

## ABSTRACT

JOHNSON, ANDREA MICHELLE. Estimation and Sampling Properties of Gene Diversity, Heterozygosity and  $F_{ST}$ . (Advisor: Bruce S. Weir)

Estimates of the coancestry coefficient  $F_{ST}$ , gene diversity and heterozygosity have been used in many fields, including conservation and evolutionary biology, and forensic studies. Although the sampling properties of estimators of these parameters could affect inferences to be made, these continue to be frequently overlooked in published analyses. This dissertation characterizes the estimators of these measures by presenting relevant theoretical developments, approaches to estimation, and results regarding evaluations of the sampling properties of these three measures. Making inferences about the genetic variation among populations of a species, rather than some larger, between-species scope, will be the biological focus.

The accuracy and precision of the method of moments and maximum likelihood estimators of population-specific  $F_{ST}$  developed by Weir and Hill (2002) are evaluated through population simulation and analysis of an empirical data set. Of the two estimators considered, the method of moments estimator for population-specific  $F_{ST}$  is found to be relatively unbiased with a large sampling variance, which increases as coancestry increases in a population. Sampling more loci has a much stronger effect on reducing this sampling variance than sampling more individuals. The other estimator evaluated here obtained by maximum-likelihood poorly estimates the coancestry in a population for two iterative approaches and a non-iterative approach, and is not recommended for future analyses. Problems with estimates obtained from individual loci with very low polymorphism levels for

both estimators are discussed and practical measures for proceeding with analyses are suggested.

Properties of several methods for inferring the variances of sample heterozygosity or gene diversity are evaluated, including the use of a new random model for the total variance of sample heterozygosity. Large differences with a previous mixed model are observed for a case where there is a large variance component due to loci. Several approximations are evaluated and compared to variances obtained from exact expressions. Different results with unbalanced data for the total variance of sample heterozygosity are obtained with four variance component methods, as expected by statistical theory. The likelihood-based methods considered here are shown to be robust to violations of assumptions of normality, even for very small sample sizes.

ESTIMATION AND SAMPLING PROPERTIES OF GENE  
DIVERSITY, HETEROZYGOSITY AND  $F_{ST}$

by

Andrea Michelle Johnson

A dissertation submitted to the Graduate Faculty of  
North Carolina State University  
in partial fulfillment of the  
requirements for the Degree of  
Doctor of Philosophy

GENETICS

Raleigh

2004

APPROVED BY:

---

---

Co-Chair of Advisory Committee

Co-Chair of Advisory Committee

## DEDICATION

To my parents, Dr. Douglas and Dr. Margaret Johnson.

Your example has been my guide for the entire way.

Your support has meant that I reached my destination.

## BIOGRAPHY

Andrea Johnson was born on June 14, 1974 in Mesa, Arizona and was raised in Tempe, Arizona. Her interests included reading every book she could get her hands on, the writing and production of short videos, drawing, horseback riding and hiking in the mountains of northern Arizona.

After graduation from McClintock High School in 1992, Andrea attended the University of California, San Diego and majored in Molecular Biology. While in San Diego, Andrea focused on class work, gaining experience working in a molecular laboratory setting and making movies. During this time she spent a year as a research apprentice at Scripps Institute of Oceanography working on the biochemistry of a type of anaerobic clam with Dr. Horst Felbeck and a year at the Salk Institute working on a yeast model of neurological protein interactions with Dr. Damon Getman.

Andrea next obtained a Master's degree in Genetics from the University of California, Davis working with Dr. Juan Medrano. Her thesis work included touring northern Californian dairies collecting milk from different breeds of dairy cattle in order to extract mRNA from mammary gland cells and resulted in determining a SNP associated with the different amounts of saturated fat in the milk of different cattle breeds. During her studies at UC Davis, Andrea also had the unusual experience for a city girl of being a teaching assistant for Animal Science 101, where the final exam required that she grade students on their abilities in milking cattle, among other tasks.

Through recommendations of the faculty at UC Davis, Andrea decided to apply

and was accepted into the doctoral program of the Genetics department at North Carolina State University. At this point in her career she decided to refocus her energies towards working on the theoretical aspects of genetics and after choosing as her advisor Dr. Bruce Weir, she has not looked back. For the future, Andrea has accepted a position as a Biostatistician with Roche Molecular Diagnostics in Alameda, California and is looking forward to returning to northern California and entering the world of industrial research.

## ACKNOWLEDGEMENTS

This work was produced with the help of the consistently insightful advice and large amounts of patience of my advisor, Bruce Weir. I would like to thank my committee Bill Atchley, Jeff Thorne and Michael Purugganan. Thanks also to Ken Olsen, for supplying me with a data set that had linkage disequilibrium, as well as John Nason and Serge Planes for kindly responding to my request for their previously published data sets. I am also grateful for financial support from an NIEHS training fellowship.

I would like to thank everyone at the Bioinformatics Research Center for creating an excellent place for pursuing graduate studies on a day to day basis, particularly with the support of Jb Briseno and Debbie Hibbard. Appreciation for their friendship and computing help goes out to Doug Robinson, who helped me think out some difficult problems in translating abstractions of genes in populations into code, Frank Mannino, for always being ready to help with R or Perl syntax, and Errol Strain, for providing timely computer support and steady encouragement to be focused on graduation. Finally, the support, interest and love regularly shown to me by my parents Doug and Peggy Johnson, and sister Stacy, has been invaluable in pursuing my degree.

# Contents

<b>LIST OF TABLES</b>	<b>ix</b>
<b>LIST OF FIGURES</b>	<b>xi</b>
<b>1 REVIEW</b>	<b>1</b>
Introduction . . . . .	2
Parameterizations of $F_{ST}$ . . . . .	7
Foundations: Models and Measures With Allele Frequency Data . . . . .	8
Extensions: New Types of Molecular Information . . . . .	10
Methods for Obtaining Estimators of Gene Diversity, Heterozygosity and $\theta$	16
Obtaining Estimators of $\theta$ . . . . .	17
Dominant Data . . . . .	25
Estimators of Gene Diversity . . . . .	26
Sampling Properties of Estimators of Population Structure . . . . .	28
Analytical Approximations and Resampling Approaches . . . . .	28
Bias . . . . .	31
Variance . . . . .	33
Confidence Interval Estimators . . . . .	35

Hypothesis Testing . . . . .	36
Sampling Properties of Estimators of Gene Diversity and Heterozygosity	37
Literature Cited . . . . .	40
<b>2 SAMPLING PROPERTIES OF POPULATION-SPECIFIC <math>F_{ST}</math></b>	
<b>ESTIMATORS</b>	<b>48</b>
Abstract . . . . .	49
Introduction . . . . .	50
Theory . . . . .	55
Methods . . . . .	59
Results . . . . .	62
Evaluation of Estimator Properties by Simulation . . . . .	62
Effect of Low Polymorphism on the Method of Moments Estimator	67
Discussion . . . . .	76
Literature Cited . . . . .	80
Appendix 1 . . . . .	83
Appendix 2 . . . . .	89
<b>3 THE VARIANCE OF SAMPLE HETEROZYGOSITY AND GENE</b>	
<b>DIVERSITY</b>	<b>92</b>
Abstract . . . . .	93
Introduction . . . . .	94
Theory . . . . .	99
Point Estimators of Gene Diversity and Heterozygosity . . . . .	99
Variances, Models and Scope of Inference . . . . .	100

Variance Component Methods . . . . .	106
Materials and Methods . . . . .	108
Results . . . . .	110
Discussion . . . . .	121
Literature Cited . . . . .	124
Appendix . . . . .	127

# List of Tables

2.1	Effect of different amounts of coancestry and sample sizes on the MOME and MLE of $\hat{\theta}_i$ . . . . .	63
2.2	Mean bias for $\hat{\theta}_i$ obtained from population simulations . . . . .	66
2.3	Estimates and standard deviations of different approaches to combining information about $\theta$ over loci and populations for six simulated data sets . . . . .	68
2.4	Effects of low polymorphism at individual loci on the method of moments estimator of $\hat{\theta}_{il}$ . . . . .	70
2.5	Effects of incorporating different polymorphism criteria into estimation of the population-specific method of moment estimator of $\hat{\theta}_i$ for the Planes data set . . . . .	73
2.6	Pairwise estimates of $\theta$ for the Planes data set . . . . .	74
2.7	Estimates of $\hat{\theta}_{il}$ from a subset of loci with very low polymorphism . . . . .	91
2.8	Table 2.7 Continued . . . . .	91

3.1	Relationships between variances of sample gene diversity and heterozygosity for data without significant composite linkage disequilibrium . . . . .	111
3.2	Relationships between variances of sample gene diversity and heterozygosity for data with significant composite linkage disequilibrium	114
3.3	Comparison of within-population standard deviations of sample gene diversity and heterozygosity of simulations to their true values . . .	116
3.4	Example of the relationships between the total standard deviation of sample heterozygosity obtained with two linear models and four estimation methods for relatively balanced data . . . . .	118
3.5	Example of the relationships between the total standard deviation of sample heterozygosity obtained with two linear models and four estimation methods for relatively balanced data with minimal sample sizes . . . . .	119
3.6	Example of the relationships between the total standard deviation of sample heterozygosity obtained with two linear models and four estimation methods for unbalanced data . . . . .	120
3.7	Comparison of the total standard deviation of sample gene diversity approximated as a function of descent measures to the standard deviation of the sample gene diversities at individual loci . . . . .	129

# List of Figures

2.1	Empirically derived sampling distributions for MOME and MLE of $\hat{\theta}_i$	64
2.2	Values of maximum likelihood estimates of $\hat{p}_u$ and $\hat{\phi}_i$ for each iteration of a converged replicate of simulated data . . . . .	85
2.3	Values of maximum likelihood estimates of $\hat{p}_u$ and $\hat{\phi}_i$ for each iteration of a replicate of simulated data where the estimates exceeded the possible parameter range of the parameters they are estimating	86
2.4	Maximum likelihood estimates of $\hat{p}_u$ and $\hat{\phi}_i$ for a case of simulated data with periodicity . . . . .	87
2.5	Change in the log likelihood for a converged replicate of simulated data and a nonconverged replicate that showed periodicity . . . . .	88

# Chapter 1

## REVIEW

# Introduction

Frequently published in a wide range of journals, estimates of  $F_{ST}$  or  $\theta$ , gene diversity and heterozygosity have been used for many different purposes including conservation biology, evolutionary biology and forensic profiling studies. Although the sampling properties of estimators of these parameters could affect inferences to be made, these properties continue to be frequently overlooked in analyses of this type. This dissertation characterizes the sampling properties of these three measures of genetic variation. To accomplish this, summarization of the theoretical developments, approaches to estimation, and results regarding the sampling properties of these three measures composes the statistical focus of this review. Making inferences about variation among populations of a species, rather than some larger, between-species scope, will be the biological focus.

When estimates of  $F_{ST}$ , gene diversity and heterozygosity are combined with an understanding of the statistical uncertainty involved in this type of analysis, a full picture of the genetic variation of a species or population can be obtained. Inferences made with the parameters discussed here can be used to characterize the genetic variation of populations, both in amount with gene diversity and heterozygosity, and in distribution, as reflected by the  $F$ -statistics. The distribution of genetic variation can be characterized as the relative amount of alleles shared and exchanged between a group of populations in a species. The extent of reproduction between the members of different populations is termed the level of differentiation, or structure, of a group of populations. The populations of most species are structured to some extent (Balloux and Lugon-Moulin, 2002). These

authors postulate that isolation through distance could occur frequently between populations of a species as a result of the physical space occupied by a species being greater than the distance a single individual can migrate.

Structure is frequently modelled as associations between alleles and these can occur on different levels, and to different extents, such as within individuals, between individuals within the same population, and between individuals in different populations. In a hierarchical population model there are typically three  $F$ -statistics described;  $F_{ST}$ ,  $F_{IT}$  and  $F_{IS}$ . These can be defined as the correlations between genes sampled from different levels in the population hierarchy due to shared ancestry. Their subscripts refer to the levels they are concerned with, where  $I$  stands for individuals,  $S$  stands for subpopulations and  $T$  for the total population.  $F_{IT}$  is called the fixation index because it measures the progress of neutral loci towards fixation in a finite-sized population for a single allele under the influence of random genetic drift in simple population models.  $F_{ST}$  is frequently referred to as the coancestry coefficient, and measures the degree of relationship, or extent of common ancestry, between individuals within populations, relative to the amount of coancestry found in the total population.

Cockerham introduced new notation in his work by renaming the  $F$ -statistics  $\theta$ ,  $F$  and  $f$ , respectively, from the list above. While the  $I, S, T$  subscripts allow a quick intuitive understanding of the place of an  $F$ -statistic in the population parameter hierarchy, Cockerham's notation will be used for the remainder of this work. This is done to reduce confusion between  $F$ -statistics and the  $F$  distribution, and to emphasize that these measures are population parameters, rather than statistics that are functions of observed data.

Inbreeding within populations occurs when individuals are more related to each other than the relatedness of a random set of individuals from the total population. Two factors contribute to the total amount of inbreeding in a set of populations and these can be generally thought of as contributing to  $\theta$  or  $f$  values, respectively. Both random genetic drift causing differentiation between populations in the group, and any assortative mating occurring within populations, can increase the amount of inbreeding of the whole group of populations. This whole value can be expressed as  $F$ , the inbreeding coefficient, where

$$F = f + \theta(1 - f).$$

This relationship demonstrates that  $F$  is a sum of the variation due to genes that are alike in individuals, summarized in  $f$ , and the variation due to unrelated genes in a population, the remaining product term. Generally,  $f$  tends to reflect more the immediate history of the population, while  $\theta$  reflects the long-term effects of demographic and evolutionary forces on the group of populations as a whole (Cockerham, 1973). For this reason, this discussion focuses on  $\theta$  as being of greater interest to investigators.

Research regarding making inferences about population structure has led to the development of different parameterizations of  $\theta$ , multiple approaches to obtaining estimators of these quantities and work at characterizing the sampling properties of these estimators. There are many different parameterizations of  $\theta$ , each based on the combination of different population and mutation models, measures of relatedness or similarity and the molecular characteristics of different types of data assumed in the development of that parameterization. Several estimators for  $\theta$  have

been developed with both frequentist approaches, such as the method of moments (MOM) and maximum likelihood (ML) methods, and Bayesian approaches. These statistical methods for obtaining estimators differ in the assumptions made about the underlying distribution of allele frequencies, and in the sampling properties of their estimators. To determine the variance, bias and confidence interval estimators for the various estimators of  $\theta$ , both analytical and resampling approaches have been applied. The large number of demographic and evolutionary forces acting on alleles in populations tend to make sampling variances of estimates of  $\theta$  quite large. Estimators can be distinguished by their approach to the minimization of this variance. Due to the complexity of the sampling distributions of  $\theta$  estimators, no single estimator appears to have the best sampling properties for the entire parameter range of  $\theta$ .

The other measures we consider here are the less complex parameters of gene diversity and heterozygosity. Heterozygosity is simply the proportion of individuals with a heterozygous genotype in a population at a single locus, often termed observed heterozygosity or  $H_o$ . For the case of species with some degree of selfing, the presence of many different types of homozygous genotypes would not be captured by the frequency of heterozygotes, so that the term heterozygosity can be inadequate to describe the amount of genetic variation in a population. The measure of gene diversity, proposed by Nei (1973), has been frequently and incorrectly termed average heterozygosity, or  $H_e$  in the literature due to the fact that it is the expected frequency of heterozygote in a population with Hardy-Weinberg proportions. This can be misleading terminology in that gene diversity is not the same as the theoretical expectation of heterozygosity, which might be written as

$E(H_i)$ .

As a measure of genetic variation, Nei's gene diversity should be particularly used for selfing species. However, gene diversity will also have the same value as the expected frequency of heterozygosity in a random mating population not undergoing selfing. The relationship between gene diversity, heterozygosity and  $\theta$  can be expressed exactly for certain specific population and mutation models but may be more complicated in real life. Exact expressions for the variance of sample gene diversity and heterozygosity were determined by Weir (1989) and Weir et al. (1990).

## Parameterizations of $F_{ST}$

The form of a parameterization of  $F_{ST}$  is determined in an interconnected way by selection of population and mutation models, type of data and the measure of relatedness or similarity between different alleles. Different combinations of choices of these factors have resulted in the development of a number of parameterizations of  $F_{ST}$ , also sometimes termed analogues. Some parameterizations are more suited to theoretical studies of stochastic processes while others lead more naturally to estimation.

The population and mutation models of a parameterization determine the dynamics of similarity and relatedness of alleles while the data type assumed dictates the extent similarity and relatedness can be distinguished or detected. The choice of population model determines how relatedness or dependencies due to common ancestry between pairs of alleles arise through drift and mating system and diminish through gene flow by specifying components such as mating system, amount of gene flow between populations and both number and size of populations. The mutation model deals instead with similarity, describing both probabilities of loss through changes occurring between ancestrally related alleles and of gain of similarity between unrelated alleles. Data type determines the type of evolutionary information available about relatedness between alleles and can affect the choice of specific mutation model due to the variety of mutational processes at different types of marker loci.

Choosing between different measures of relatedness should be done in order to utilize the distinct types of information available about population differentiation

specific to each molecular marker. Particular data types may be better suited to different measures of relatedness due to the nature of the evolutionary information they contain, whether it be the number of repeats of a microsatellite allele, or the allelic type of an allozyme. Additionally, the type of research to be pursued may dictate the choice of measure, since some, such as identity by descent, can facilitate development in a theoretical framework, while others may be better suited to estimation-based studies.

## **Foundations: Models and Measures With Allele Frequency Data**

Initial studies of population structure defined and determined the effects of different population and mutation models and measures of relatedness on inferences. Focus later shifted to considering the impacts of different types of molecular data on parameterizations as new types of molecular markers became available. The models first considered involved pedigrees of individuals (Wright, 1921, 1931) and these were generalized to groups of individuals with regular mating systems by (Cockerham, 1967). All individuals at a particular level in a population hierarchy would then have the same theoretical expectation of the  $F$ -statistics. Research about population structure not concerned primarily with estimation became increasingly centered on finding theoretical results in different population models using stochastic processes approaches. Ohta and Kimura (1969), Nagylaki (1975), and Weir and Cockerham (1984), among others, dealt particularly with effects such as unequal mutation or migration rates, linkage disequilibrium and the relationship

between selection and geography.

Interest in studies of this type has remained over a number of years because of the number of possible combinations of the four components of a parameterization, in addition to the desire to model other complex aspects of biological reality. In a recent example of this, Rousset (1999) looked at applying the results of inferring population structure based on methods with simple drift-only models to models with more complicated sources of population structure. He gave a large sample approximation showing how to transform an estimate from a simple pure-drift model into one consistent with results from a more complex model such as a model with separate sexes or non-discrete generations. By determining exact results for the mean and covariance of allele frequencies given a finite-island model with migration and mutation, Fu et al. (2003) found that populations exchanging alleles have strongly correlated allele frequencies. In essence, these results imply genetic drift is acting on the level of the total population.

Correlations of pairs of alleles (Wright, 1951; Cockerham, 1969), identity by descent (Malécot, 1948) and variance components for linear models of allele frequencies (Cockerham, 1969) are the measure paradigms developed during the initial period of work on population structure. Cockerham (1973) synthesized these paradigms, showing explicitly the equivalence in parameterizations obtained from each of these measures. Although the identity by descent parameterization is the most flexible approach to determining transition equations of interest, identity cannot be directly observed, requiring identity in state formulations of the biological parameter of interest to be used in practice (Cockerham and Weir, 1987, 1993). However, for the sorts of short evolutionary distances where mutation and

migration can be assumed to be negligible, the identity in state and identity by descent probabilities will be equivalent, given a pure drift model. The correlational definitions of the  $F$ -statistics accommodate work with selection, migration and other evolutionary forces since they can be used in a more general sense than the probabilities of identity parameters (Cockerham, 1973).

## **Extensions: New Types of Molecular Information**

The next major emphasis in the body of work on population structure resulted from the introduction of new molecular markers including RFLPs, microsatellites and DNA sequence data. New types of data contained different sorts of information about differentiation of populations and so inspired more parameterization development. Issues associated with the new data types included the need to account for the effect of different underlying mutational processes and rates, the desire to capture additional information about evolutionary relationships between alleles, and a growing interest in determining the relationship between adaptive variation associated with QTL and neutral genetic variation.

Within this period, the development of new measure paradigms for parameterizations included genetic distances, adaptive variation and coalescent times. All of the parameterizations incorporating these measures relied on the partitioning of the variation of their particular measure into between and within population components and constructed a function analogous to the coancestry coefficient from these components, as developed previously for allele frequency measures. Problems with applying frequency-based parameterizations to sequence data brought into use

new measure paradigms of different genetic distances. This type of approach had the unappealing characteristic that longer sequences, potentially containing more information about their evolutionary relationships, gave estimates with inferior properties relative to those based on shorter sequences (Hudson, 2000).

### **Nucleotide diversity**

By quantifying the mean number of substitutions per site between pairs of haplotypes and treating sites as individual loci, parameterizations based on diversity measures can be used to partition nucleotide level variation into between and within population components. The application of frequency-based measures to haplotype data can create problems due to the characteristics typical of haplotype data, and these include the inability to increase the accuracy of estimates by increasing the length of the sequences sampled (Hudson et al., 1992). Diversity parameterizations better summarize the type of information contained in sequence data, accounting for long sequences of tightly linked sites and relatively smaller sample sizes. Nucleotide diversity-based structure parameterizations are likely to differ from frequency-based parameterizations if haplotypes or alleles within a group of populations have a wide range in the number of differing sites between pairs of sequences (Lynch and Crease, 1990).

