journal of Human Genetics*, 81(3):559–575, 2007.

Kl Raphael, Ra Murphy, MG Shlipak, S Satterfield, Hk Huston, A Sebastian, De Sellmeyer, Kv Patel, AB Newman, Mj Sarnak, Jh Ix, and Lf Fried. Bicarbonate concentration, acid-base status, and mortality in the health, aging, and body composition study. *Clinical Journal Of The American Society Of Nephrology*, 11(2):308–316, February 2016.

Kermit Ritland. Estimators for pairwise relatedness and individual inbreeding coefficients. *Genetics Research*, 67(2):175–185, 1996.

Scientific Working Group on DNA Analysis Methods. Interpretation guidelines for y-chromosomes STR typing, 2014.

Steven G. Self and Kung-Yee Liang. Asymptotic properties of maximum likelihood estimators and likelihood ratio tests under nonstandard conditions. *Journal of the American Statistical Association*, 82(398):605–610, June 1987.

The 1000 Genomes Project Consortium. An integrated map of genetic variation from 1,092 human genomes. *Nature*, 491(7422):56–65, October 2012.

E. A. Thompson. The estimation of pairwise relationships. *Annals of Human Genetics*, 39 (2):173–188, October 1975.

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

T. Thornton, Q. Zhang, X. Cai, C. Ober, and M. S. McPeek. XM: association testing on the X-chromosome in case-control samples with related individuals. *Genetic Epidemiology*, 36 (5):438–450, July 2012.

Mei-Ling Ting Lee. Properties and applications of the sarmanov family of bivariate distributions. *Communications in Statistics-Theory and Methods*, 25(6):1207–1222, 1996.

A. W. van der Vaart. Asymptotic statistics, 1998.

Jon Wakefield. *Bayesian and frequentist regression methods*. Springer series in statistics. Springer, New York, NY, 2013. ISBN 9781441909251.

Larry Wasserman. *All of nonparametric statistics*. Springer texts in statistics. Springer, New York; London, 2006. ISBN 9780387306230.

B. S. (Bruce S.) Weir. Genetic data analysis II: methods for discrete population genetic data, 1996.

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

Bruce S Weir and William G Hill. Estimating f-statistics. *Annual review of genetics*, 36(1): 721–750, 2002.

Bruce S Weir, Lon R Cardon, Amy D Anderson, Dahlia M Nielsen, and William G Hill. Measures of human population structure show heterogeneity among genomic regions. *Genome research*, 15(11):1468–1476, 2005.

BS Weir and C. Clark Cockerham. Estimation of linkage disequilibrium in randomly mating populations. *Heredity*, 42(1):105–111, 1979.

Jon Wellner. Chapter 4 Efficient likelihood estimation and related tests. URL `https://www.stat.washington.edu/jaw/COURSES/580s/581/LECTNOTES/ch4-rev4.pdf[Accessed:August5,2018]`.

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

Sewall Wright. Evolution in mendelian populations. *Genetics*, 16(2):97–159, 1931.

C. F. Jeff Wu. On the convergence properties of the EM algorithm. *The Annals of Statistics*, 11(1):95–103, 1983.

Jian Yang, S. Hong Lee, Michael E. Goddard, and Peter M. Visscher. Gcta: A tool for genome-wide complex trait analysis. *The American Journal of Human Genetics*, 88(1):76 – 82, 2011.

Loic Yengo, Zhihong Zhu, Naomi R Wray, Bruce S Weir, Jian Yang, Matthew R Robinson, and Peter M Visscher. Detection and quantification of inbreeding depression for complex traits from snp data. *Proceedings of the National Academy of Sciences of the United States of America*, 114(32):8602–8607, August 2017.

Qian S Zhang, Brian L Browning, and Sharon R Browning. Genome-wide haplotypic testing in a finnish cohort identifies a novel association with low-density lipoprotein cholesterol. *European Journal of Human Genetics*, 23(5):672–677, 2015.

Qian S Zhang, Brian L Browning, and Sharon R Browning. ASAFE: ancestry-specific allele frequency estimation. *Bioinformatics*, 32(14):2227–2229, 2016.