Parameters of this type include  $\gamma_{ST}$  (Nei, 1982) for the fixed population scope of inference, and  $N_{ST}$  (Lynch and Crease, 1990) and  $\langle F_{ST} \rangle$  (Hudson et al., 1992) for random populations. All three of these parameterizations are appropriate for use with analyses of RFLP or DNA sequence data. The two random scope parameters  $N_{ST}$  and  $\langle F_{ST} \rangle$  differ in the mutation models used in their development.  $\langle F_{ST} \rangle$

assumes an infinite sites model of mutation and  $N_{ST}$  assumes the Jukes-Cantor substitution model. In practice, the effect of the difference between these assumptions should be small for within-species variation (Lynch and Crease, 1990), and use of the infinite-sites model simplifies the calculation of estimates.

For the case of RFLP data, Lynch and Crease (1990) showed that the variance of nucleotide diversity due to populations and nucleotides is not negligible. Estimates from fixed scopes not accounting for this variation are likely to be quite different than random population estimators of population structure. This result can be qualified by the fact that estimates of population structure made from sequence data are likely to have a minimized sampling variance due to the large increase in the number of nucleotides sampled per sequence relative to sites for RFLPs (Lynch and Crease, 1990). Additionally, Lynch and Crease concluded that increasing the number of loci, rather than the number of individuals sampled, is likely to be more effective at reducing sampling variation of the  $N_{ST}$  estimator due to statistical sampling.

### **The coalescent**

Slatkin (1991) developed an alternative coalescent parameterization of  $F_{ST}$  based on the relationship between the probability of identity by descent and the probability of no mutation before coalescence for pairs of alleles in structured populations. This relies on the coalescent times being assumed to be proportional to the probability of identity in an evolutionary model. Where the probability of identity by descent can be assumed to be a linear function of time, differentiation is a function of drift only. In a subsequent paper, Slatkin (1995) determined another related

parameterization termed  $R_{ST}$ , involving the coalescent and assuming the stepwise mutation model and microsatellite data in the form of numbers of repeats per allele.  $R_{ST}$  is determined by a similar relationship between the variances of size differences in the number of repeats between alleles, for a given level of a population hierarchy, and the probabilities of identity in state of those alleles. The complementary relationship between the probabilities of identity by descent and the coalescent developed in (Slatkin, 1991) was used to justify the assumption that coalescence times are proportional to variances of differences in allele sizes. The parameterization of Slatkin made in the context of a fixed population scope of inference, was complemented by the equivalent term  $\rho_{ST}$  for random populations given by Rousset (1996) a year later.

For the probabilities of identity to differ from the probabilities of coalescence before mutation, at least two mutation events must occur within the time to coalescence (Slatkin, 1995; Rousset, 1996). Leading to the hypothesis that this rate of mutation would be the point where the probability of identity by descent would no longer be linear with time and inferences made from estimates of repeat-based measures, such as  $R_{ST}$ , would differ from frequency-based inferences. As a result, although coalescent-based approaches for estimating gene diversity and population structure can be powerful for types of tightly-linked marker sets, their use for describing and inferring population differentiation for shorter times scales, such as for groups within a species, should be avoided.

### **Adaptive variation**

Inspiration behind attempts to quantify the differentiation between populations

due to the genetic variation associated with local adaptation has sprung from breeding programs wishing to avoid the problem of outbreeding depression where populations are maladapted to their parental environments. By partitioning the genetic variation associated with a quantitative trait, the same ratios of within and between populations variance components used for  $F_{ST}$  could be used to create a parameterization of the term called  $Q_{ST}$  (Spitze, 1993). Evidence for selection acting on a particular QTL could be inferred by testing the hypothesis that  $Q_{ST}$  of a particular trait is not significantly different from  $F_{ST}$  for neutral loci at the species level (Rogers, 1986). Empirical results reviewed by McKay and Latta (2002) showed poor correlations between  $Q_{ST}$  and  $F_{ST}$ , suggesting instead a difference between the processes shaping genetic variation at QTL and neutral molecular markers. They conclude  $F_{ST}$  is a more appropriate parameter for making inferences about differentiation at QTL in random mating populations than  $Q_{ST}$ .

## Synthesis

Marker driven parameterizations and parameterizations based on identities, correlations or variance components, were synthesized into a common framework called analysis of molecular variance, or AMOVA by (Excoffier et al., 1992; Michalakis and Excoffier, 1996). The name of this procedure refers to the close relationship to ANOVA method of moments estimation as given by Weir and Cockerham (1984). The basic approach is to first reduce the data to a vector of indicator variables using one of the various measures of phylogenetic distance, geographic distances or allele types. Estimation then proceeds with the linear models approach of Weir

and Cockerham (1984) to infer population structure. Despite the similarity of AMOVA analyses with different measures, it should be realized inferences based on alternative parameterizations of  $F_{ST}$  can differ because the underlying molecular processes generating genetic variation can differ for each type of molecular marker (Charlesworth, 1998).

# Methods for Obtaining Estimators of Gene Diversity, Heterozygosity and $\theta$

In this section the statistical methods used to obtain estimators of gene diversity, heterozygosity and  $\theta$  parameterized as a function of correlations of allele frequencies are reviewed. To obtain estimators of these three parameters, the frequentist approaches of method of moments and maximum likelihood have been used, while Bayesian methods have been used only to obtain estimates of  $\theta$ . An advantage of frequentist approaches lies in the reduced computational intensity relative to the intensity required with Bayesian methods. In contrast, Bayesian approaches have the benefit of systematic incorporation of prior information about the data into an analysis and potentially increasing the ability to capture important information within data sets about parameters with complex distributions. To date a rigorous and independent study of the actual performance of frequentist estimators relative to Bayesian estimators remains to be completed. Results should be fairly similar for the case of inferring structure from dominant markers for either method, unless a significant amount of prior information is available to be included in the analysis (Hill and Weir, 2004).

Frequentist approaches take their name from a focus on determining the probability or frequency of an event, given the assumption of a large number of hypothetical, identical and independent replicate populations from which statistical expectations of quantities of interest are determined. Alternatively, Bayesian theory relies on the use of Bayes rule to update the choices of the prior probability distribution of parameters in the model with any information available in the data

into what is called the posterior distribution of these parameters. The posterior is obtained generally through Markov Chain Monte Carlo (MCMC) simulation. Currently, the use of several hierarchical models with a Bayesian approach have been described and a newly developed likelihood-free approximation methods to approximate posterior densities may potentially be of use in obtaining estimates of  $\theta$ .

The distribution used by both maximum likelihood and Bayesian hierarchical approaches to model allele frequencies is crucial because it determines the properties of the estimators developed from it (Weir and Hill, 2002). The method of moments and likelihood-free MCMC approaches may be able to better capture the complexity of the biological reality generating population allele frequencies since they do not require the simplifying assumption of a specific form of a distribution. However, estimators obtained through methods assuming the distribution of allele frequencies, such as maximum likelihood and hierarchical Bayes models, have the advantage of better understood statistical properties such as efficiency and minimized variances.

## **Obtaining Estimators of $\theta$**

### **Method of moments estimators**

Estimators of  $\theta$  obtained through a method of moments approach include the bivariate estimators of Weir and Cockerham (1984) and Robertson and Hill (1984) and the multivariate estimator of Long (1986). Construction of the bivariate estimators proceeds through combining estimates from individual alleles linearly over

all alleles and loci, while the multivariate estimator is combined only over loci. See Yang (1998) for a generalization of the Weir and Cockerham estimator to an arbitrary number of levels in a population hierarchy. All three of these estimators are functions of variance components of a linear model of allele frequencies. Long's estimator is equivalent to the Robertson-Hill and Weir-Cockerham estimators for bi-allelic data from a single locus. For multi-locus, multi-allelic data, the Long estimator differs from the bivariate estimators by using all the information in the data simultaneously to determine the value of  $\theta$ . However, none of these methods account for the linkage disequilibrium between loci when combining information over loci. At this time the sampling properties of Long's estimator remain to be determined.

The best possible way to combine bivariate estimators over alleles has remained an issue in this area of research stemming from the complicated nature of the sampling distribution of these statistics. The approach of Weir and Cockerham was to combine estimates by taking the ratio of the sum of the numerators of each estimator to the sum of the denominators of each estimator. Alternatively, Robertson and Hill combined the single allele estimates by taking the weighted average of the ratios over all alleles. The different approaches of (Weir and Cockerham, 1984) and (Robertson and Hill, 1984) to weighting multiple alleles and loci have been shown to minimize the variance of the estimator for different ranges of the true underlying parameter  $\theta$ . The variance of estimates from populations with a high level of differentiation will be minimized for the Weir and Cockerham estimator while the Robertson and Hill approach has a minimized variance for populations with low to medium differentiation (Raufaste and Bonhomme, 2000). Raufaste

and Bonhomme thus recommended the use of the different estimators be governed by the level of differentiation found in a given study.

### **Population-specific estimates**

For simplicity, the parameterization of most classical estimators of  $\theta$  has assumed that populations have the same size as each other over successive generations, so that all populations have the same  $\theta$  value. In practice, an estimate of  $\theta$  may effectively be considered to be an estimate of the average  $\theta$  of a group of populations. Nicholson and Donnelly (2002) and Weir and Hill (2002) and Balding (2003) have increased the number of parameters in their models to allow for population-specific values of  $\theta$ . The enormous amount of markers available for many SNP projects increase the potential degrees of freedom hugely and make this parameterization possible. As is usual when a model is made more general, it can better reflect the biological complexities, and thus better estimates and inferences should be the result.

Weir and Hill (2002) relaxed the assumption of equal population sizes, defined population-specific  $\theta$ , and determined estimators for these quantities using both method of moments and maximum likelihood approaches for multi-allelic, multi-locus data. Additionally, for future analysis, a more general formula for a moment estimator of  $\theta$  allowing for unbalanced samples was given in Eqn. 9 of that paper. Weir and Hill (2002) gave method of moments estimators for both the coancestry of each population and for the coancestry between pairs of populations, which is essentially the extent of shared relatedness between a pair of populations. They obtained as well a maximum likelihood estimator of population-specific  $\theta$ , based

on a multivariate normal approximation of sample allele frequencies. The quality of the population-specific estimators of Weir and Hill improves with an increase in the number of alleles per locus, and this was also a characteristic of the Bayesian population-specific  $\theta$  developed later by (Balding, 2003).

### **Bayesian estimators**

An increasing number of investigators have been applying Bayesian methods to the problem of inferring population structure in the past few years as a result of the rising availability in affordable computing power. Although frequently computationally demanding, Bayesian approaches have the advantage of the ease of methodical inclusion of any available prior information. By using knowledge about populations gained in the past, the robustness of estimates from extreme data sets sampled from the present can be increased (Lange, 1995). Bayesian methods also offer the ability to make simultaneous inferences about other quantities of interest. Examples of this would be the calculation of statistics summarizing model fit and estimating the number of distinct populations in a group of populations. The sensitivity of Bayesian estimates to the choice of prior, the performance of these estimators relative to frequentist estimators of the same parameters, and the choice between hierarchical models are concerns with the use of Bayesian approaches for making inferences about differentiation.

Current Bayesian approaches to estimation of  $\theta$ , both population-specific and as a common value over all populations, involve the assumption of hierarchical models including the forms of the prior parameter and likelihood distributions. The Dirichlet (see Balding and Nichols, 1995; Corander et al., 2003; Lange, 1995;

Rothman et al., 1974; Holsinger, 1999), and the Multivariate Normal, (see Weir and Hill, 2002) are two commonly used forms for the distribution of population allele frequencies with multiple alleles at a locus. For bi-allelic data such as SNP loci, the bivariate forms of these distributions are the Beta (see Holsinger et al., 2002; Nicholson and Donnelly, 2002) and Normal distributions (see Balding, 2003; Holsinger and Wallace, 2004; Nicholson and Donnelly, 2002; Smouse and Williams, 1982; Tufto and Hindar, 1996), are used instead. Of these studies, the work of Nicholson and Donnelly (2002) and Balding (2003) parallels the frequentist population-specific estimators of Weir and Hill (2002) using a Bayesian approach, where the estimates are the posterior mean of the conditional distributions of the parameters generated by using MCMC based rejection sampling.

Distributions of allele frequencies vary in their fit with different population models and the time since divergence of populations. The Beta is the stationary distribution of allele frequencies in the stochastic process models of the island model (Wright, 1931) and has been found to be a reasonable approximation of the stationary distribution of the finite stepping-stone model (Maruyama, 1977). The use of the normal distribution, rather than arising as the stationary distribution of a model of interest, instead has been justified by appeal to large sample theory (Weir and Hill, 2002; Nicholson and Donnelly, 2002), and should become an increasingly good approximation as sample sizes increase. The normal distribution has generally been used for non-equilibrium populations which are likely to have shorter times since divergence (Nicholson and Donnelly, 2002). For populations with weak drift and migration, the Dirichlet distribution may be a poor fit for allele frequency distributions because this increases the time to reach equilibrium

(Tufto and Hindar, 1996). Distributions in populations with high stepwise mutation rates fit the Dirichlet poorly as well (Graham et al., 2000). Additionally, making inferences about high levels of differentiation using the Dirichlet may be associated with increased uncertainty due to the long upper tail for this likelihood (Balding, 2003).

Analysis of population differentiation using Bayesian estimates raises the concerns of assessment of the sensitivity of Bayesian estimates to the choice of priors and likelihoods and the performance of these estimators relative to frequentist estimators. Some indications of sensitivity of the population-specific estimates of  $\theta$  to their hyperparameter priors were found by Nicholson and Donnelly (2002). Sensitivity to Nicholson's prior on allele frequencies was found in the interpretation by Marchini and Cardon (2002), of the results of Nicholson and Donnelly.

Regarding sensitivity of estimators to likelihood choice, method of moments estimation which relies on the form of the first two moments of the sampling distribution, should be less restrictive than Bayes estimators obtained with hierarchical models or maximum likelihood estimators requiring the assumption of a full distribution (Weir and Hill, 2002). This is especially true for the first two moments of the Beta and truncated normal distributions used by Balding (2003) and Nicholson and Donnelly (2002), which were shown to be particularly similar for intermediate allele frequencies and low values of  $\theta$  (Balding, 2003). For these distributions the frequentist approach would then be equivalent for both the equilibrium and non-equilibrium population models.

Comparisons between the sampling properties of Bayes and frequentist estimators of  $\theta$  and population-specific  $\theta$  have typically been concerned with using

data sets simulated under the population conditions assumed in their development, and so may not provide rigorous enough conditions for realistic assessment of performance. A comparison by Nicholson and Donnelly (2002) of the variances of Nicholson's Bayesian estimator and a biased method of moments estimator of population-specific  $\theta$  gave mixed results for data simulated from a normal distribution. The method of moments estimator had a smaller variance for the data sets simulated with the smallest number of populations at all levels of  $\theta$  simulated. For data sets simulated with the largest number of populations, the method of moments estimator had the smallest variance for those populations with the higher levels of differentiation.

Some recent innovations with making inferences about population structure using Bayesian approaches include new summary statistics for evaluating the contributions of model and data, the addition of simultaneous estimation of new parameters of interest, promising new likelihood-free estimation methods, and work on inferring the number of populations represented in a sample of individuals. Model fit of the Beta and Normal likelihoods has been assessed by comparing the density of the standardized residuals of allele frequencies to the standard normal distribution (Nicholson and Donnelly, 2002) and by looking at a model choice criteria (Holsinger and Wallace, 2004) that assesses the fit of the observed pattern of variation to inbreeding within populations and differentiation between populations. Holsinger and Wallace (2004) also used the difference in entropy of the posterior and prior distributions of a parameter to determine the amount of information in the data about the parameter of interest.

Corander et al. (2003) introduced the inclusion of an additional parameter for

the number of populations with different allele frequency distributions to analyses of this sort, by adding a matrix of population structure to the prior parameters and estimating from a joint Dirichlet-Multinomial posterior. In a related development, Pritchard et al. (2000) described a Bayesian clustering approach for assigning individuals to populations which simultaneously estimates the number of populations in a sample and gives associated confidence values for these estimates. While Pritchard et al. does not describe a new estimator for  $F_{ST}$ , the authors suggest that their method could be used to more precisely estimate population allele frequencies and from these the structure estimator of choice could be calculated. This might reduce problems in analyses where distinguishing between populations due to recent migration or admixture is an issue. The clustering approach has been implemented in the program STRUCTURE which is freely available to download at <http://pritch.bsd.uchicago.edu/software.html>.

Likelihood-free based rejection sampling and MCMC approaches have recently been applied to population genetics problems, as reviewed by Marjoram et al. (2003), (see also Tavaré et al., 1997; Fu and Li, 1997; Beaumont et al., 2002). Although these approaches have not yet been directly specifically to estimating  $\theta$ , these approaches potentially eliminate the need for a closed distributional form for allele frequencies and the need to deal with the nuisance parameters of the allele frequencies. The heavy reliance of these approximation approaches on population and mutational models for simulation and their still heavy computational load, which increases exponentially as the dimensionality of posterior density of the parameters increases, are issues to be considered for these approaches (Beaumont et al., 2002). Additionally, how best to assess the performance of these estima-

tors, the adequacy of the approximate posteriors generated by these likelihood-free methods, and the general small sample properties of these estimators remains to be studied.

## Dominant Data

Another area of interest for inferring population differentiation has been to extend codominant markers work and results to dominant markers, such as RAPDs and AFLPs. The short development time and low cost of dominant markers makes them particularly favored in studies of non-model organisms, and results in increased numbers of loci, populations and individuals that can be sampled. Inferring the extent of population structure with dominant marker data is complicated by the inability to distinguish heterozygous individuals and this further complicates estimation of the variance of allele frequencies from codominant data. Some knowledge about the within-populations inbreeding coefficient  $f$  is then needed, to measure of the extent of the departure from Hardy-Weinberg proportions of genotype frequencies.

In order to account for associations between alleles in individuals in inferring population structure, three different approaches have been taken in making assumptions about the value of  $f$ . The approach used by (Lynch and Milligan, 1994; Zhivotovsky, 1999) assumed estimates of  $f$  will have been obtained from previous studies. Rather than assume a value through prior results or other reasons, Hill and Weir (2004), among others, assumed  $f = 0$ . In effect, this assumes Hardy-Weinberg proportions. In a third alternative, Holsinger et al. (2002) gave a

Bayesian approach to simultaneously estimating  $f$  and  $\theta$  from the joint posterior of the allele frequencies. Although this is a more flexible approach to estimation regarding assumptions about mating systems in populations, problematically there can be little information in the data regarding  $f$ , particularly for the case of small numbers of loci and populations sampled (Holsinger et al., 2002). They concluded estimates of  $f$  are heavily affected by their priors as a consequence.

Due to the robustness and simplicity of calculation of the direct estimation of  $\hat{\theta}$  from variances of genotype frequencies of the method of moments estimator of Hill and Weir (2004), for situations where little or no prior information is available about  $f$  this estimator may be preferred, while the Bayesian estimator of Holsinger et al. should be used if any real prior information is available (Hill and Weir, 2004). Note that the method of moments estimator of Hill and Weir differs from the Weir and Cockerham (1984) estimator based on codominant data, in that it is both an iteratively obtained estimator and a nonlinear combination across genotypes and loci.

## **Estimators of Gene Diversity**

Estimators of gene diversity and heterozygosity can be distinguished by whether the effect of population structure was incorporated into their development, as is the case for the random effects models used by the estimators of (Weir, 1989; Weir et al., 1990), or populations are treated as independent, as is the case for the fixed effects models used in the development of the estimators of (Nei, 1973; Shete, 2003). Assuming genotype counts are multinomially distributed, the maximum

likelihood estimator of gene diversity within population  $i$ , summed over alleles  $u$  at locus  $l$ ,

$$\tilde{d}_{il} = 1 - \sum_{u=1}^{m_l} \tilde{p}_{ilu}^2, \quad (1.1)$$

where the  $\tilde{p}_{ilu}$  are the sample allele frequencies. The moments of sample gene diversity and heterozygosity are functions of expected genotype frequencies so they can be equivalently expressed in terms of allele frequencies and their correlation coefficients of population structure. The estimators of gene diversity and heterozygosity are used particularly to determine the extent of departure from Hardy-Weinberg proportions and thus the incorporation of the effect of population structure into the development of these estimators is particularly important.

# Sampling Properties of Estimators of Population Structure

## Analytical Approximations and Resampling Approaches

The complexity of the sampling distribution of  $\hat{\theta}$  make exact expressions for the bias and sampling variance of estimators of population structure difficult to obtain. Studies have typically relied instead on the use of resampling and analytical approaches to obtain approximate values and expressions for these quantities, as well as for confidence interval estimators of  $\theta$ . This section reviews work in determining the sampling properties of frequentist estimators of  $\theta$  and hypothesis testing for population differentiation using parametric and nonparametric approaches.

Analytical expressions for the bias and variance of the Weir and Cockerham (1984) method of moments estimator for  $\theta$  have been determined by several studies including (Reynolds, 1981; Dodds, 1986; Jiang, 1987; Li, 1996). Because of the complexity of expectations of  $\hat{\theta}$  involved in the bias and variance of this term, a number of simplifications have been made in order to obtain these analytical expressions. These simplifying strategies include working with simpler bivariate, single-locus estimators of  $\hat{\theta}$ , ratios of related quantities that are easier to find expectations of, and large sample approximations. In order to obtain an analytical expression for the bias or variance of a multi-allelic estimator of  $\theta$ , the effects of covariance between different alleles would need to be accounted for, increasing the degree of difficulty in obtaining such expressions significantly (Li, 1996).

The bivariate estimator most frequently used for determining analytical expres-

sions can be written as a simple intraclass correlation

$$\hat{\theta}^{(1)} = \frac{MSB}{\bar{p}_u(1 - \bar{p}_u)},$$

where

$$MSB = \frac{\sum_{i=1}^r n_i (\tilde{p}_{iu} - \bar{p}_u)^2}{r - 1},$$

$\tilde{p}_{iu}$  is the sample allele frequency for the  $u$ th allele in the  $i$ th population,  $\bar{p}_u$  is the average frequency of allele  $u$  across all populations,  $r$  is the number of populations, and  $n_i$  is the number of individuals sampled from population  $i$ . The bias of  $\theta^{(1)}$  has been shown to be approximately  $-\theta^2/r$  (Li, 1996; Jiang, 1987).

Raufaste and Bonhomme (2000) instead used the estimator

$$\hat{\theta}^{(2)} = \frac{MSB}{\bar{p}_u(1 - \bar{p}_u) + \frac{MSB}{r}},$$

which corrects the bias of  $\theta^{(2)}$ , but is more tractable than the most commonly used estimator, the method of moments estimator from Weir and Cockerham (1984).

This can be written as

$$\hat{\theta}^{(3)} = \frac{MSB - MSW}{MSB + (n_c - 1)MSW},$$

where

$$MSW = \frac{\sum_{i=1}^r n_i \tilde{p}_{iu} (1 - \tilde{p}_{iu})}{\sum_{i=1}^r (n_i - 1)} \quad \text{and} \quad n_c = \frac{1}{r - 1} \left( \sum_{i=1}^r n_i - \frac{\sum_{i=1}^r n_i^2}{\sum_{i=1}^r n_i} \right).$$

$\hat{\theta}^{(3)}$  is essentially unbiased in that it is a ratio of two unbiased quantities (Weir, 1996). The estimator  $\hat{\theta}^{(1)}$  is equivalent to estimator  $\hat{\theta}^{(3)}$  with the assumption of

both large sample sizes and a large number of populations so that terms of  $1/r$  and  $1/\bar{n}$  can be ignored (Weir and Cockerham, 1984).

Taylor's expansion has been used to transform expectations of ratios of tractable quantities into expectations of moments of  $\hat{\theta}$ , and then further manipulated to produce expressions for the bias and variance of  $\hat{\theta}$ . These more tractable quantities include higher order identities by descent (Reynolds, 1981) and sample allele frequencies (Jiang, 1987; Li, 1996). More simplifications of the analytical expressions based on assumptions of large numbers of populations (Reynolds, 1981) or large sample sizes for each population (Jiang, 1987; Li, 1996) were then made, to generalize predictions of the behavior of the sampling properties of  $\hat{\theta}$ . Unfortunately, simulations indicate these simplifications based on large samples do not reflect the true value of the bias and variance of  $\hat{\theta}$  as well as the original expressions (Li, 1996).

Resampling methods are an alternative to analytical expressions for determining approximate bias, variance and confidence intervals of  $\hat{\theta}$  estimates. Applications of this approach include Dodds (1986), who found that the bootstrap resampling method gives a better approximation of the bias of  $\hat{\theta}$  than the jackknife, and Weir and Hill (2002), who recommended resampling be done with a bootstrap over loci, for non-small numbers of loci, with their population-specific method of moments estimators of  $\theta$ . The recommendation of Dodds that levels of a factor should be weighted proportionately when resampling with unbalanced data, might be usefully applied to determining the sampling properties of  $\hat{\theta}$  using resampling methods.

Application of resampling methods to genetical data sets must be done with

caution to ensure that an adequate amount of loci and populations were sampled in order to obtain a reasonable approximation of the probability distribution of interest, and to preserve any population structure that might be present in the data. In consideration of these concerns, resampling should be done at the locus level across populations rather than resample individuals (Dodds, 1986; Weir, 1996). Several authors who have studied these methods including Dodds (1986), Van Dongen (1995) and Rousset and Raymond (1997), warn about the limitations of resampling methods for studies with either small numbers of populations or loci sampled due to the unpredictable effects on the probability distribution obtained by the resampling if there is a very limited number of possible different resamples. Resampling with less than twenty loci is clearly inappropriate, but how many more loci than twenty is necessary to get good results has not yet been determined stringently (Van Dongen, 1995).

## Bias

Results from bootstrap and jackknife resampling methods (Dodds, 1986) and analytical expressions (Reynolds, 1981; Jiang, 1987; Li, 1996), indicate  $\hat{\theta}^{(1)}$  is negatively biased on the order of  $\theta^2/r$ . The expression obtained by Jiang (1987) using a Taylor expansion with an approximation of moments of sample allele frequencies is

$$\text{Bias}(\hat{\theta}^{(1)}) \approx \left[ \frac{1}{n-1} + 1 \right] \left\{ \frac{2(r-2)\theta'^2(1-\theta)}{r(r-1)} - \frac{1}{p_u(1-p_u)} \left( \frac{\theta'^2}{2r} - \frac{\theta^3(r^2+4r-4)}{2r^3} \right) \right\}$$

where

$$\theta' = \frac{1}{n}((n-1)\theta + 1). \tag{1.2}$$

From this Jiang concluded that bias will be low for a moderately large number of populations sampled and alleles of intermediate frequency, and will remain relatively small for data sets with five or fewer populations, if allele frequencies are within the range of (0.1, 0.9).

Using exact moments of sample allele frequencies, instead of the approximations that Jiang used, Li found an approximation for the bias

$$\text{Bias}(\hat{\theta}^{(1)}) \approx -\frac{2(1-\theta)}{r-1} \left( \frac{1+(n-1)\theta}{n} \right)^2,$$

with the same order of magnitude as Jiang's expressions. Both approximations and large sample simplifications of Li indicate  $\hat{\theta}^{(1)}$  is negatively biased. Simulations run by Li (1996) showed that this bias becomes increasingly negative as  $\theta$  goes to one. However, bias is overall generally low, as is typical of method of moments estimators.

It has been shown by Jiang (1987), and others, that  $\hat{\theta}^{(3)}$  is essentially unbiased, in the sense that it is a ratio of unbiased expressions. However, recent results of Fu et al. (2003) appear to indicate that estimators developed from a parameterization that assumes complete reproductive isolation of populations could have very large biases, if this assumption is violated. In summary, although analytical expressions for the bias of  $\hat{\theta}^{(1)}$  differ in accuracy, because the sampling variance is generally so much greater than the bias, determination of the most accurate approximation of the bias of estimators of  $\theta$  is likely to be of less concern than minimizing the sampling variance of these estimators (Dodds, 1986).

## Variance

The very large sample variances of estimates of  $\hat{\theta}$  are due to both statistical and genetical sampling. Improving sampling design by increasing the number of loci or individuals sampled can only reduce the sample variance to some extent, due to the sampling occurring between generations (Weir and Cockerham, 1984). Nicholson and Donnelly (2002) found that increasing either the number of populations or loci sampled decreased the variance of their population-specific Bayesian estimator of  $\theta$ , but that increasing the number of loci had more of an effect after a certain point. Regardless, the issue of large sample variance of these estimators is likely to remain a central problem that cannot be satisfactorily resolved within the area of inferring the extent of population differentiation.

Jiang (1987) determined this approximation of the sample variance of  $\hat{\theta}$

$$\text{Var}(\hat{\theta}) \approx \frac{2\theta'^2}{r-1} - \frac{(8r-4)\theta'^3}{r(r-1)} + \frac{1}{p_u(1-p_u)} \left[ \frac{\theta'^3}{r} - \frac{(2r^2-2r+3)\theta'^4}{r^3} \right],$$

where this can be simplified to get an approximate variance of  $\theta'^2/r$ , where  $\theta'$  is defined as in (1.2), which agrees with others results in the magnitude of the variance. Li (1996) obtained asymptotically the simpler form of

$$\text{Var}(\hat{\theta}) \approx \frac{2(1-\theta)^2}{r-1} \left( \frac{1+(n-1)\theta}{n} \right)^2,$$

which is similar to her expression for the bias of  $\hat{\theta}$ .

Weir and Cockerham (1984) advocate the jackknife for estimating the sampling variance of  $\hat{\theta}^{(3)}$ , following the procedure first described in Reynolds et al. (1983), where the sample variance of  $\hat{\theta}$  is the variance among the set of jackknifed  $\theta$  estimates, one per locus. This approach relies on the loci sampled being nearly

completely independent, especially for a large number of loci sampled (Weir and Cockerham, 1984). Chakraborty and Danker-Hopfe (1991) found that the sampling variance of  $\hat{\theta}$  was very similar for both an expression obtained through a Taylor expansion for a fixed model and the jackknife procedure of Weir and Cockerham (1984) applied to a random model for the empirical data set they analyzed.

Different weights of  $\hat{\theta}^{(3)}$  were given by Robertson and Hill (1984) and Weir and Cockerham (1984). Each minimizes the sampling variance of the estimator for a different range of parameter values of  $\theta$  (Weir and Cockerham, 1984; Raufaste and Bonhomme, 2000). Weights were given by Robertson and Hill (1984) and can be written as

$$w_{RH} = \frac{1 - \bar{p}_u}{m - 1},$$

where  $m$  is the number of alleles at the locus, minimize the sampling variance of the estimator for low to medium levels of differentiation. The weights of Weir and Cockerham (1984)

$$w_{WC} = \frac{\bar{p}_u(1 - \bar{p}_u)}{\sum_{u' \neq u} p_{u'}(1 - p_{u'})}$$

instead minimize the sampling variance for highly differentiated populations. Although the Robertson-Hill estimator is negatively biased, Raufaste and Bonhomme (2000) found that the bias did not affect the minimized variance property of this estimator and also determined a bias correction for  $\theta$  outside the range of (0.05, 0.10). Additionally, Weir and Cockerham (1984) looked at the properties of a matrix estimator similar to the estimator given in Long (1986), which had the smallest sampling variance for all cases studied, but was only the best estimator in terms of mean squared error as a measure of overall performance for the lowest

level of differentiation studied.

## Confidence Interval Estimators

Confidence interval estimators of  $\hat{\theta}$  have been obtained through analytical approximation (Li, 1996), bootstrapping (Rousset and Raymond, 1997) and exact expressions for  $\chi^2$  distributed quadratic forms (Weir and Hill, 2002). Given lower ( $X_L$ ) and upper ( $X_U$ ) percentiles of a  $\chi_{r-1}^2$  distribution where

$$(L, U) = \left( \frac{\alpha}{2}, 1 - \frac{\alpha}{2} \right),$$

Li (1996) determined the form of an approximate confidence interval estimator for  $\hat{\theta}$  to be

$$\text{C.I.}(\hat{\theta}) = \left( \frac{(r-1)\hat{\theta}}{X_U(1-\hat{\theta}) + (r-1)\hat{\theta}}, \frac{(r-1)\hat{\theta}}{X_L(1-\hat{\theta}) + (r-1)\hat{\theta}} \right),$$

which has an advantage in the ease of evaluation of this interval for a given set of data compared to the bootstrap interval given by Dodds (1986). The lower bound of Li's confidence interval may be negative, while the upper bound is at 1. For their overall maximum likelihood estimator of  $\hat{\theta}$ , Weir and Hill (2002) gave the interval estimator

$$\left( \frac{d}{X_U} \left[ \hat{\theta} + \frac{1}{n-1} \right] - \frac{1}{n-1}, \frac{d}{X_L} \left[ \hat{\theta} + \frac{1}{n-1} \right] - \frac{1}{n-1} \right),$$

based on the properties of a  $\chi_d^2$  distribution where

$$d = (r-1) \sum_{l=1}^L (m_l - 1).$$

## Hypothesis Testing

Exact tests, bootstrap approaches and parametric tests have been used to test for the presence of significant population differentiation. Parametric hypothesis testing requires that the sampling distribution of the test statistic be known under the null hypothesis. The parametric  $\chi^2$  contingency table test of Nei (1987), based on a fixed population scope, has been used extensively for this purpose. Allele counts with very small expected values due to low observed frequencies may produce extreme test statistic values for this test. As haplotype data came into increasing use, the problem of extreme test statistics became exacerbated since sequence data has often been characterized by having a much greater numbers of alleles relative to sample size. The assumption of the  $\chi^2$  distributional form may be violated for complicated population models.

Nonparametric tests are often more powerful if the distribution assumed in a parametric test poorly approximates the populations of interest. Due to the complicated nature of the sampling distribution of  $\hat{\theta}$ , several groups have developed nonparametric procedures for the purpose of hypothesis testing of population differentiation. The actual null hypothesis tested varies between fixed and random population scopes. The fixed population scope is necessarily concerned with differences between mean allele or genotype frequencies, while in the random population scope, concern lies with testing variances of allele frequencies, as quantified by  $\theta$ , for significant differentiation between populations.

Several exact tests in the fixed population scope have been introduced (Hudson et al., 1992; Raymond and Rousset, 1995; Goudet et al., 1996; Hudson, 2000;

Sundell and Durrett, 2001). Of these, the recent nearest-neighbor statistic,  $S_{nn}$ , of Hudson, which uses the frequency with which the nearest neighbor in sequence space appears within the same population, was the most powerful under data simulated from several population models, independent of the true underlying value of population differentiation. For random populations, both resampling and parametric approaches have been described. Weir (1996) advocates testing for population differentiation using the nonparametric bootstrap over loci. This requires large numbers of loci for good sampling properties, but allows more biologically interesting inferences at the species level to be made (Dodds, 1986).

## Sampling Properties of Estimators of Gene Diversity and Heterozygosity

Exact expressions for fixed and random population scopes (Weir, 1989; Weir et al., 1990), and bootstrap resampling approaches for fixed populations (Shete, 2003; Weir, 1996), have been used to determine the sampling properties of gene diversity and heterozygosity. The bootstrap procedure should not be applied to a random population model because the resampling would disrupt associations between genes in individuals (Weir, 1989). The commonly used estimator of gene diversity defined in (1.1) has the expectation in the total population context of

$$E(\tilde{d}_{it}) = d_{it} \left[ (1 - \theta) - \frac{1}{2n}(1 + F - 2\theta) \right]$$

(Weir, 1989). It can be seen from this equation that the bias of sample gene diversity will be minimized as sample size increases. This bias has been found to be

generally small, particularly relative to the large sampling variance of the estimator (Shete, 2003). However, Shete has also determined the form of a uniform minimum variance unbiased estimator (UMVUE) of gene diversity for a fixed population model, by correcting the bias of (1.1). The estimator of heterozygosity given by Weir et al. (1990) is unbiased.

The magnitude of the total variance of sample gene diversity is a result of the scope of inferences to be made, the number of loci and individuals per population sampled, mating system, and the distribution of alleles across the populations of interest. Differences in these factors can result in differences in the sampling variance of gene diversity from equivalent sized samples of different species. As with  $\hat{\theta}$ , the sampling variance of the estimators of gene diversity and heterozygosity in a random population model has a component due to statistical sampling of individuals, which can be minimized through sampling design, and a component due to the genetical sampling individuals over time, which cannot be minimized, even by complete census of populations.

Expressions for the total variance of sample gene diversity and heterozygosity must account for both within and between population variation for the case of more than one population sampled. Weir (1989) and Weir et al. (1990) determined the total variance of sample gene diversity and heterozygosity, respectively, using exact expressions involving genotype frequencies and descent measures under a variety of different evolutionary models. The total variance of sample heterozygosity is minimized as sample size increases (Weir et al., 1990). For a mixed-mating or random mating system, this variance results mostly from associations of genes between individuals. Because of this, the total variance of sample heterozygosity

is minimized most efficiently by sampling more individuals, rather than increasing the number of loci sampled. In contrast, increasing the number of loci sampled, rather than the number of individuals sampled, has the strongest minimization of variance effect for unlinked loci in populations at migration-drift equilibrium.

## Literature Cited

- Balding, D. J. 2003. Likelihood-based inference for genetic correlation coefficients. *Theor. Popul. Biol.* 63:221–230.
- Balding, D. J. and Nichols, R. A. 1995. A method for characterizing differentiation between populations at multi-allelic loci and its implications for establishing identity and paternity. *Genetica* 96:3–12.
- Balloux, F. and Lugon-Moulin, N. 2002. The estimation of population differentiation with microsatellite markers. *Mol. Ecol.* 11:155–165.
- Beaumont, M. A., Zhang, W. Y., and Balding, D. J. 2002. Approximate Bayesian computation in population genetics. *Genetics* 162:2025–2035.
- Chakraborty, R. and Danker-Hopfe, H. 1991. Analysis of population structure: a comparative study of different estimators of Wright's Fixation indices. Pp. 203–254 *in* C. R. Rao and R. Chakraborty, eds. *Handbook of Statistics*, Vol. 8. Elsevier, New York.
- Charlesworth, B. 1998. Measures of divergence between populations and the effect of forces that reduce variability. *Mol. Biol. Evol.* 15:538–543.
- Cockerham, C. C. 1967. Group inbreeding and coancestry. *Genetics* 56:89–104.
- Cockerham, C. C. 1969. Variance of gene frequencies. *Evolution* 23:72–84.

- Cockerham, C. C. 1973. Analyses of gene frequencies. *Genetics* 74:679–700.
- Cockerham, C. C. and Weir, B. S. 1987. Correlations, descent measures: Drift with migration and mutation. *Proc. Natl. Acad. Sci. U.S.A.* 84:8512–8514.
- Cockerham, C. C. and Weir, B. S. 1993. Estimation of gene flow from  $F$ -statistics. *Evolution* 47:855–863.
- Corander, J., Waldmann, P., and Sillanpää, J. M. 2003. Bayesian analysis of genetic differentiation between populations. *Genetics* 163:367–374.
- Dodds, K. G. 1986. Resampling Methods in Genetics and the Effect of Family Structure in Genetic Data. Ph.D. thesis, North Carolina State University, Raleigh, NC.
- Excoffier, L., Smouse, P. E., and Quattro, J. M. 1992. Analysis of Molecular Variance inferred from metric distances among DNA haplotypes: application to human mitochondrial DNA restriction data. *Genetics* 131:479–491.
- Fu, R., Gelfand, A. E., and Holsinger, K. E. 2003. Exact moment calculations for genetic models with migration, mutation, and drift. *Theor. Popul. Biol.* 63:231–243.
- Fu, Y.-X. and Li, W.-H. 1997. Estimating the age of the common ancestor of a sample of DNA sequences. *Mol. Biol. Evol.* 14:195–199.
- Goudet, J. M., Raymond, M., De Meeus, T., and Rousset, F. 1996. Testing differentiation in diploid populations. *Genetics* 144:1933–1940.

- Graham, J., Curran, J., and Weir, B. S. 2000. Conditional genotypic probabilities for microsatellite loci. *Genetics* 155:1973–1980.
- Hill, W. G. and Weir, B. S. 2004. Moment estimation of population diversity and genetic distance from data on recessive markers. *Mol. Ecol.* 13:895–908.
- Holsinger, K. E. 1999. Analysis of genetic diversity in geographically structured populations: a Bayesian perspective. *Hereditas* 130:245–255.
- Holsinger, K. E., Lewis, P. O., and Dey, D. K. 2002. A Bayesian approach to inferring population structure from dominant markers. *Mol. Ecol.* 11:1157–1164.
- Holsinger, K. E. and Wallace, L. E. 2004. Bayesian approaches for the analysis of population genetic structure: an example from *Platanthera leucophaea* (*Orchidaceae*). *Mol. Ecol.* 13:887–894.
- Hudson, R. R. 2000. A new statistic for detecting genetic differentiation. *Genetics* 155:2011–2014.
- Hudson, R. R., Slatkin, M., and Maddison, W. P. 1992. Estimation of levels of gene flow from DNA sequence data. *Genetics* 132:583–589.
- Jiang, C.-J. 1987. Estimation of  $F$ -statistics in Subdivided Populations. Ph.D. thesis, North Carolina State University, Raleigh, NC.
- Lange, K. 1995. Applications of the Dirichlet distribution to forensic match prob-

- abilities. *Genetica* 96:107–117.
- Li, Y.-J. 1996. Characterizing the Structure of Genetic Populations. Ph.D. thesis, North Carolina State University, Raleigh, NC.
- Long, J. C. 1986. The allelic correlation structure of Gaij- and Kalam-speaking people. I. The estimation and interpretation of Wright's  $F$ -statistics. *Genetics* 112:629–647.
- Lynch, M. and Crease, T. J. 1990. The analysis of population survey data on DNA sequence variation. *Mol. Biol. Evol.* 7:377–394.
- Lynch, M. and Milligan, B. G. 1994. Analysis of population genetic structure with RAPD markers. *Mol. Ecol.* 3:91–99.
- Malécot, G. 1948. *Les Mathématiques de l'Hérédité*. Masson, Paris.
- Marchini, J. L. and Cardon, L. R. 2002. Discussion on the meeting on 'Statistical modelling and analysis of genetic data'. *J. R. Statist. Soc. B* 64:740–741.
- Marjoram, P., Molitor, J., Plagnol, V., and Tavaré, S. 2003. Markov chain Monte Carlo without likelihoods. *Proc. Natl. Acad. Sci. U.S.A.* 100:15324–15328.
- Maruyama, T. 1977. *Stochastic Problems in Population Genetics*. Lecture Notes in Biomathematics, Vol. 17. Springer, Berlin.
- McKay, J. K. and Latta, R. G. 2002. Adaptive population divergence: markers, QTL and traits. *Trends in Ecology and Evolution* 17:285–291.

- Michalakis, Y. and Excoffier, L. 1996. A generic estimation of population subdivision using distances between alleles with special reference for microsatellite loci. *Genetics* 142:1061–1064.
- Nagylaki, T. 1975. Conditions for the existence of clines. *Genetics* 80:595–615.
- Nei, M. 1973. Analysis of gene diversity in subdivided populations. *Proc. Natl. Acad. Sci. U.S.A.* 70:3321–3323.
- Nei, M. 1982. Evolution of human races at the gene level. Pp. 167–181 *in* B. Bonne-Tamir, T. Cohen and R. Goodman, eds. *Human Genetics, Part A: The Unfolding Genome*. Alan R. Liss, New York.
- Nei, M. 1987. *Molecular Evolutionary Genetics*. Columbia University Press, New York.
- Nicholson, G. and Donnelly, P. 2002. Assessing population differentiation and isolation from single-nucleotide polymorphism data. *J. R. Statist. Soc. B* 64:1–21.
- Ohta, T. and Kimura, M. 1969. Linkage disequilibrium in subdivided populations. *Genetics* 13:47–55.
- Pritchard, J. K., Stephens, M., and Donnelly, P. 2000. Inference of population structure using multilocus genotype data. *Genetics* 155:945–959.
- Raufaste, N. and Bonhomme, F. 2000. Properties of bias and variance of two multiallelic estimators of  $F_{ST}$ . *Theor. Popul. Biol.* 57:285–296.

- Raymond, M. and Rousset, F. 1995. An exact test for population differentiation. *Evolution* 49:1280–1283.
- Reynolds, J. 1981. Genetic Distance and Coancestry. Ph.D. thesis, North Carolina State University, Raleigh, NC.
- Reynolds, J., Weir, B. S., and Cockerham, C. C. 1983. Estimation of the coancestry coefficient: basis for a short-term genetic distance. *Genetics* 105:767–779.
- Robertson, A. and Hill, W. 1984. Deviations from Hardy-Weinberg proportions: sampling variances and use in estimation of inbreeding coefficients. *Genetics* 107:703–718.
- Rogers, A. R. 1986. Population differences in quantitative characters as opposed to gene frequencies. *Am. Nat.* 127:729–730.
- Rothman, E. D., Sing, C. F., and Templeton, A. R. 1974. A model for analysis of population structure. *Genetics* 78:943–960.
- Rousset, F. 1996. Equilibrium values of measures of population subdivision for stepwise mutation processes. *Genetics* 142:1357–1362.
- Rousset, F. 1999. Genetic differentiation in populations with different classes of individuals. *Theor. Popul. Biol.* 55:297–308.
- Rousset, F. and Raymond, M. 1997. Statistical analyses of population genetic data: new tools, old concepts. *Trends in Ecology and Evolution* 12:313–317.

- Shete, S. 2003. Uniformly minimum variance unbiased estimation of gene diversity. *J. Hered.* 94:421–424.
- Slatkin, M. 1991. Inbreeding coefficients and coalescence times. *Genetical Research (Cambridge)* 58:167–175.
- Slatkin, M. 1995. A measure of population subdivision based on microsatellite allele frequencies. *Genetics* 139:457–462.
- Smouse, P. E. and Williams, R. C. 1982. Multivariate analysis of HLA-disease associations. *Biometrics* 38:757–768.
- Spitze, K. 1993. Population structure in *Daphnia obtusa*: quantitative genetic and allozyme variation. *Genetics* 135:367–374.
- Sundell, N. M. and Durrett, R. T. 2001. Exponential distance statistics to detect the effects of population subdivision. *Theor. Popul. Biol.* 60:107–116.
- Tavaré, S., Balding, D. J., Griffiths, R. C., and Donnelly, P. 1997. Inferring coalescence times from dna sequence data. *Genetics* 145:505–518.
- Tufto, J. Engen, S. and Hindar, K. 1996. Inferring patterns of migration from gene frequencies under equilibrium conditions. *Genetics* 144:1911–1921.
- Van Dongen, S. 1995. How should we bootstrap allozyme data? *Heredity* 74:445–447.

- Weir, B. S. 1989. Sampling properties of gene diversity, Ch. 2 *in* Brown, A. H. D., Clegg, M. T., Kahler, A. L. and B. S. Weir, eds. Plant Population Genetics, Breeding and Genetic Resources. Sinauer, Sunderland, MA.
- Weir, B. S. 1996. Genetic Data Analysis. Sinauer, Sunderland, MA, 2 edition.
- Weir, B. S. and Cockerham, C. C. 1984. Estimating  $F$ -statistics for the analysis of population structure. *Evolution* 38:1358–1370.
- Weir, B. S. and Hill, W. G. 2002. Estimating  $F$ -statistics. *Ann. Rev. Gen.* 36:721–750.
- Weir, B. S., Reynolds, J., and Dodds, K. G. 1990. The variance of sample heterozygosity. *Theor. Popul. Biol.* 37:235–253.
- Wright, S. 1921. Systems of mating. *Genetics* 6:111–178.
- Wright, S. 1931. Evolution in Mendelian populations. *Genetics* 16:97–159.
- Wright, S. 1951. Genetical structure of populations. *Annals of Eugenics* 15:323–354.
- Yang, R.-C. 1998. Estimating hierarchical  $F$ -statistics. *Evolution* 52:950–956.
- Zhivotovsky, L. A. 1999. Estimating population structure in diploids with multi-locus dominant DNA markers. *Mol. Ecol.* 8:907–913.

## Chapter 2

# SAMPLING PROPERTIES OF POPULATION-SPECIFIC $F_{ST}$ ESTIMATORS

Johnson AM and BS Weir

## Abstract

The accuracy and precision of two estimators of population-specific  $F_{ST}$  developed by Weir and Hill (2002) are evaluated through both simulation and analysis of an empirical data set. Estimating different values of coancestry for individual populations, instead of using the classical method of moments estimator of  $\theta$  described by Weir and Cockerham (1984) as the average coancestry for a group of populations, provides a more detailed characterization of the population structure of groups of populations. Of the two estimators considered, results from population simulations showed that a method of moments estimator for population-specific  $\theta$  is relatively unbiased with a large sampling variance, which increases as coancestry increases in a population. Sampling more loci has a much stronger effect on reducing this sampling variance than sampling more individuals. The other estimator evaluated, obtained through a maximum-likelihood approach, fails to estimate the coancestry in a population for two iterative approaches and a non-iterative approach and is not recommended for future analyses. Additionally four different approaches to combining information about  $\theta$  over loci and populations are evaluated, confirming expectations given by Weir and Hill (2002) of the equivalency of the classical  $\theta$  and the average of the population-specific method of moments estimators. Analysis of empirical data revealed problems with loci with very low polymorphism levels. Using polymorphism cutoffs for excluding loci from analyses is rejected in favor of obtaining estimates from combining data over a reasonably large number of loci, which should both make the estimates more robust to effects of extreme data and decrease the sampling variance of the estimates.

## Introduction

The estimation of  $F_{ST}$  and analogues of  $F_{ST}$  has been discussed widely in the literature for many years (including Nei, 1973; Weir and Cockerham, 1984; Robertson and Hill, 1984; Slatkin, 1995; Nagylaki, 1998). Until recently, all estimation methods assumed  $F_{ST}$  to be the same across all populations, and estimates were then an average of the actual population values of  $F_{ST}$ . As Balding (2003) points out, the usual demographic variation in a collection of populations and the effect of different sizes of these populations makes this average approach an oversimplification. Estimates of  $F_{ST}$  specific to individual populations have the potential to increase the amount of information about differentiation that population studies can produce by providing a more detailed summary of the distribution and amount of genetic variation in a group of populations.

$F_{ST}$  has been frequently used in fields ranging from forensics, to evolutionary and conservation biology for a variety of species, as a measure for describing the extent of common ancestry in a group of populations. Thus often called the coancestry coefficient,  $F_{ST}$  was defined by Wright in 1951 as the correlation of pairs of alleles within different individuals in a population, relative to a random pair of alleles from the total group of populations. This so-called  $F$ -statistic, is in fact a population parameter, an unobservable true characteristic of a population that can be estimated by statistics. To emphasize the status of  $F_{ST}$  as a parameter, for the remainder of this chapter we will use Cockerham's equivalent notation of  $\theta$ .

For the purposes of this chapter we will also refer to within-species groups as populations, each of which has a true population-specific  $\theta$  value. The total

population will be the species as a whole about which we wish to make inferences. Different values of population-specific  $\theta$  arise from both genetic and statistical sampling. These differences in  $\theta$  are the results of the differing action of evolutionary forces such as drift, selection and migration on populations over time, and also the result of the sampling processes used to collect the data.

Problems with classical analyses of population structure introduced by the simplifying assumption of a common value of  $\theta$  across all populations were illustrated in detail by Long and Kittles (2003). Their results showed that overall estimates of  $\theta$  from global human data sets were essentially meaningless, in that they failed to describe important local patterns and amounts of genetic variation, leading to the problem of important differentiation being commonly overlooked by researchers. The sequential model-fitting approach to estimating  $\theta$  described by these authors can be used to determine the extent of the effect of the classical assumptions of independently and equally diverged populations on coancestry estimates.

Several groups of investigators have now begun relaxing the constraint that  $F_{ST}$  be the same across populations using both frequentist and Bayesian approaches. Nicholson and Donnelly (2002) and Weir and Hill (2002) devised new parameterizations of population models that defined a parameter specific to each population, allowing inferences of different amounts of coancestry for different populations. Nicholson and Donnelly (2002) approached this expanded parameterization from a Bayesian perspective, in the context of an application to SNP data by modelling allele frequencies as normally distributed. The authors justify this model as having a reasonable fit to recently diverged, non-equilibrium populations. They also described how to obtain an associated measure of fit of the data to their hierarchical

model.

A complementary approach was given by Balding (2003), which differed in that the allele frequencies were modelled as Beta distributed for the bivariate case. Beta has been shown to be the stationary distribution of allele frequencies for the island model (Wright, 1931) and to be a reasonable approximation for the stationary distribution of the finite stepping-stone model (Maruyama, 1977). The approach of Balding (2003) also differed from that of Weir and Hill (2002) in that  $F_{ST}$  was defined as a scaled variance, where  $F_{ST} \geq 0$ , rather than a correlation where  $F_{ST}$  can be negative, as used here. Holsinger and Wallace (2004) gave an extension to the hierarchical model of Balding by describing a summary statistic that compared the entropy of the posterior to the prior distribution of the coancestry parameters in order to determine the amount of information in a data set about a given parameter.

In contrast to the Bayesian approaches of these other investigators, two estimators were obtained with a frequentist approach by Weir and Hill (2002), extending the estimator of  $\theta$  of Weir and Cockerham to allow for population-specific parameters  $\theta_i$ , where  $i$  indexes the sampled populations. The first estimator (MOME), obtained through a method of moments approach, was a direct extension of the previous Weir and Cockerham moment estimator of  $\theta$ . As a moment estimator, it is a ratio of unbiased estimates and is therefore expected to be essentially unbiased, but to also have a large sampling variance. The method of moments approach to obtaining estimates has the benefit that the estimator requires the knowledge of only the first two moments rather than the form of the entire distribution of the parameter of interest (Weir and Hill, 2002).

The second estimator of  $\theta_i$  described by Weir and Hill (2002) was a maximum likelihood based estimator (MLE) whose development included the assumption that the sample allele frequencies be multivariate normally distributed as had been previously made by Smouse and Williams (1982), Long (1986), and Nicholson and Donnelly (2002). The method of obtaining a point estimator by maximizing the likelihood of the parameter of interest given the data has several desirable properties, such as invariance to transformation. However, it is well known that maximum likelihood estimators can be highly unstable in their sampling properties for extremely flat likelihood functions and this can particularly be a problem for maximum likelihood estimators that must be obtained through numerical maximization rather than having an explicit solution (Casella and Berger, 1990), as is the case with the Weir and Hill (2002) maximum likelihood estimator.

In this study the sampling properties of the population-specific  $\theta_i$ , MOME and MLE estimators of Weir and Hill (2002), are evaluated through population simulation and the analysis of a previously published empirical data set. The bias and variance of these estimators is characterized in order to determine which has the most desirable sampling properties, and to determine if these properties are uniformly consistent for varying levels of differentiation, number of loci and individuals sampled, and amounts of recombination. Because the distribution of estimates of  $\theta_i$  is a complicated function, it is not obvious what properties these estimators will have, and is therefore worth exploring to see if our theoretical assumptions hold up in different situations. We approach these questions by applying the estimators to simulated data, empirically determining the sampling properties of the estimators since exact results are difficult to come by for ratio estimators. In addition, we

evaluate two proposed methods of iterating the ML estimates using simulated data and four ways of combining information over populations to estimate  $\hat{\theta}$ . These approaches include a newly proposed maximum-likelihood based estimator of  $\hat{\theta}$  that constrains populations to have the same value of coancestry, but allows for sample sizes to vary.

## Theory

The population model used consists of  $r$  independent, totally isolated populations that differ from each other due only to the effects of random genetic drift. Define  $\theta_{iu}$  to be the correlation of the indicator variables for allele type  $u$  of two alleles sampled from different individuals of the same population. We then assume that this correlation is the same for all allele types  $u$  and  $u'$  so that  $\theta_{iu} = \theta_{iu'}$ . To continue with the description of the population model, in our notation, the assumption of independence of populations is equivalent to  $\theta_{ii'} = 0$ , for all possible pairs of populations  $i$  and  $i'$ . The populations are descended from a single common ancestral population with a constant population size across each generation although the populations may differ in size from each other. Random samples of  $n_{il}$  monoeucous, haploid individuals are obtained from population  $i$  at locus  $l$ , at which there are  $m_l$  allele types.

The method of moments estimator obtained by Weir and Hill (2002) can be written as

$$\hat{\theta}_{il} = 1 - \frac{x_{il}}{y_l} \quad (2.1)$$

where

$$\begin{aligned} x_{il} &= \frac{n_{il}}{n_{il} - 1} \sum_{u=1}^m \tilde{p}_{ilu}(1 - \tilde{p}_{ilu}), \\ y_l &= \frac{1}{\sum_{i=1}^r n_{ilc}} \sum_{u=1}^m \sum_{i=1}^r [n_{il}(\tilde{p}_{ilu} - \bar{p}_{lu})^2 + n_{ilc}\tilde{p}_{ilu}(1 - \tilde{p}_{ilu})], \end{aligned}$$

$\tilde{p}_{ilu}$  indicates the sample allele frequency,  $\bar{p}_{lu}$  is the average allele frequency across

all populations sampled and an adjusted sample size can be calculated as

$$n_{ilc} = n_{il} - \frac{n_{il}^2}{\sum_{i=1}^r n_{il}}.$$

The MOM estimates over all populations, all loci, or over both, are explicitly

$$\hat{\theta}_i = 1 - \frac{\sum_{l=1}^L x_{il}}{\sum_{l=1}^L y_l}, \quad \hat{\theta}_l = 1 - \frac{\sum_{i=1}^r x_{il}}{r y_l} \quad \text{and} \quad \hat{\theta} = 1 - \frac{\sum_{l=1}^L \sum_{i=1}^r x_{il}}{r \sum_{l=1}^L y_l}. \quad (2.2)$$

By assuming that the randomly sampled allele frequencies from a set of populations have a multivariate normal distribution, Weir and Hill (2002) derived estimation equations to be numerically solved for a MLE of  $\theta_i$  as

$$\hat{\phi}_{il} = \frac{1}{(m_l - 1)} \sum_{u=1}^{m_l} \frac{(\tilde{p}_{ilu} - p_{lu})^2}{p_{lu}} \quad (2.3)$$

$$\hat{p}_{lu} = \frac{\sum_{i=1}^r \left(1 - \frac{\tilde{p}_{ilu}^2}{\phi_{il} p_{lu}}\right)}{\sum_{u=1}^{m_l} \sum_{i=1}^r \left(1 - \frac{\tilde{p}_{ilu}^2}{\phi_{il} p_{lu}}\right)}, \quad (2.4)$$

where the relationship between  $\phi_{il}$  and  $\theta_{il}$  can be written as

$$\hat{\theta}_{il} = \frac{n_{il} \hat{\phi}_{il} - 1}{n_{il} - 1}. \quad (2.5)$$

Weir and Hill (2002) suggested iterating between the Eqns. (2.3) and (2.4) until the estimates for both sets of parameters converge. Other possibilities would be to iterate between estimating  $\phi_i$ 's and the variance-covariance matrix  $\Omega$  of the allele frequencies, as given by Weir and Hill (2002) or to evaluate the total likelihood using numerical optimization software. These three methods for obtaining the ML estimates of  $\theta_{il}$  were applied to simulated data, but unfortunately neither iterative method produced estimates that converged consistently to reasonable

values within the appropriate parameter ranges and similar problems were observed with the likelihood optimization. These results are further described in Appendix 1. Because of the failure of these iteration methods for practical estimation, in the remainder of this chapter we use the first iteration estimate of  $\hat{\phi}_{il}$  of Eq. (2.3) and assume that the population value  $p_{lu}$  can be estimated by the sample value  $\bar{p}_{lu}$ . Estimates of  $\theta_{il}$  can then be obtained by transforming  $\hat{\phi}_{il}$  with (2.5). Although this will be referred to with the notation  $\hat{\theta}_{i,ML}$ , it is not actually then a maximum likelihood estimate, in that the first iterates are not being maximized over the likelihood surface.

Overall ML estimates are obtained by combining estimators as weighted averages over loci to produce an estimate of  $\theta_i$ , combining over populations to produce  $\hat{\theta}_l$ , or combining over both to obtain  $\hat{\theta}$ , which was shown to be equivalent algebraically by Weir and Hill (2002) to the previous multiple locus moment estimator described by Weir and Cockerham (1984) for the case of two alleles. The weighted estimators summed over the appropriate indices can be expressed as

$$\hat{\theta}_i = \frac{\sum_{l=1}^L n_{il} \hat{\theta}_{il}}{\sum_{l=1}^L n_{il}}, \quad \hat{\theta}_l = \frac{\sum_{i=1}^r n_{il} \hat{\theta}_{il}}{\sum_{i=1}^r n_{il}} \quad \text{and} \quad \hat{\theta} = \frac{\sum_{l=1}^L \sum_{i=1}^r n_{il} \hat{\theta}_{il}}{\sum_{l=1}^L \sum_{i=1}^r n_{il}}. \quad (2.6)$$

### Maximum likelihood estimator for unequal sample sizes

A more general approach to estimating the overall populations coancestry coefficient  $\theta$  than the method of moments estimator of Weir and Cockerham (1984) can be obtained by a maximum likelihood approach. For this development we constrain the coancestry of a group of populations to be a single value  $\theta$ , but allow for unequal sample sizes. This extension to unbalanced sampling could be useful

for the case of bi-allelic data, such as SNPs, where there are not enough degrees of freedom to estimate  $\theta_i$  for each population.

The log-likelihood for independent sample sizes given these assumptions is then

$$\ln L = -\frac{m-1}{2} \sum_i \ln(\phi_i) - \frac{r}{2} \sum_u \ln(p_u) - \frac{1}{2} \sum_i \sum_u \frac{(\tilde{p}_{ilu} - \bar{p}_u)^2}{\phi_i p_u}. \quad (2.7)$$

As an approximation, suppose that  $p_u = \bar{p}_u$ , the sample mean value. Because the  $\theta_i = \theta_{i'}$  for all populations  $i$  and  $i'$ , this leads to the expression

$$\phi_i = \frac{1}{n_i} [1 + (n_i - 1)\theta].$$

Then

$$\frac{\partial \ln L}{\partial \theta} = -\frac{m-1}{2} \sum_i \frac{n_i - 1}{1 + (n_i - 1)\theta} + \frac{1}{2} \sum_i \sum_u \frac{n_i(n_i - 1)(\tilde{p}_{ilu} - \bar{p}_u)^2}{\bar{p}_u [1 + (n_i - 1)\theta]^2}.$$

Let  $X_i = \sum_u (\tilde{p}_{ilu} - \bar{p}_u)^2 / \bar{p}_u$ , and write

$$f(\theta) = \sum_i \left( \frac{(m-1)(n_i - 1)}{1 + (n_i - 1)\theta} - \frac{n_i(n_i - 1)X_i}{[1 + (n_i - 1)\theta]^2} \right). \quad (2.8)$$

Setting  $f(\theta) = 0$  gives the MLE. This requires numerical methods to solve for  $\theta$ .

Note that

$$\begin{aligned} f(0) &= \sum_i (n_i - 1)(m - 1 - n_i X_i) \\ f(1) &= \sum_i \frac{n_i - 1}{n_i} (m - 1 - X_i). \end{aligned}$$

## Methods

Both simulation and analysis of a previously published empirical data set were employed to study the sampling properties of the  $\theta_i$  estimators. Population data were simulated with software written in  $C^{++}$ , with 1000 replicates performed for each set of population parameters simulated. Each replicate simulation consisted of three independent populations of haploid individuals with 1 to 20 loci, where each locus had three alleles and the population sizes varied from 200 to 1995 individuals. These populations were initialized by sampling from an infinite-sized ancestral population to make up three populations. The ancestral population was constructed with equal frequencies for each allele type, at each locus. Generations were constant in size and new generations were simulated until a specific  $\theta_i$  value was reached in each population. After the target theoretical  $\theta_i$  value for each population had been reached, a sample was taken from these final generations. Sample sizes were either balanced, meaning the samples were all the same size, or unbalanced, where the sample size was allowed to vary. In either case, samples of size 25 to 100 were taken. From these samples, the values of the two estimators of  $\theta_i$  were calculated for each replicate, these estimates providing collectively over the 1000 replicates an empirical sampling distribution of the estimators for the particular parameter set used in the simulations.

For the simulated loci, an amount of recombination was simulated ranging from 0 for effectively completely linked loci, to 0.5, where the loci were effectively independent. The genotypes of individual offspring were created by first sampling randomly, with replacement, two parental gametes from a population. With

equal probability, one of the parental gametes was then selected to be the primary parental gamete. The allele type of the first locus of the primary parental gamete then was assigned to be the allelic identity of the first locus of the offspring. Next, either the primary gamete contributes an allele to the second locus or, with a specific probability of recombination, it is determined that a recombination event occurs. If this happens, the secondary parental gamete then becomes the primary parental gamete, and contributes the allele to the second locus of the offspring. The allelic identities of the remaining loci in the offspring are thus assigned from the current primary parental gamete (whose identity could potentially switch again if another recombination event occurs), moving in sequence through the remaining loci on the simulated chromosome.

Population-specific estimates of  $\theta_i$  were also obtained from the analysis of a data set previously collected and analyzed by Planes and Fauvelot (2002). The data set is composed of 18 polymorphic allozyme loci genotyped in 727 surgeonfish (*Acanthurus triostegus*), collected at 15 locations throughout the Pacific Ocean, and a single group in the Indian Ocean (Mozambique). Planes and Fauvelot (2002) found that the sites had significant, geographically structured population differentiation, with  $\hat{\theta} = 0.0199$ .  $\hat{H}_e$  was estimated to be 0.117 with a range of 0.045 to 0.181 across groups, which makes these fish some of the most polymorphic marine fish so far studied (Planes and Fauvelot, 2002). The populations fell into five general groups that were relatively undifferentiated within themselves. These are the Western Pacific, composed of Guam, the Philippines, Palau and the Great Barrier Reef samples; the central Pacific, composed of the Solomons, New Caledonia and Fiji; the Hawaiian Archipelago, composed of the Marquesas, Oahu and Hawaii; French

Polynesia and Clipperton Island, composed of Rangiroa, Fangataufa, Moorea and Bora-Bora; and lastly the Indian Ocean group composed solely of the Mozambique population.

Evaluating the maximum likelihood estimator of  $\theta$  given in Eq. (2.8) for unequal sample sizes, and termed  $\theta_{U,ML}$  in the results presented here, was done using the R numerical optimization function ‘optim’ using the unconstrained default Nelson-Mead optimization algorithm, initializing the search to  $\theta = 0.5$  and using sample values from population simulations over 1000 replicates for each data set considered to evaluate  $X_i$ . For all cases studied the optimization converged to a single estimate. The other optimization methods, both constrained and unconstrained, available with the optim function were also initially used to obtain estimates for  $\theta_U$ , and a number of different initialization values were tried for this optimization procedures and all converged to the same estimates as the Nelson-Mead procedure that was used for the major part of this analysis.

The estimates of  $\theta_{U,ML}$  and their sampling variance was compared to three other approaches to combining information over populations to obtain estimates for  $\theta$ . These included the method of moments estimator of Weir and Cockerham (1984), and the weighted average of the MOME and MLE of  $\theta_i$  of Weir and Hill (2002). This average was calculated as

$$\theta_W = \frac{\sum_{i=1}^r n_i \hat{\theta}_i}{\sum_{i=1}^r n_i}$$

for both of the  $\theta_i$  estimators.

# Results

## Evaluation of Estimator Properties by Simulation

This section is an evaluation of the properties of the population-specific  $\theta_i$  MOME given in Eqns. (2.1) and (2.2), and the first iterate MLE given in Eqns. (2.3), (2.5) and (2.6). The accuracy of the estimators of  $\theta_i$  is given by the bias of the estimator, while the precision is given by the variance. We are particularly concerned with the effects of the levels of differentiation, the amount of recombination, the number of loci sampled and the sample sizes on the accuracy and precision of the estimators. In addition, we are interested in determining how robust the estimator properties will be to unequal amounts of coancestry and unbalanced sampling across populations.

Table 2.1 and Figure 2.1 give a general idea of how the performance of the estimators of  $\theta_i$  obtained by the MOM and ML approaches are affected by unbalanced sampling and different amounts of coancestry in the simulated populations. For the point estimates of  $\theta_i$  presented in Table 2.1, the MOM estimates are unaffected by unbalanced sampling and the MOME is able to produce estimates that reveal differences in the levels of population structure of the populations simulated. Figure 2.1 gives an idea of the effects of unbalanced samples and coancestry levels on the sampling properties of the estimators. The MOME can be seen to be slightly negatively biased, which is essentially unaffected by unbalanced sampling, while the variance of the MOME responds somewhat to increased sample sizes. However, the variance of the MOME also becomes extremely large for high levels of population differentiation.

Table 2.1: Effect of different amounts of coancestry ( $\theta_i$ ) and sample sizes ( $n_i$ ) on the method of moments (MOM) and first iterate maximum likelihood (ML) estimates of  $\hat{\theta}_i$ . The data shown is from four simulations of three populations each, with 20 loci sampled. Each simulation was replicated 1000 times. Results shown in this table are the average of the estimates of  $\theta_i$  over all replicates.

	Pop 1	Pop 2	Pop 3
$n_i$	50	50	50
$\theta_i$	0.010	0.010	0.010
$\hat{\theta}_{i,MOM}$	0.010	0.010	0.010
$\hat{\theta}_{i,ML}$	0.000	0.000	0.000
$n_i$	50	50	50
$\theta_i$	0.005	0.010	0.050
$\hat{\theta}_{i,MOM}$	0.005	0.010	0.050
$\hat{\theta}_{i,ML}$	0.003	0.004	0.017
$n_i$	30	50	70
$\theta_i$	0.010	0.010	0.010
$\hat{\theta}_{i,MOM}$	0.010	0.010	0.010
$\hat{\theta}_{i,ML}$	0.003	0.000	-0.002
$n_i$	30	50	70
$\theta_i$	0.005	0.010	0.050
$\hat{\theta}_{i,MOM}$	0.005	0.010	0.051
$\hat{\theta}_{i,ML}$	0.010	0.010	0.010

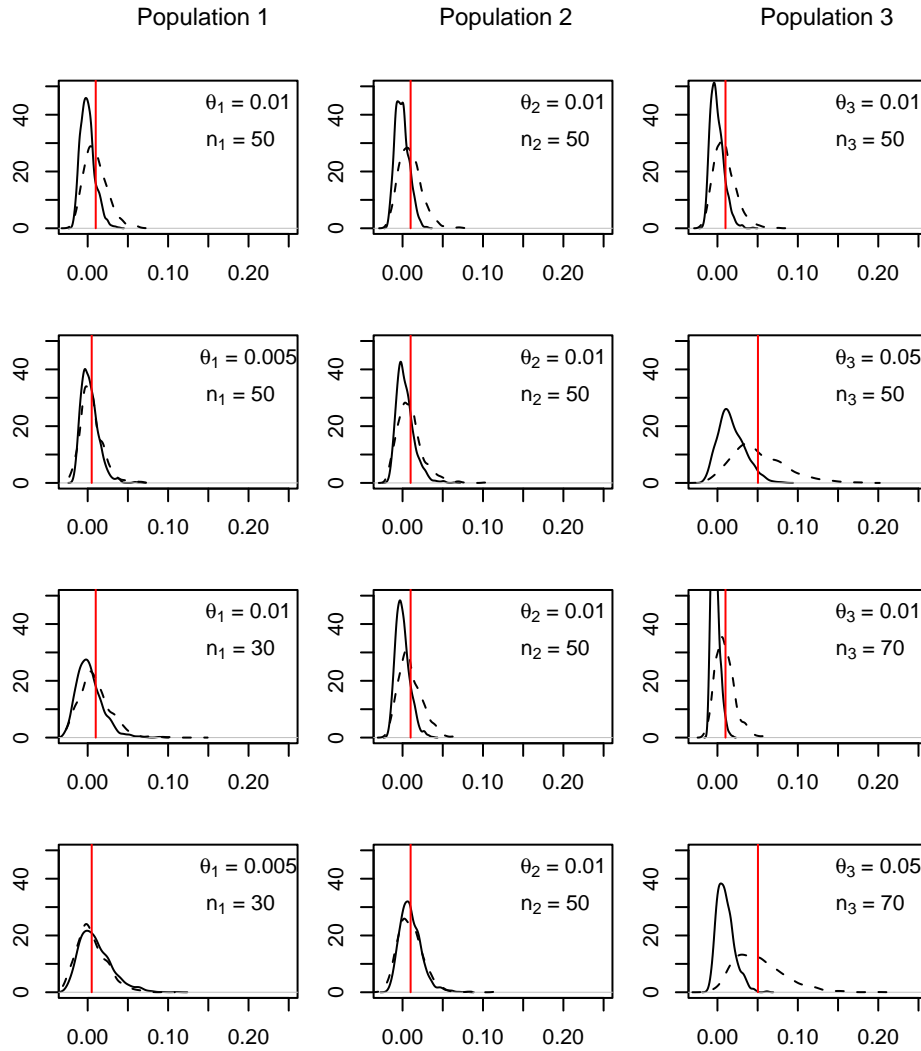


Figure 2.1: Empirically derived sampling distributions for estimates of  $\theta_i$  obtained from four sets of data simulated with three populations and five loci. The solid line is the sampling distribution of the first iterate MLE of  $\theta_i$ , while the dashed line is the distribution of the MOME. The vertical lines in the plots indicate the true  $\theta_i$  values of the simulations, and the y-axis represents the frequency out of the 1000 replicates that a particular estimate value was obtained.

In contrast, while the MLE has a much smaller variance than the MOME, this estimator is very poorly estimating the quantity of interest. Further study of the sampling properties of the MLE presented in Figure 2.1 and Table 2.2 show that the undesirable behavior of this estimator becomes particularly pronounced for the case of unbalanced samples. For example, with the case of the fourth data set shown in Figure 2.1, the bias of the MLE becomes very large as the value of  $\theta_i$  in the simulated population increases because the mean estimates are roughly equivalent across all samples of this data set and not close to the value of  $\theta_i$  in that particular population. This can be seen directly as well in Table 2.2.

The more quantitative results in Table 2.2 were used to determine additionally the effects of simulated recombination and of increasing the number of loci sampled on the sampling properties of the estimators. A more detailed evaluation of the performance of the MOME shows that this estimator produces estimates with relatively small, negative bias that remains unaffected by the number of loci sampled or the rate of simulated recombination. These results also confirm that the sampling variance of the MOME is inversely proportional to the number of loci sampled, as expected.

Regarding sampling design, results indicate that increasing the number of loci sampled will have a much stronger effect on decreasing the sampling variance of the estimates of  $\theta_i$  than an increase in the number of individuals, which has little effect (Table 2.2). Note that only the variance due to the effect of statistical sampling can be reduced by increasing the number of observations. Variance due to the genetical sampling occurring in the populations over generations will remain, as has been pointed out previously (Weir and Cockerham, 1984; Weir, 1996).

Table 2.2: Mean bias and standard deviation for  $\hat{\theta}_i$  obtained from 1000 replicate population simulations. For those simulations with multiple loci, the recombination between the loci was simulated with a probability of  $c$ , with  $c = 0.5$  simulating effectively unlinked loci.

	$n_i$	True		MOME			MLE		
		$\theta_i$	$c$	Number of Loci Sampled			Number of Loci Sampled		
				1	5	20	1	5	20
Bias	30	0.005	0.5	-0.0003	-0.0008	0.0004	0.0042	0.0040	0.0048
			0.1	-0.0004	-0.0003	0.0001	0.0037	0.0044	0.0050
			0.01	-0.0001	-0.0006	0.0005	0.0039	0.0056	0.0052
			0.0	-0.0016	0.0001	0.0001	0.0014	0.0049	0.0052
	50	0.01	0.5	0.0010	-0.0003	0.0005	0.0011	-0.0009	-0.0001
			0.1	-0.0006	-0.0009	-0.0001	0.0001	-0.0010	-0.0008
			0.01	-0.0013	-0.0003	0.0000	-0.0009	-0.0012	-0.0005
			0.0	-0.0019	0.0010	0.0003	-0.0013	-0.0001	-0.0005
	70	0.05	0.5	0.0023	-0.0007	0.0006	-0.0402	-0.0413	-0.0407
			0.1	0.0003	-0.0003	-0.0003	-0.0409	-0.0413	-0.0412
			0.01	-0.0016	-0.0010	0.0001	-0.0417	-0.0415	-0.0410
			0.0	-0.0044	-0.0008	-0.0002	-0.0425	-0.0412	-0.0411
SD	30	0.005	0.5	0.0406	0.0192	0.0099	0.0430	0.0193	0.0100
			0.1	0.0435	0.0189	0.0099	0.0435	0.0191	0.0103
			0.01	0.0436	0.0197	0.0104	0.0415	0.0210	0.0100
			0.0	0.0425	0.0201	0.0103	0.0394	0.0206	0.0109
	50	0.01	0.5	0.0365	0.0153	0.0081	0.0295	0.0130	0.0068
			0.1	0.0357	0.0154	0.0082	0.0305	0.0125	0.0067
			0.01	0.0343	0.0163	0.0083	0.0296	0.0128	0.0069
			0.0	0.0343	0.0158	0.0087	0.0268	0.0137	0.0073
	70	0.05	0.5	0.0660	0.0285	0.0148	0.0233	0.0101	0.0053
			0.1	0.0654	0.0290	0.0149	0.0229	0.0100	0.0054
			0.01	0.0615	0.0291	0.0165	0.0226	0.0102	0.0058
			0.0	0.0580	0.0289	0.0188	0.0210	0.0101	0.0063

Several approaches for obtaining overall estimates of  $\theta$  by combining information over loci were also evaluated using population simulation. A newly proposed, alternative MLE that constrains populations to have the same value of  $\theta$ , but allows sample sizes to vary given in (2.8), appears to be positively biased, particularly for higher  $\theta$  values, and to have a greater variance than the overall MOME for  $\theta$  also evaluated here (Table 2.3). The maximum likelihood weighted average  $\hat{\theta}_{W,ML}$ , appears to be as poor an estimator as the population-specific MLE of  $\theta_i$ . Combining method of moment estimates of  $\theta_i$  over populations as a weighted average with  $\hat{\theta}_{W,MOM}$  gives equivalent results to the classical ANOVA estimator of  $\theta$ , as was expected from the algebraic results in Eq. (7) of Weir and Hill (2002). There appears to be a small amount of bias for  $\hat{\theta}_{W,MOM}$  and  $\hat{\theta}_{MOM}$  for several of the data sets as well.

## **Effect of Low Polymorphism on the Method of Moments Estimator**

The analysis of the Planes data set brought to light cases where extreme estimates for the MOM and the MLE were obtained from loci with very low polymorphism levels, and highlighted the need for practical procedures to deal with these cases. The Planes data differed from the human forensic data and simulated data sets previously analyzed in Weir and Hill (2002) and the simulated data of this study in that the overall level of polymorphism at the loci sampled was much lower than the polymorphism of the other sets of loci examined. Unexpected values of estimates of  $\theta_i$  can be seen for several population and locus combinations in Table

Table 2.3: Estimates and standard deviations of different approaches to combining information about  $\theta$  over loci and populations for six simulated data sets. Data shown had three populations and twenty unlinked loci, each with the same value of  $\theta$  so that  $\theta_i = \theta_{i'}$ . The measures shown here include  $\hat{\theta}_W$ , the average  $\hat{\theta}$  value weighted by sample size, obtained with both the MOME and MLE,  $\hat{\theta}_U$ , a ML estimator that constrains  $\theta$  to be constant across populations, but allows for sample sizes to vary, and the classical MOM estimator given in (Weir and Cockerham, 1984).

True $\theta$	$n_1$	$n_2$	$n_3$	$\hat{\theta}_{W,MOM}$	$\hat{\theta}_{W,ML}$	$\hat{\theta}_{U,ML}$	$\hat{\theta}_{MOM}$
0.01	50	50	50	0.011 (0.005)	0.001 (0.004)	0.022 (0.007)	0.011 (0.005)
	30	50	70	0.010 (0.005)	0.000 (0.004)	0.018 (0.008)	0.010 (0.006)
	25	50	100	0.010 (0.004)	0.000 (0.003)	0.015 (0.007)	0.010 (0.005)
0.10	50	50	50	0.098 (0.016)	0.059 (0.010)	0.139 (0.021)	0.098 (0.016)
	30	50	70	0.103 (0.018)	0.060 (0.012)	0.150 (0.026)	0.102 (0.018)
	25	50	100	0.097 (0.020)	0.053 (0.011)	0.155 (0.030)	0.097 (0.019)

2.4, one of the most extreme being the estimate for the Clipperton population at locus *Me-2*. Because of the other, more serious problems with the maximum likelihood estimates, details for the problems with low polymorphism with the MLE are given in Appendix 2 and the remainder of this section is concerned solely with the MOME.

Table 2.4: Effects of low polymorphism at individual loci on the method of moments estimator of  $\hat{\theta}_{il}$  for 16 populations of surgeonfish at 18 allozyme loci.

Locus	Population															
	GBR	Guam	Palau	Philip.	Hawaii	Oahu	Marq.	Fiji	New Cal.	Solo.	Bora	Clipp.	Fanga.	Moorea	Rang.	Mozam.
<i>Aat-1</i>	0.221	0.221	1.000	-1.418	-1.299	-2.029	-0.539	1.000	1.000	1.000	1.000	-1.595	1.000	1.000	-0.476	-2.039
<i>Aat-2</i>	0.896	0.157	0.702	1.000	1.000	0.896	1.000	1.000	0.768	0.400	-1.064	-1.108	-1.030	0.055	-0.965	1.000
<i>Aat-3</i>	-0.299	-0.072	-0.458	-0.465	1.000	0.409	1.000	-0.363	-0.083	-0.433	0.769	0.466	0.599	0.145	0.226	0.855
<i>Ada</i>	0.906	0.158	0.730	1.000	0.814	-0.946	0.814	0.809	0.879	0.810	0.775	-1.515	-0.896	-0.708	-0.898	0.906
<i>Gda</i>	0.543	0.735	0.247	0.667	-0.183	-0.155	0.140	-0.004	-0.053	-0.118	0.061	0.077	-0.123	0.681	0.447	-0.157
<i>G3pdh</i>	1.000	1.000	1.000	1.000	-2.220	1.000	1.000	-3.887	1.000	-0.673	-2.911	1.000	1.000	1.000	1.000	-2.220
<i>Idhp-2</i>	1.000	1.000	-7.058	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	-7.105	1.000	1.000	1.000	-0.274
<i>Idhp-1</i>	0.344	0.344	-0.265	0.309	0.344	1.000	-2.781	0.327	-0.055	-0.329	1.000	0.250	-1.506	0.344	0.375	0.344
<i>Ldh-1</i>	1.000	1.000	1.000	-17.676	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000
<i>Ldh-2</i>	1.000	-16.793	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000
<i>Ldh-3</i>	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000
<i>Mdh</i>	0.345	0.443	0.421	0.385	0.716	0.806	1.000	0.432	0.374	0.432	0.076	0.190	-0.007	0.009	0.036	0.305
<i>Mep-1</i>	0.425	0.147	0.167	-2.686	1.000	0.711	1.000	0.133	0.445	-0.429	-0.366	-0.902	-0.366	0.147	-0.580	0.711
<i>Mep-2</i>	1.000	0.111	0.133	1.000	1.000	1.000	1.000	1.000	1.000	0.088	0.274	-11.972	1.000	0.111	1.000	1.000
<i>Pepb</i>	0.610	-0.500	0.620	0.190	-3.999	-3.880	0.610	0.600	0.747	0.206	0.682	1.000	1.000	1.000	0.629	0.610
<i>Pgdh</i>	-0.030	0.290	0.117	0.324	-0.366	-0.296	1.000	0.284	0.423	0.117	0.395	0.467	0.252	0.468	0.549	-0.385
<i>Gpi-1</i>	0.202	-0.032	-0.473	-0.134	0.620	0.901	0.901	-0.241	-0.247	-0.110	-0.342	0.469	-0.522	-0.414	-0.171	0.901
<i>Gpi-2</i>	0.765	1.000	1.000	1.000	1.000	0.765	-1.427	0.290	0.475	-0.142	0.620	1.000	0.108	0.307	1.000	-3.756
<i>Pgm</i>	-0.304	0.303	-0.131	0.109	-0.508	-0.347	1.000	-0.334	0.062	0.121	-0.227	0.037	0.321	-0.465	0.337	-0.016
Overall	0.346	0.295	0.169	0.281	0.328	0.152	0.679	0.222	0.300	0.119	0.147	-0.173	-0.088	0.127	0.061	0.202

To study more generally the situation where these extreme estimates occur, for simplicity consider a case where all but one of the populations are completely fixed and that all population samples are of the same size  $n$ . The single unfixed population has some number  $x$  of an allele at a locus, with all other alleles observed in that population,  $n - x$ , being of a second type of allele. Assume also that all the remaining populations sampled at this locus are completely fixed for the second allele. It can then be shown that the MOME of  $\hat{\theta}_{il}$  for this situation is

$$\hat{\theta}_{il} = 1 - \frac{r(n - x)}{(n - 1)}$$

for the one unfixed population, and  $\hat{\theta}_{il} = 1.0$  otherwise. Since  $(n - x) \leq (n - 1)$ , the term  $r(n - x)/(n - 1) \leq r$ . In the most extreme case, if only a single allele of the first type is observed in the one unfixed population, and all other alleles observed are of type 2, that is, if  $x = 1$ , then the expression above simplifies to

$$\hat{\theta}_{il} = 1 - r,$$

which is clearly undesirable, because only for the situation where  $r = 2$  would the estimate be in the range of a correlation appropriate for the parameter we are estimating. As  $x$  increases, it approaches  $n$ , so that  $r(n - x)/(n - 1) \leq 1$ , and  $\hat{\theta}_{il}$  will then be within the range of a correlation between  $[-1, 1]$ .

On a practical level, a possible remedy for this might be to set some cut off point of polymorphism for loci to be included in an analysis. The effect of the introduction of several possible cutoff points on the method of moment estimates applied to this data set appears to be fairly strong, where the estimates increase about twofold as a result (Table 2.5). Comparing these population-specific estimators to the pairwise  $\theta_i$ 's given in Table 2.6, indicates that the pairwise estimates are

similar in magnitude to the MOM estimates obtained without any polymorphism criteria.

Table 2.5: Effects of incorporating different polymorphism criteria into estimation of the population-specific method of moment estimator of  $\hat{\theta}_i$  for the Planes data set. Loci were excluded from the analysis for a given criterion if the most frequent allele in a sample had a frequency greater than 0.95 or 0.99.

Population	Num. Loci Included			$\hat{\theta}_{i,MOM}$		
	No Cutoff	0.99	0.95	No Cutoff	0.99	0.95
GBR	18	14	7	0.346	0.633	0.709
Guam	18	15	10	0.295	0.610	0.685
Palau	18	14	7	0.169	0.557	0.671
Philip.	18	12	8	0.281	0.556	0.639
Hawaii	18	11	7	0.328	0.609	0.688
Oahu	18	13	9	0.152	0.528	0.602
Marq.	18	8	4	0.679	0.802	0.849
Fiji	18	13	7	0.222	0.532	0.641
New Cal.	18	12	7	0.300	0.621	0.704
Solo.	18	15	10	0.119	0.447	0.546
Bora-Bora	18	14	8	0.147	0.470	0.578
Clipp.	18	14	12	-0.173	0.319	0.420
Fanga.	18	12	9	-0.088	0.348	0.473
Moorea	18	13	9	0.127	0.457	0.554
Rang.	18	13	9	0.061	0.424	0.524
Mozam.	18	14	6	0.202	0.587	0.676

Table 2.6: Classic ANOVA estimates of  $\theta$  for pairs of populations in the Planes data set. Planes found that populations are grouped into five geographic regions within which populations are relatively unstructured. Populations show higher amounts of differentiation between the groups.

	GBR	Guam	Palau	Philip.	Hawaii	Oahu	Marq.	Fiji	New Cal.	Solo.	Bora	Clipp.	Fanga.	Moorea	Rang.	Mozam.
GBR	–	0.021	0.083	0.088	0.395	0.417	0.412	0.090	0.057	0.121	0.190	0.247	0.187	0.105	0.139	0.209
Guam		–	0.074	0.077	0.370	0.381	0.365	0.086	0.051	0.114	0.127	0.173	0.128	0.056	0.071	0.220
Palau			–	0.012	0.356	0.385	0.370	0.032	0.079	0.053	0.168	0.211	0.148	0.075	0.104	0.213
Philip.				–	0.391	0.416	0.394	0.042	0.112	0.068	0.201	0.244	0.183	0.083	0.122	0.255
Hawaii					–	0.096	0.263	0.371	0.343	0.366	0.318	0.367	0.278	0.294	0.304	0.371
Oahu						–	0.326	0.399	0.388	0.392	0.356	0.309	0.287	0.309	0.310	0.397
Marq.							–	0.380	0.326	0.382	0.314	0.375	0.260	0.239	0.264	0.413
Fiji								–	0.096	0.000	0.188	0.233	0.158	0.087	0.126	0.222
New Cal.									–	0.122	0.148	0.240	0.162	0.116	0.132	0.162
Solo.										–	0.185	0.224	0.154	0.112	0.133	0.222
Bora											–	0.125	0.064	0.118	0.049	0.248
Clipp.												–	0.067	0.128	0.065	0.284
Fanga.													–	0.051	0.016	0.235
Moorea														–	0.020	0.236
Rang.															–	0.244
Mozam.																–

Estimating  $\theta$  for pairs of populations as in Table 2.6 has been a popular way to try to determine essentially the value of  $\theta_i$  in populations of interest by using the classical ANOVA estimator of  $\theta$  that gives an average value across populations. A drawback to this approach compared to the use of the population-specific estimators evaluated here concerns the fact that  $\theta$  is a relative measure. The value estimated is relative to the amount of relatedness between the least related pair of alleles in the population hierarchy. The pairwise approach could be adversely affected if  $\hat{\theta}$  are compared between pairs of populations with different minimum levels of relationship. If instead population-specific estimates are obtained, they will all be based on the same level of relatedness, and thus can be reasonably compared. Regardless of this argument, the pairwise  $\hat{\theta}$  for the Planes data set are given here in order to indicate approximate levels of differentiation in order to compare the effects of low polymorphism on the MOME of  $\theta_i$  in Table 2.5.

## Discussion

This study has explored the sampling properties of two estimators for population-specific  $\theta_i$ . In doing so, we have shown that the MOME has a reasonably low bias and a sampling variance that can be decreased by increasing the number of loci sampled. Investigators must be cautious about the use of the MOME with loci in data sets with very low polymorphism. Increasing the number of polymorphic loci sampled should make this estimator more robust to polymorphism problems, in addition to having the side benefit of minimizing the sampling variance of the estimates. The maximum likelihood approaches based on normality considered for obtaining an estimator of population structure for the average over populations  $\theta$ , and population-specific  $\theta_i$ , gave undesirable estimates for both iterative and non-iterative approaches, and are not recommended for general use in analysis.

The bias of the MOME of  $\theta_i$  is relatively small in magnitude, and negative in direction. This result is consistent with previous theoretical results obtained from numerical approximations of Li (1996) and Jiang (1987), who worked with analytical approximations of the MOME of  $\hat{\theta}$  assuming large samples. The bias of the MOME of  $\theta_i$  increases positively as a side-effect of the large increase in the variance of this estimator that accompanies increased differentiation levels in a population, but was found to be unaffected by the number of loci sampled, the amount of linkage between loci, and unbalanced sampling.

Relative to the size of the sampling variance, which is quite large, as is typical of estimators of this type, the bias of the MOME is a negligible effect. The sampling variance of the MOME increases as differentiation increases, but is not affected by

unbalanced sampling. There is a strong effect of reduction of the sampling variance of the MOME as a result of increasing the number of loci sampled which is encouraging, but variances remain fairly large on the whole due to variance from genetic sampling occurring in populations that cannot be reduced by sampling design. Increasing the number of loci sampled is likely to have a stronger effect on reducing the sampling variance of the MOME than increasing the number of individuals sampled because increasing the number of individuals sampled increases the precision of the population allele frequency estimates, while increasing the number of loci sampled increases the precision of the estimate of population structure by providing more information about the genetic sampling process occurring in the populations. Therefore increasing the number of loci sampled will probably be a more desirable approach to maximizing limited genotyping resources for empirical studies.

The issue of obtaining extreme estimates exceeding the range of a correlation with the MOM due to data from loci with low polymorphism has been discussed from a practical perspective, and it appears to be adequate to use estimates of population-specific  $\theta_i$ , averaged over a reasonably large number of sampled loci. On a more theoretical level, the extreme estimates are a reminder that the parameter we are actually estimating is, in the notation of Weir and Hill (2002),

$$\hat{\beta}_i = \frac{\theta_i - \theta_A}{1 - \theta_i},$$

where  $\theta_A$  is the amount of covariance between the alleles among populations. The quantity estimated will be a function of both the association of alleles in individuals within populations and among populations, which is not necessarily a correlation

and therefore is not constrained by that particular range.

The combination of analysis of both simulated and real data for evaluation of the performance of estimators with complicated sampling distributions appears to have been validated again by this study, in that complexity present in the real data highlighted an issue with the performance of the estimator that was not observed in previous analyses. That the low polymorphism issue had not previously been detected in either simulations, or the analysis of a human data set presented in Weir and Hill (2002), was partly because, although the fish species the Planes data was collected from is highly polymorphic for a marine fish (Planes and Fauvelot, 2002), the set of loci routinely genotyped in human forensics are used specifically because of their high polymorphism in human populations, and effects of this nature should perhaps be taken into consideration in future studies of estimators used with natural population data.

Approximating sample allele frequencies with the multivariate normal distribution and proceeding with maximum likelihood estimation for either  $\theta$  or  $\theta_i$  estimators appears to be unreasonable for all approaches to estimating population structure examined here. The failure to consistently converge to estimate values within the possible parameter range of the correlation based statistic of both of the two proposed iteration methods for the MLE discussed is disheartening. Possibly future work with different numerical optimization procedures might be able to solve this problem of convergence and provide better estimates of  $\theta_i$ . On the other hand, numerical optimization for the MLE of  $\theta$ , where populations were constrained to have the same  $\theta$  values but sample sizes were allowed to vary did not produce any reasonable estimates although convergence of the optimization

occurred for every estimate. In summary, the unpredictable behavior of the iterative and non-iterative MLE suggests that the MOME is a much better choice to be used in analyses of this type.

It is possible that if allele frequencies were simulated with a multivariate normal distribution, instead of the individual-based approach used here, that the performance of the MLE of  $\theta_i$  would improve. It has been asserted by Nicholson and Donnelly (2002) that this distribution reasonably approximates the sampling distribution of allele frequencies for non-equilibrium populations. These type of populations may be characterized by shorter times since divergence or by low amounts of drift and migration which could increase the time needed to reach equilibrium (Tufto and Hindar, 1996). This assertion could be interesting to further pursue in future work.

## Literature Cited

- Balding, D. J. 2003. Likelihood-based inference for genetic correlation coefficients. *Theor. Popul. Biol.* 63:221–230.
- Casella, G. and Berger, R. L. 1990. *Statistical Inference*. Wadsworth Publishing Co., Belmont, CA.
- Holsinger, K. E. and Wallace, L. E. 2004. Bayesian approaches for the analysis of population genetic structure: an example from *Platanthera leucophaea* (*Orchidaceae*). *Mol. Ecol.* 13:887–894.
- Jiang, C.-J. 1987. Estimation of  $F$ -statistics in Subdivided Populations. PhD thesis, North Carolina State University, Raleigh, NC.
- Li, Y.-J. 1996. Characterizing the Structure of Genetic Populations. PhD thesis, North Carolina State University, Raleigh, NC.
- Long, J. C. 1986. The allelic correlation structure of Gainj- and Kalam-speaking people. I. The estimation and interpretation of Wright's  $F$ -statistics. *Genetics* 112:629–647.
- Long, J. C. and Kittles, R. A. 2003. Human genetic diversity and the nonexistence of biological races. *Hum. Biol.* 75:449–471.
- Maruyama, T. 1977. *Stochastic Problems in Population Genetics*. Lecture Notes in Biomathematics, Vol. 17. Springer, Berlin.

- Nagylaki, T. 1998. Fixation indices in subdivided populations. *Genetics* 148:1325–1332.
- Nei, M. 1973. Analysis of gene diversity in subdivided populations. *Proc. Natl. Acad. Sci. U.S.A.* 70:3321–3323.
- Nicholson, G. and Donnelly, P. 2002. Assessing population differentiation and isolation from single-nucleotide polymorphism data. *J. R. Statist. Soc. B* 64:1–21.
- Planes, S. and Fauvelot, C. 2002. Isolation by distance and vicariance drive genetic structure of a coral reef fish in the pacific ocean. *Evolution*. 56:378–399.
- Robertson, A. and Hill, W. 1984. Deviations from Hardy-Weinberg proportions: sampling variances and use in estimation of inbreeding coefficients. *Genetics* 107:703–718.
- Slatkin, M. 1995. A measure of population subdivision based on, microsatellite allele frequencies. *Genetics* 139:457–462.
- Smouse, P. and Williams, R. C. 1982. Multivariate analysis of HLA-disease associations. *Biometrics* 38:757–768.
- Tufto, J. Engen, S. and Hindar, K. 1996. Inferring patterns of migration from gene frequencies under equilibrium conditions. *Genetics* 144:1911–1921.
- Weir, B. S. 1996. *Genetic Data Analysis*. Sinauer, Sunderland, MA.

Weir, B. S. and Cockerham, C. C. 1984. Estimating  $F$ -statistics for the analysis of population structure. *Evolution* 38:1358–1370.

Weir, B. S. and Hill, W. G. 2002. Estimating  $F$ -statistics. *Ann. Rev. Gen.* 36:721–750.

Wright, S. 1931. Evolution in Mendelian populations. *Genetics* 16:97–159.

Wright, S. 1951. Genetical structure of populations. *Annals of Eugenics* 15:950–956.

# Appendix 1

## Iteration methods for the MLE of $\theta_i$

We have considered two iteration methods for obtaining the MLE of  $\theta_i$ . Neither method produced estimates that converged consistently to reasonable values within the possible parameter ranges of the quantities being estimated. Weir and Hill (2002) proposed the first iteration method we examined, suggesting iterating between estimating the  $\phi_i$ 's and the total population allele frequencies  $p_u$ , using Eqns. (2.3) and (2.4). This process was initialized by using the mean sample allele frequencies  $\bar{p}_u$  for the  $p_u$ . Alternatively, a second iteration method was proposed where iteration was between estimating the  $\phi_i$ s and the variance-covariance matrix  $\mathbf{\Omega}$ , as defined below.

For the second iteration method, we assume that  $\mathbf{p}$  and  $\mathbf{\Omega}$  are distinct parameters that will be estimated separately, and therefore substitute the mean of the sample allele frequencies  $\bar{\mathbf{p}}$  for  $\mathbf{p}$  in these equations as an intuitive estimate of the  $p_u$ s. If we have  $r$  independent populations sampled, the log-likelihood for the sample allele frequencies can be expressed equivalently to (2.7) as

$$\ln L \propto -\frac{r}{2} \ln(|\mathbf{\Omega}|) - \frac{m}{2} \sum_i \ln(\phi_i) - \frac{1}{2} \sum_i \frac{1}{\phi_i} (\tilde{\mathbf{p}}_i - \mathbf{p})' \mathbf{\Omega}^{-1} (\tilde{\mathbf{p}}_i - \mathbf{p}),$$

where

$$\mathbf{p} = \begin{bmatrix} p_1 \\ p_2 \\ \dots \\ p_{m-1} \end{bmatrix}$$

and

$$\mathbf{\Omega} = \begin{bmatrix} p_1(1-p_1) & -p_1p_2 & \cdots & -p_1p_{m-1} \\ -p_1p_2 & p_2(1-p_2) & \cdots & -p_2p_{m-1} \\ \cdots & \cdots & \cdots & \cdots \\ -p_1p_{m-1} & -p_2p_{m-1} & \cdots & p_{m-1}(1-p_{m-1}) \end{bmatrix}.$$

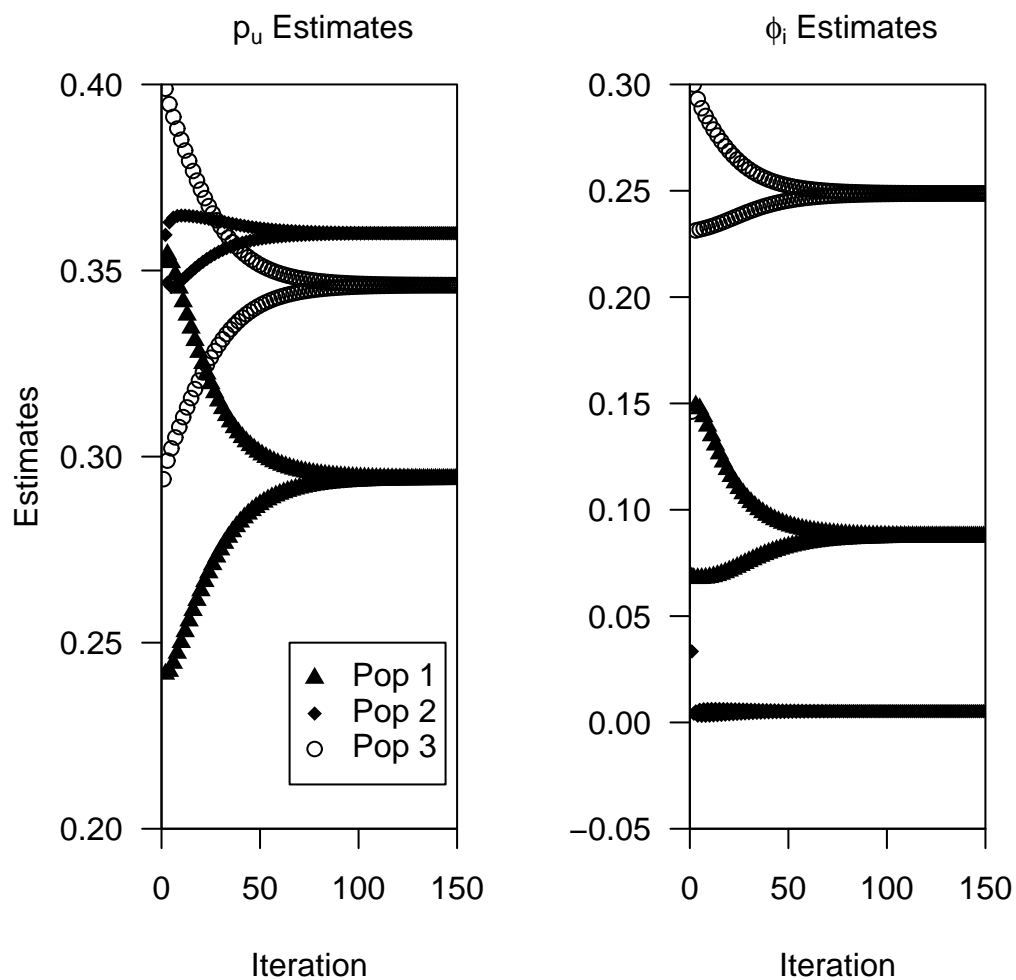
Taking the partial derivatives of this likelihood with respect to  $\phi_i$ ,  $\mathbf{p}$ , and  $\mathbf{\Omega}$ , and setting them equal to zero gives the following two equations which are iterated between for the second iterative method as

$$\hat{\phi}_i = \frac{1}{(m-1)}(\tilde{\mathbf{p}}_i - \bar{\mathbf{p}})' \mathbf{\Omega}^{-1}(\tilde{\mathbf{p}}_i - \bar{\mathbf{p}}) \quad (2.9)$$

$$\hat{\mathbf{\Omega}} = \frac{1}{r-1} \sum_i \frac{1}{\phi_i} (\tilde{\mathbf{p}}_i - \bar{\mathbf{p}})(\tilde{\mathbf{p}}_i - \bar{\mathbf{p}})'. \quad (2.10)$$

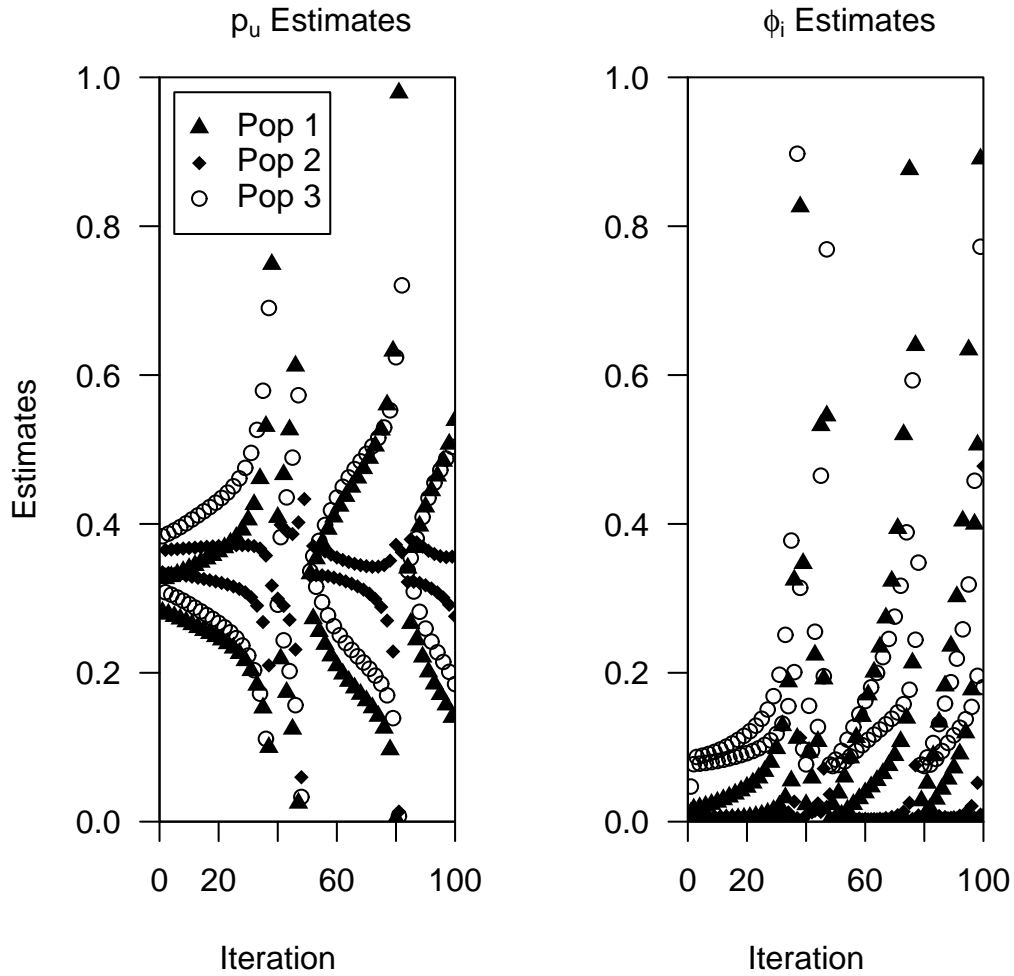
When the first iteration method was applied to the simulated data, the estimates of  $\phi_i$  for each simulation replicate fell into one of three categories: convergence to a single estimate value (roughly 10%), failure to converge due to periodicity of the estimates (about 5%), and the remaining replicate estimates failed to converge due to exceeding the possible range of the parameters  $\phi_i$  or  $p_u$ . These possible parameter ranges for the correlations  $\phi_i$  are (-1,1) and are (0,1) for the total population allele frequencies  $p_u$ . About 5% of the replicates produced estimates of  $\theta_i$  that stayed within the parameter ranges but did not converge, and instead assumed what looked to be a stable periodicity, alternating between a small number of estimate values with sequential iterations.

Figures 2.2, 2.3 and 2.4 give examples of replications whose estimates fell into each of these categories using three different replicates from a simulated data set. For the particular data set used to produce these figures, about 10% of the repli-



..

Figure 2.2: Values of maximum likelihood estimates of  $\hat{p}_u$  and  $\hat{\phi}_i$  for each iteration of a converged replicate of simulated data. The estimates for each of the three allele types  $u$  are shown on the left plot with different characters for each type, while the  $\hat{\phi}_i$  for each of the three sampled populations are shown on the right plot.



..

Figure 2.3: Values of maximum likelihood estimates of  $\hat{p}_u$  and  $\hat{\phi}_i$  for each iteration of a replicate of simulated data where the estimates exceeded the possible parameter range of the parameters they are estimating. Shown are only the estimates that remained within the possible parameter ranges. The iteration number is not reflective of the actual iteration number, but is sequential for the estimates within the ranges. Although the  $\phi_i$ s are correlations and so can possibly range from -1 to 1, in this case, no estimates between -1 and 0 were obtained so the plot of the  $\phi_i$ s do not show this region on its axis.

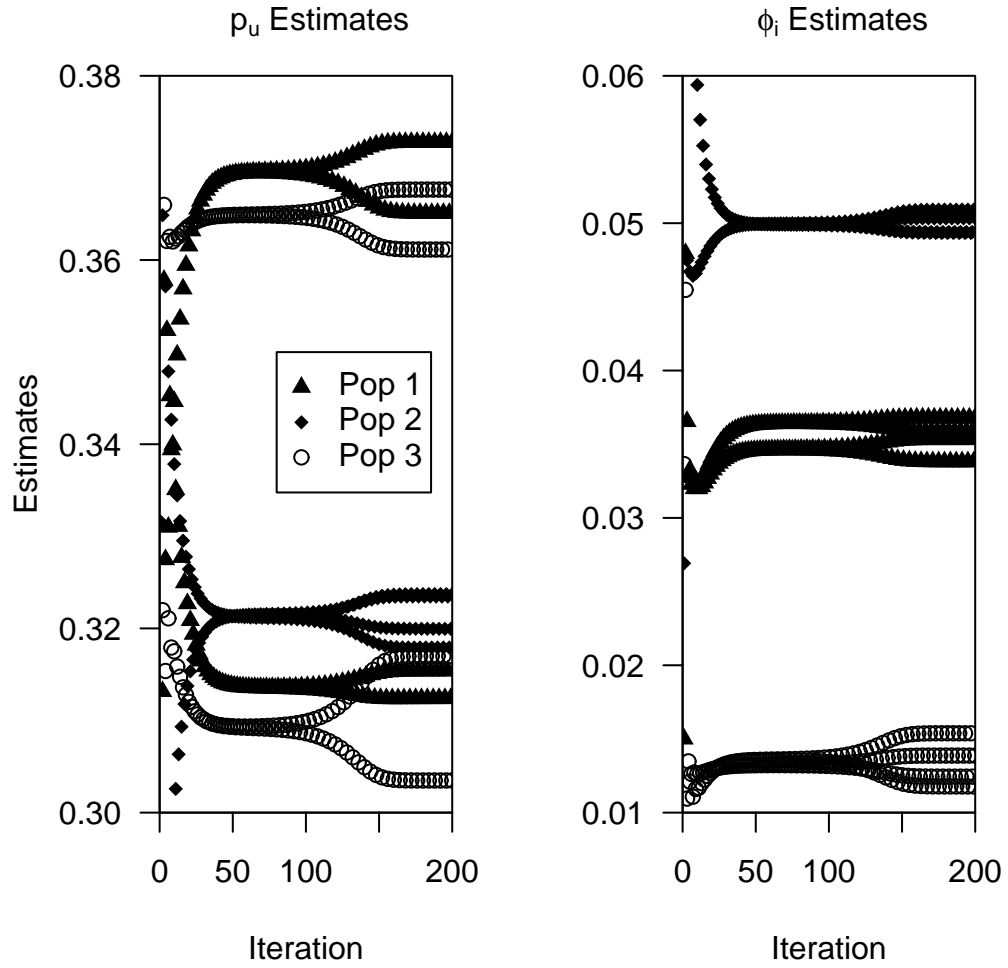


Figure 2.4: Maximum likelihood estimates of  $\hat{p}_u$  and  $\hat{\phi}_i$  for a case of simulated data where the estimates failed to converge to a single value, but remained within the possible parameter ranges.

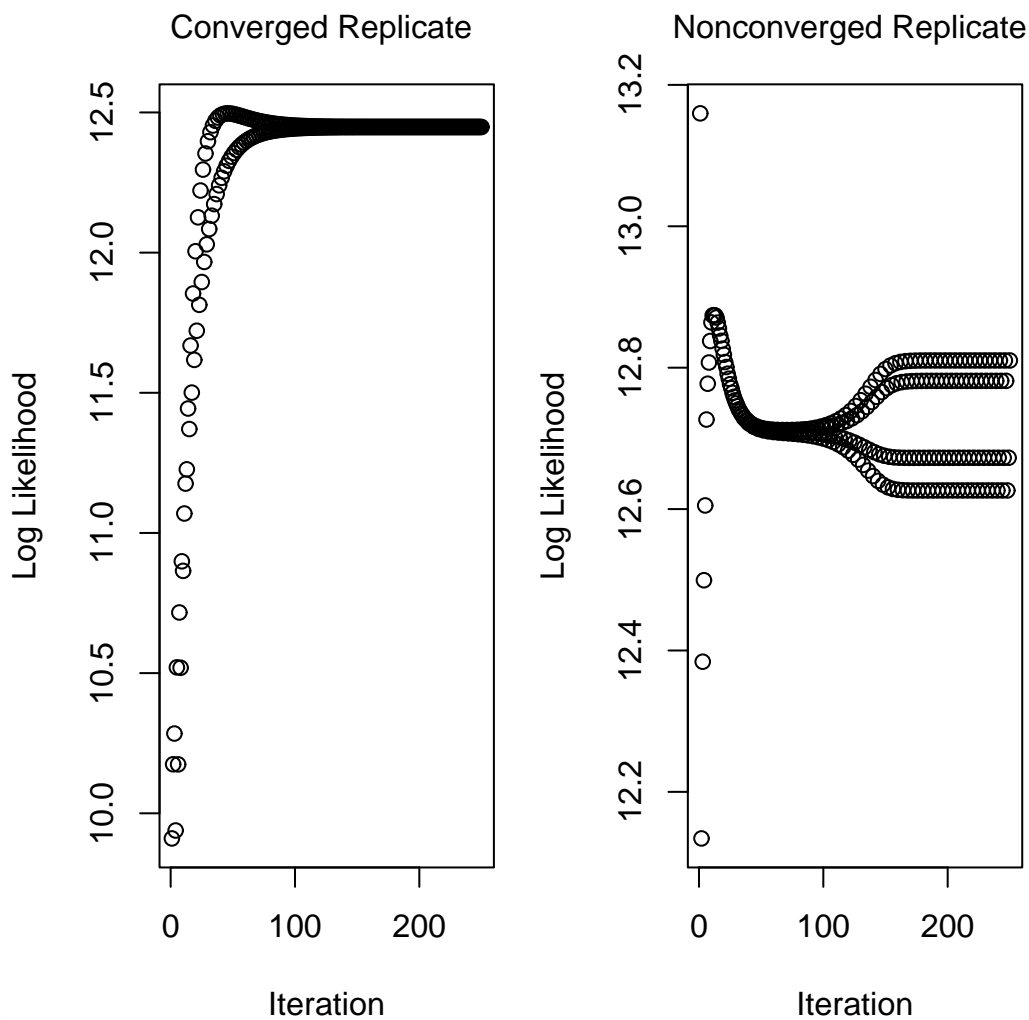


Figure 2.5: Change in the value of the log likelihood iterating between estimation of  $\hat{p}_u$  and  $\hat{\phi}_i$  using the first iteration method, for a converged replicate of simulated data and a nonconverged replicate that showed periodicity. The results of evaluating the log likelihood 2 using the  $\hat{p}_u$  and  $\hat{\phi}_i$  for each iterate are shown here. Not pictured is the log likelihood evaluated with extreme estimates since negative  $\phi_i$  values will have an undefined log value and cannot be used to evaluate the likelihood. No other negative  $\hat{\phi}_i$  values were observed in the other two cases shown here.

cates converged within 10,000 iterations, and most of the replicates that did converge did so within approximately the first 200 iterations. Figure 2.5 illustrates a summary of the previous figures by showing the effect of evaluating the likelihood for the observed  $\phi_i$  estimates and confirming that for the cases where the  $\hat{\phi}_i$  estimates are not converging, the likelihood values as well do not converge as the iterations continue.

The second iterative method studied was no more successful at providing desirable estimates of the maximum likelihood  $\phi_i$ s than the first. The use of two possible initial values to start the iterations was explored. These initial values resulted in different estimates for the first few iterations, but identical estimates for both initial values were obtained after this period. The first initial values used were the mean allele frequencies  $\bar{p}_u$  substituted into the equation for  $\phi_i$  to replace the  $p_u$ s in  $\Omega$  and  $\mathbf{p}$ . This resulted in obtaining estimates that were the same as the first iterate estimate of the  $\phi_i$ s of the first iteration method discussed here. Alternatively, using  $\phi_{i,0} = 1$  and  $\bar{\mathbf{p}}_u$ s was tried instead, resulting in beginning the iterations with  $\hat{\Omega}$ . The estimates of  $\phi_i$  converged quickly to the same value for all three populations sampled using these initial values as well, and then in subsequent iterations this value of the estimates decreased in all cases to 0.

## Appendix 2

### Effect of low polymorphism on the MLE of $\theta_i$

In addition to the low polymorphism behavior of the MOME discussed previously, the analysis of an empirical data set highlighted some additional undesirable be-

havior of the MLE that were not observed in the simulated data, and the need for practical procedures to deal with these cases. The case of undesirable estimates with the MLE also occurs when loci that are completely fixed in a particular population are not fixed in other populations sampled. In this case the situation is reversed, in that the problem lies in the ML estimates of  $\theta_{il}$  for the fixed populations, rather than in the estimates for the single unfixed population, as with the MOME. For the MLE, where we would expect the estimate in the fixed population to be  $\hat{\theta}_{il} = 1.0$ , as is true of the MOM, we do not obtain this value if we use the first iterate estimator of Eqs. (2.3) and (2.5). An instance of this problem can be seen in Table 2.7 at locus *Aat-2* in the Hawaiian population.

The value of  $\hat{\theta}_{il,MLE}$  is not 1.0 for this case due to the fact that the numerator of  $\hat{\phi}_{il}$  which has the form

$$\hat{\phi}_{il} = \frac{1}{(m_l - 1)} \sum_{u=1}^{m_l} \frac{(\tilde{p}_{ilu} - p_{lu})^2}{p_{lu}},$$

will be non-zero. This is because  $\bar{p}_{lu} \neq \tilde{p}_{ilu} = 1.0$  for the allele  $u$  observed in the fixed population, since  $\bar{p}_{lu}$  is an average over all populations sampled and not all populations were fixed for this allele. In contrast with the MLE, for this case  $\theta_{il,MOM} = 1.0$ , because the numerator  $x_{il}$  from (2.1) is just a function of  $\tilde{p}_{ilu}$  values, which are 1.0 for the fixed population.

To more explicitly describe the behavior of the MLE, for the case where  $\bar{p}_{lu} < 1/m_l$ ,  $\hat{\theta}_{il} > 1.0$ , while if  $1/m_l < \bar{p}_{lu}$  is true,  $\hat{\theta}_{il}$  will be bounded by zero and one. Therefore as a special case, we could require that in this situation of a population fixed at a particular locus,  $\hat{\theta}_{il} = 1.0$  for the unfixed population, and no further calculation is necessary.

Table 2.7: Estimates of  $\hat{\theta}_{il}$  from a subset of loci with very low polymorphism of the Planes data set. Those estimates where the MOME is 1.0 reflects fixation of the population at that locus. Population and loci combinations where the MOME is not 1.0 were not fixed.

Locus	$\hat{\theta}_{il}$	GBR	Guam	Palau	Philip.	Hawaii	Oahu	Marq.	Fiji
<i>G3pdh</i>	MOM	1.000	1.000	1.000	1.000	-2.220	1.000	1.000	-3.887
	ML	-0.009	-0.009	-0.008	-0.009	0.017	-0.009	-0.009	0.073
<i>Idhp-2</i>	MOM	1.000	1.000	-7.058	1.000	1.000	1.000	1.000	1.000
	ML	-0.008	-0.008	0.322	-0.008	-0.008	-0.008	-0.008	-0.008
<i>Ldh-2</i>	MOM	1.000	-16.793	1.000	1.000	1.000	1.000	1.000	1.000
	ML	-0.012	0.193	-0.012	-0.013	-0.012	-0.012	-0.012	-0.012
<i>Mep-2</i>	MOM	1.000	0.111	0.133	1.000	1.000	1.000	1.000	1.000
	ML	-0.008	0.022	-0.009	-0.009	-0.008	-0.008	-0.008	-0.008

Table 2.8: Table 2.7 Continued

Locus	$\hat{\theta}_{il}$	New Cal.	Solo.	Bora	Clipp.	Fanga.	Moorea	Rang.	Mozam.
<i>G3pdh</i>	MOM	1.000	-0.673	-2.911	1.000	1.000	1.000	1.000	-2.220
	ML	-0.000	-0.009	0.039	-0.011	-0.008	-0.009	-0.008	0.196
<i>Idhp-2</i>	MOM	1.000	1.000	1.000	-7.105	1.000	1.000	1.000	-0.274
	ML	0.001	-0.008	-0.005	0.323	-0.007	-0.008	-0.007	0.095
<i>Ldh-2</i>	MOM	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000
	ML	-0.003	-0.012	-0.010	-0.014	-0.012	-0.012	-0.011	-0.012
<i>Mep-2</i>	MOM	1.000	0.088	0.274	-11.972	1.000	0.111	1.000	1.000
	ML	0.001	-0.009	0.013	1.153	-0.008	-0.009	-0.007	-0.008

## Chapter 3

# THE VARIANCE OF SAMPLE HETEROZYGOSITY AND GENE DIVERSITY

Johnson AM and BS Weir

## Abstract

The properties of several methods for obtaining the variances of sample heterozygosity or gene diversity are studied in order to evaluate different procedures for calculating these quantities. This work is motivated primarily by the frequently overlooked importance of reporting the sampling properties of estimators in studies of the genetic variation of populations. Analysis of empirical data sets presented here demonstrates that dependencies between gene diversities or heterozygosities at different loci can lead to very large differences between variances determined by an approach accounting for the stochastic process of the evolution of populations and a commonly used approximation of the variances. Approximating the total variance of sample gene diversity by the total variance of sample heterozygosity is considered in order to reduce the computational complexity of this term, but simulation shows this to be a poor approximation. Different results with unbalanced data for the total variance of sample heterozygosity are obtained with four variance component methods, as expected by statistical theory. Additionally, the likelihood based methods considered here are shown to be robust to violations of assumptions of normality, even for very small sample sizes. Finally, variance obtained from a previously proposed mixed model are compared to a new random effects model. Large differences are observed for a case where there is a large variance component due to loci.

## Introduction

The statistics of gene diversity and heterozygosity are basic tools for summarizing the patterns of genetic variation in a group of populations. Characteristics of population genetic variation are of key interest in studies of evolution, and for commercial and conservational breeding programs that seek to develop and maintain desirable variation in an efficient way (Grenier et al., 2000). At the most basic level, this interest lies in the fact that the amount of variation present in a population or species determines the capacity that group has for heritable change. While the statistics of gene diversity and heterozygosity can be very useful descriptive measures for populations, a number of studies published each year reporting estimates of these parameters fail to also report variances of their estimates. For the descriptive purposes of these studies, it would be better to consider and report the sampling properties of the estimators in order to increase the quality of resulting inferences. In this paper we examine several approaches to obtaining the variance of sample heterozygosity and gene diversity in order to clarify how best to determine these variances, and to illustrate their importance.

The measure of genetic variation generally called gene diversity was first described by Marshall and Allard (1970) as a term called the polymorphic index. Subsequent work for determining variances of gene diversity was published by Nei (1973) and Nei and Roychoudhury (1974), and it is in this work that the name ‘gene diversity’ for this measure was first used. Their approach focused on a single random-mating population and so did not consider variation due to differentiation between populations. Instead, they chose to work with genetic distance measures

to account for variation between pairs of populations. This variation can alternatively be summarized as the total variance of heterozygosity or gene diversity.

Weir (1989) and Weir et al. (1990) developed extensive theory for the variances of sample gene diversity and heterozygosity, respectively. The development in these papers emphasized that appropriate expressions for these variances are dependent upon the scope of inference that an investigator wishes to make. The total population scope is used when making inferences about a larger group of populations from which the populations in the data set were randomly sampled. For this scope, the evolutionary history of the populations must be accounted for by including dependencies between individuals within populations. In contrast, the within-population scope is for the case where inferences are to be made only about the specific populations sampled and individuals can be regarded as being independent. For both the within and total scopes, Weir (1989) and Weir et al. (1990) showed how to properly account for dependencies between loci.

The purpose of this paper is to advocate the regular inclusion and consideration of the variances of the estimators of gene diversity and heterozygosity in studies that estimate these quantities. For the years 2001-2003, the journals *Evolution* and *Journal of Evolutionary Biology*, two journals that often publish population surveys, were examined to determine the frequency of studies reporting estimates of either gene diversity or heterozygosity. Generally, one to two papers per issue presented estimates of either of these two statistics, but only around 35% of *Evolution* papers and 50% of such papers in the *Journal of Evolutionary Biology*, also reported variances with their point estimates.

Because of the inclusion of the effects of the evolutionary history and the abil-

ity to make inferences about larger groups such as species, the total population scope is likely to have more biological relevance for making inferences about the genetic variation of structured populations. However, of those studies that gave variances of their estimates, typically only the variance due to the effects of statistical sampling was reported, with no accounting for dependencies in the data due to the evolutionary history of the sampled populations and loci. Heterozygosity estimates presented on the web site of the HapMap project are indicative of this general practice (<http://www.ncbi.nlm.nih.gov/SNP/Hetfreq.html>).

Recently Shete (2003) derived a uniformly minimum variance unbiased estimator (UMVUE) of gene diversity by correcting the bias of the classical estimator. Shete suggested a bootstrap procedure to estimate the variance of sample gene diversity for single populations at multiple loci, which is equivalent to what we term here the within-population variance. Resampling approaches are not recommended for estimating the total variance because resampling over populations disrupts dependencies due to common ancestry within populations, and has thus been shown to fail to approximate the sampling distribution of allele frequencies (Dodds, 1986; Weir, 1996). Alternatively, it would be difficult to obtain an exact expression for the total variance of the UMVUE of sample gene diversity, because of the complexity of this expression, which is reduced for the within-population variance by the assumption of independence of alleles between individuals. However, the UMVUE estimator should give similar values to the classical estimator of gene diversity for large samples.

Since many studies that report estimates of gene diversity and heterozygosity are surveys of natural populations, data sets are frequently unbalanced, with differ-

ent numbers of individuals observed in different populations according to the availability of individuals that can be genotyped during data collection. Unbalanced data sets introduce difficulties in determining analytically the statistical properties of sample variances, particularly for ANOVA estimates of variances. While for balanced data, the familiar ANOVA procedure gives estimates of variance components with desirable and known statistical properties such as uniform best unbiasedness, problems for ANOVA estimates with unbalanced data include the lack of a unique set of sums of squares that are uniformly best, and no known general analytic properties of different types of sums of squares for adequate comparisons (Searle et al., 1992). For these reasons, the methods of MIVQUE (minimum variance quadratic unbiased estimation), ML (maximum likelihood) estimation and REML (restricted maximum likelihood), have been advocated by (Rao and Kleffe, 1988) and (Searle et al., 1992), among others, for analyses with unbalanced data. Implementations of these methods are available in many statistical software packages, including Proc Mixed in SAS.

In addition to generally motivating the need for the consideration of the sampling properties of heterozygosity and gene diversity estimators, a number of issues related to determining the best procedures for obtaining these variances are addressed here. These include the illustration of concerns with a commonly used approximation of the total variance of sample gene diversity and heterozygosity by the re-analysis of two previously published data sets (Nason et al., 2002; Olsen and Schaal, 2002). Motivated by a need to reduce the computational complexity of the total variance of sample gene diversity, population simulation is used as well to evaluate the approximation for this term by the total variance of sample

heterozygosity. Throughout, it will be demonstrated that failure to account for sources of variation in populations could strongly affect the quality of inferences that can be made in an analysis. Differences between random and mixed effects models for estimating the total variance of sample heterozygosity will be discussed, and the relative merits of four variance component methods used for the case of unbalanced data will be used to obtain variance estimates of statistics of interest.

# Theory

## Point Estimators of Gene Diversity and Heterozygosity

Sample heterozygosity is simply the observed frequency of heterozygotes in a data set. While often denoted as  $H_o$ , for this paper the term  $H_i$  will be used instead to mean the proportion of heterozygotes in the  $i$ th population sampled. Properties of estimates of  $H_i$  are most easily derived using indicator variables. Let  $x_{ijl}$  take the value of 1 if the  $j$ th individual in a sample from the  $i$ th population is a heterozygote at locus  $l$ , and 0 otherwise. We use the terms  $\tilde{H}_{il}$  and  $\tilde{H}_i$  for the sample values of heterozygosity for a given population and locus, and over all loci  $L$ , respectively. These can be written as functions of the indicator variables:

$$\tilde{H}_{il} = \frac{\sum_{j=1}^{n_i} x_{ijl}}{n_{il}}, \quad \tilde{H}_i = \frac{\sum_{l=1}^L \sum_{j=1}^{n_i} x_{ijl}}{\sum_{l=1}^L n_{il}},$$

where  $n_{il}$  is the size of the sample from population  $i$  at locus  $l$ . Note that the symbol  $\sim$  is used here to distinguish sample values from the population values of the quantities under consideration.

Gene diversity, often called expected heterozygosity and written as  $H_e$ , is the frequency of heterozygotes expected for a population with Hardy-Weinberg genotype proportions. Rather than using this terminology, we will instead use the notation  $d_i$  for the parameter of gene diversity. This is done in order to remove any confusion between the quantity “expected heterozygosity” and the expectation of sample heterozygosity  $E(\tilde{H}_i)$ . Using subscripts as before, the gene diversity for

a single locus and population, and for multiple loci can be expressed as

$$\tilde{d}_{il} = 1 - \sum_{u=1}^{m_l} \tilde{p}_{ilu}^2, \quad \tilde{d}_i = \frac{\sum_{l=1}^L n_{il} \tilde{d}_{il}}{\sum_{l=1}^L n_{il}}, \quad (3.1)$$

where  $\tilde{p}_{ilu}$  is the frequency of allele  $u$  at locus  $l$  in a sample from population  $i$ , and  $m_l$  is the number of alleles at locus  $l$ .

Gene diversity and heterozygosity are both measures that summarize the amount and distribution of genetic variation found in populations, but they differ in that gene diversity quantifies variation at the allelic level, while heterozygosity summarizes variation at the genotypic level. The simplicity of the measure of heterozygosity as the observed proportion of heterozygotes in a sample makes its usage both desirable and straightforward in many studies. However, the proportion of heterozygotes can fail to capture the true extent of genetic variation in populations with high amounts of selfing or asexual reproduction, such as is found in many plant species and simpler organisms. Alleles tend to associate within individuals in a mixed mating system that includes some percentage of selfing, so that there can be many types of alleles present in a population, but very few heterozygous individuals. In this case gene diversity would be the more appropriate measure to use.

## Variations, Models and Scope of Inference

The general expression for the variance of both  $\tilde{d}_i$  and  $\tilde{H}_i$ , independent of within or total-population scope, includes the variance of the population estimates at single loci,  $\tilde{d}_{il}$  or  $\tilde{H}_{il}$ , and the covariance between these estimates. As an example of this,

the within-population variance of sample heterozygosity can be written as

$$\text{Var}_W(\tilde{H}_i) = \frac{1}{L^2} \left[ \sum_{l=1}^L \text{Var}_W(\tilde{H}_{il}) + \sum_{l=1}^L \sum_{l' \neq l}^L \text{Cov}_W(\tilde{H}_{il}, \tilde{H}_{il'}) \right]. \quad (3.2)$$

This value could be estimated through bootstrapping the data from the population of interest as suggested by Shete (2003), or by using sample values for the quantities in Eq. (3.2) giving the exact expression

$$\text{Var}_W(\tilde{H}_i) = \frac{1}{L^2} \left[ \sum_{l=1}^L \frac{\tilde{H}_{il}(1 - \tilde{H}_{il})}{n_{il}} + \sum_{l=1}^L \sum_{l' \neq l}^L \frac{(\tilde{H}_{ill'} - \tilde{H}_{il}\tilde{H}_{il'})}{n_{ill'}} \right], \quad (3.3)$$

where  $\tilde{H}_{ill'}$  is the observed proportion of individuals that are heterozygous at loci  $l$  and  $l'$ , and  $n_{ill'}$  is the number of individuals genotyped at both the  $l$  and  $l'$  loci.

For the variance of sample gene diversity the corresponding estimate is

$$\begin{aligned} \text{Var}_W(\tilde{d}_i) \hat{=} & \frac{2}{n_i L^2} \left\{ \sum_{l=1}^L \left[ \sum_{u=1}^{m_l} (\tilde{p}_{ilu}^3 + \tilde{p}_{ilu}^2 \tilde{P}_{iluu} - 2\tilde{p}_{ilu}^4) \right. \right. \\ & \left. \left. + \frac{1}{2} \sum_{u=1}^{m_l} \sum_{u' \neq u}^{m_l} \tilde{p}_{ilu} \tilde{p}_{il'u'} (\tilde{P}_{iluu'} - 4\tilde{p}_{ilu} \tilde{p}_{il'u'}) \right] \right. \\ & \left. + \sum_{l=1}^L \sum_{l' \neq l}^L \sum_{u=1}^{m_l} \sum_{v=1}^{m_{l'}} \tilde{p}_{ilu} \tilde{p}_{il'v} \tilde{\Delta}_{ilul'v} \right\}, \quad (3.4) \end{aligned}$$

where  $\tilde{P}_{iluu}$  and  $\tilde{P}_{iluu'}$  are the sample frequencies of the  $uu$  homozygotes and  $uu'$  heterozygotes at locus  $l$ . The term  $\tilde{\Delta}_{ilul'v}$  is the sample composite linkage disequilibrium coefficient for a pair of loci and can be calculated as

$$\tilde{\Delta}_{ilul'v} = \frac{n_{ilul'v}}{n_i} - 2\tilde{p}_{ilu}\tilde{p}_{il'v}.$$

Weir et al. (1990) showed that a linear models approach can be used to obtain estimates of the total variance of sample heterozygosity. The linear model used in this treatment was

$$x_{ijl} = \alpha_i + \beta_{ij} + \gamma_l + (\alpha\gamma)_{il} + (\beta\gamma)_{ijl}. \quad (3.5)$$

From this, a mixed model approach was applied with loci as a fixed effect and all other effects random. The variance components for a total-population scope are then

$$\begin{aligned}
\text{Var}_T(\alpha_i) &= \sigma_p^2 && \text{for populations,} \\
\text{Var}_T(\beta_{ij}) &= \sigma_{i(p)}^2 && \text{for individuals within populations,} \\
\text{Var}_T[(\alpha\gamma)_{il}] &= \sigma_{lp}^2 && \text{for loci by populations and} \\
\text{Var}_T[(\beta\gamma)_{ijl}] &= \sigma_{i(p)}^2 && \text{for loci by individuals within populations.}
\end{aligned} \tag{3.6}$$

By treating the loci effect as fixed, we limit the scope of the inferences to be made to the specific loci sampled for the study. In Weir et al. (1990), it was argued that this is appropriate when the same loci are sampled for each population. In this paper, we also consider a fully random model where the locus effects are random as well. This would have the effect of adding the variance component  $\text{Var}_T(\gamma_l) = \sigma_l^2$ , to those listed in (3.6) and would allow inferences to be made about the distribution of genetic variation across the whole genome of the species under study.

The total variance of sample heterozygosity can be expressed for the mixed model as

$$\begin{aligned}
\text{Var}_T(\tilde{H}_i) &= \frac{1}{\left(\sum_{l=1}^L n_{il}\right)^2} \left[ \sigma_p^2 \left( \sum_{l=1}^L n_{il}^2 + \sum_{l=1}^L \sum_{l' \neq l} n_{ill'} \right) \right. \\
&\quad + \sigma_{i(p)}^2 \left( \sum_{l=1}^L n_{il}^2 + \sum_{l=1}^L \sum_{l' \neq l} n_{ill'} \right) \\
&\quad \left. + \sigma_{pl}^2 \left( \sum_{l=1}^L n_{il}^2 \right) + \sigma_{i(p)l}^2 \left( \sum_{l=1}^L n_{il} \right) \right], \tag{3.7}
\end{aligned}$$

$$\text{Var}_T(\tilde{H}_{il}) = \sigma_{p/l}^2 + \frac{\sigma_{i(p)l}^2}{n_{il}}, \tag{3.8}$$

where the terms in the second expression are defined as

$$\sigma_{p/l}^2 = \sigma_p^2 + \sigma_{lp}^2 \quad (3.9)$$

$$\sigma_{i(p)/l}^2 = \sigma_{i(p)}^2 + \sigma_{li(p)}^2. \quad (3.10)$$

For the random model, the total variance is the same as Eqn. (3.7) except that the third term includes the variance component for loci, which can be written as

$$(\sigma_l^2 + \sigma_{pl}^2) \left( \sum_{l=1}^L n_{il}^2 \right).$$

The total variance of  $\tilde{H}_{il}$  for the random model is as in Eq. (3.8), but Eq. (3.9) is instead

$$\sigma_{p/l}^2 = \sigma_p^2 + \sigma_l^2 + \sigma_{lp}^2.$$

The expression for the total variance of sample gene diversity given by Weir (1989) is complicated due to the nature of the moments of an estimator that is in itself a quadratic term. The variance of necessity involves terms of frequencies of two, three and four alleles at one and two loci and does not lend itself easily to a linear models approach. Accounting for all the departures from independence of these frequencies with descent measures becomes extremely cumbersome. In contrast, the expression for the total variance of sample heterozygosity is simplified in that we can use indicator variables at the genotype level, rather than using indicators at the allelic level as would be necessary with gene diversity. The use of an approximation of  $\text{Var}_T(d_i)$  by  $\text{Var}_T(\tilde{H}_i)$  would have the advantage in that readily available software could be used to estimate  $\text{Var}_T(\tilde{H}_i)$  with a linear models approach.

A separate issue for  $\text{Var}_T(\tilde{H}_i)$  is that it is necessary to have data from multiple populations to estimate the population variance component  $\sigma_p^2$ . In the absence of multiple population data, it has often been suggested that data from individual independent loci be used to approximate the genetical sampling that occurs between populations. This commonly used approach, which we will refer to in the rest of this paper as the single-locus approximation, can be written as the average of the variances of the single-locus heterozygosities

$$\text{Var}_T(\tilde{H}_i) \approx s_{\tilde{H}_i}^2 = \frac{\sum_{l=1}^L (\tilde{H}_{il} - \tilde{H}_i)^2}{L(L-1)}. \quad (3.11)$$

By taking the expectation of the single-locus approximation, it can be seen that this approximation fails to allow for dependencies in the data that are accounted for in the underlying population model used to develop the exact expressions for  $\text{Var}_T(\tilde{H}_i)$ . The expectation taken in the total context as given by Weir et al. (1990) is

$$E_T \left( s_{\tilde{H}_i}^2 \right) = \text{Var}_T(\tilde{H}_i) + \sum_{l=1}^L (H_l - H)^2 - \frac{\sum_{l=1}^L \sum_{l' \neq l} \text{Cov}_T(\tilde{H}_{il}, \tilde{H}_{il'})}{nL(L-1)}, \quad (3.12)$$

and a similar expectation can be obtained for the within-population case. This approximation can be seen to be biased by the variances among single locus heterozygosities and by the covariances between heterozygosities at different loci. Because of this, the single-locus approximation given in Eqn. (3.11) is not a reasonable approximation of the total variance of sample heterozygosity unless heterozygosities can be assumed to be equal and independent.

Similarly, a single-locus approximation of the variance of sample gene diversity

has been suggested where

$$\text{Var}_T(\tilde{d}_i) \approx s_{\tilde{d}_i}^2 = \frac{\sum_{l=1}^L (\tilde{d}_{il} - \tilde{d}_i)^2}{L(L-1)}. \quad (3.13)$$

Here the expectation of the single-locus approximation in the total scope can be written in the same form as Eq. (3.12)

$$\begin{aligned} \text{E}_T \left( s_{\tilde{d}_i}^2 \right) &= \text{Var}_T(\tilde{d}_i) \\ &+ \frac{1}{L(L-1)} \left[ \sum_{l=1}^L \left( \text{E}_T(\tilde{d}_{il}) - \frac{1}{L} \sum_{l=1}^L \text{E}_T(\tilde{d}_{il}) \right)^2 \right. \\ &\left. - \sum_{l=1}^L \sum_{l' \neq l} \text{Cov}_T(\tilde{d}_{il}, \tilde{d}_{il'}) \right], \end{aligned} \quad (3.14)$$

where the covariance of population sample gene diversities at different loci for both scopes depends on the composite linkage disequilibrium coefficients for a pair of loci,

$$\text{Cov}_T(\tilde{d}_{il}, \tilde{d}_{il'}) = \frac{2}{n_i} \sum_{u=1}^{m_l} \sum_{v=1}^{m_{l'}} p_u p_v \Delta_{lul'v}.$$

This suggests that one might reasonably test the hypothesis  $H_0 : \Delta_{ilul'v} = 0$ , to explore the appropriateness of the use of the single-locus approximation for the variance of sample gene diversity (Weir, 1989). In this paper we will demonstrate the effectiveness of this testing process for several data sets and discuss better ways to infer the total variance of sample gene diversity or heterozygosity that account for dependencies between the data caused by evolutionary and demographic forces acting on populations.

## Variance Component Methods

For balanced data, variances estimated by the MIVQUE, ML, REML and ANOVA methods should be identical to each other, but are expected to differ for unbalanced data. These differences arise from the underlying differences in the methodology of the four variance components methods and the assumptions made about the form of the data. The related methods of REML and ML, are based on assuming the full form of the probability distribution of the data. In contrast, ANOVA requires a less restrictive assumption of the form of the first two moments of this distribution, and the MIVQUE approach requires no assumptions about the form of the distribution of the data at all.

For the maximum likelihood based methods, the distribution assumed is generally the multivariate normal, due to the mathematical tractability of this distribution (Searle et al., 1992). Estimates of the parameters are found by an iterative process of maximizing the joint probability of the likelihood for the model parameters, given the observed data. REML is probably to be preferred to the ML approach in that it attempts to remedy drawbacks of the ML method. These drawbacks include that variance estimates obtained through the ML approach are not guaranteed the property of minimized variance, as with REML and MIVQUE, nor does ML account for the degrees of freedom of any fixed effects although REML does this. Additionally, REML is desirable in that it constrains the variance estimates to be non-negative. Negative estimates can be an issue with the MIVQUE procedure and might be difficult to interpret or explain for the case of a variance.

MIVQUE estimators are determined by estimation equations that were devel-

oped based on criteria of Rao (1971). MIVQUE estimators are unbiased, translation invariant and have a minimized variance. The MIVQUE approach, which does not require the normality assumption of the likelihood-based approaches, might be more desirable in this sense for genetic applications, but does have the potential drawback that negative variance estimates could be obtained, particularly in situations when the true value of the variance is very close to zero (Searle et al., 1992). A useful quality for applications to biological data is that MIVQUE does not rely on any distributional assumption of the data. This is desirable because the multivariate normal may not apply to marker data which typically fall into a few discrete classes and would generally be better described as being multinomially distributed. In any case, the maximum likelihood methods tend to be robust about violations of their assumptions and we will examine in this paper which of these methods will be acceptable in genetic variation analyses.

## Materials and Methods

Both simulation and analysis of empirical data sets were employed to study the effect of the different methods and models of interest on the variance of sample heterozygosity and gene diversity. The two previously published data sets (Nason et al., 2002; Olsen and Schaal, 2002) studied here were re-analyzed using software written in *R* and Perl, excepting the variance components methods used to estimate the total variance of sample heterozygosity obtained from SAS using Proc Mixed. Populations were simulated with software written in *R*.

The first empirical data set was collected by Nason et al. (2002), who studied the genetic variation present in a group of phytophagous (plant dwelling and eating) goldenrod elliptical-gall moths. Groups of moths within the larger group live on two different types of goldenrod plant hosts that frequently grow together, and these were studied in order to determine if sympatric speciation was occurring. These plant hosts are of the *Solidago canadensis* complex and are named *S. altissima* and *S. gigantea*. The moths were genotyped at 12 polymorphic allozyme loci and collected from both types of plant hosts at 4 geographically separate sites: Bogus Brook (BB), Crystal Lake (CL), Cone Marsh (CM) and Zimmerman Field (ZF). Mean population sample sizes  $\bar{n}_i$  were obtained by averaging the sample sizes at each locus over all loci for each population. These values vary due to a small amount of missing data for some individuals at particular loci.

The other data set was collected by Olsen and Schaal (2002), who genotyped 5 microsatellite loci in 27 populations of the plants *Manihot esculenta* ssp. *flabellifolia*, for a total of 157 individuals sampled. These plants which have been

described as the “wild progenitor of the root crop cassava”, are typically found in very small populations of less than fifteen individuals. The microsatellites were located in multiple introns of a 962-base-pair sequence of the *Glyceraldehyde 3-phosphate dehydrogenase* gene. These populations were further pooled according to geographic relationships into five groups in order to study the effects of increasing the departure from balanced sampling on the variance component method results. The five pooled groups were as follows: Tocantins included the populations Axixá, Luzinópolis, Miranorte and Dueré. The group Goiás included the populations Campos Belos, Campinorte, Rialma, Corumbá, Nerópolis, Goiás Velho, Iporá and Caiapônia. Mato Grosso was composed of Nova Xavatina, Serra Petrovina, Santa Elvira, São Vincente, Lambari d'Oeste, Pontes e Lacerda-A and -B. The group Rondônia included Vilhena, Pimenta Bueno, Jarú, Ariquemes, Teotônio, Taquaras. Finally, the group Acre was composed of the Rio Branco and Sena Madureira populations.

To test the approximation of the variance of sample gene diversity by the variance of sample heterozygosity, simulations of a single population with 1000 individuals and a single locus with two alleles were performed. The populations were created from a specified range of gene diversity and inbreeding levels which determined the population allele frequencies and genotype frequencies. A random sample of 100 individuals without replacement was then taken from the simulated population, and from this sample values of heterozygosity, gene diversity and their associated variances were calculated. For each simulated population 1000 replicate samples were generated and the results averaged over these replicate values.

## Results

We advocate the regular inclusion of the variance of estimators in statistical analyses, particularly for the estimators of heterozygosity and gene diversity, which are routinely reported solely as point estimates in population surveys. It is important to include estimates of variance in analyses of this type, because the variances can be quite large and because the inclusion of the variances more completely summarizes the data of interest, often a main goal of descriptive survey studies. Examination of the point estimates  $\tilde{d}_i$  and  $\tilde{H}_i$  in Table 3.1 illustrates the benefits of such a summarization. For this data set both groups of *altissima* and *gigantea* dwelling insects overall have high levels of genetic variation which is conveyed by the point estimates. However, the estimates for both heterozygosity and gene diversity range a great deal across both loci and populations, with underlying  $\tilde{d}_{il}$  and  $\tilde{H}_{il}$  estimates ranging from (0.0, 0.72) for both species groups and measures. This range is best summarized by including the variance of sample estimates with the point estimates for this data.

### **Approximation of the total variance of sample diversity and heterozygosity with the average of single-locus variances in a population**

We use the general term “single-locus approximation” for both estimators to mean the variance of single-locus estimates as in Eqs. (3.11) and (3.13). This is a frequently used approach for the small proportion of studies that do give variances associated with estimates of heterozygosity and gene diversity. The approximation will be very similar to the total variance of sample gene diversity and heterozygosity

Table 3.1: Relationships between several different expressions for the variances of sample gene diversity ( $\tilde{d}_i$ ) and heterozygosity ( $\tilde{H}_i$ ) for the Nason data set which does not have significant composite linkage disequilibrium. The terms given are the estimates of heterozygosity and gene diversity, their associated within and total-population standard deviations and single-locus approximations ( $s^2$ ) for each site  $i$ , over twelve loci genotyped in two species of moths. Estimates for  $Var_T(\tilde{H}_i)$  were obtained with the REML method and for the mixed and random models. The number of individuals sampled per site, averaged over loci is  $\bar{n}_i$ .

Population	$n_i$	$\tilde{d}_i$	$SD_W(\tilde{d}_i)$	$\sqrt{s_{\tilde{d}_i}^2}$	$\tilde{H}_i$	$SD_W(\tilde{H}_i)$	$\sqrt{s_{\tilde{H}_i}^2}$	$SD_T(\tilde{H}_i)$	
								Random	Mixed
<i>S. altissima</i>									
BB	47.8	0.260	0.013	0.072	0.236	0.017	0.066	0.071	0.032
CL	47.3	0.373	0.009	0.075	0.343	0.017	0.074	0.071	0.032
CM	43.3	0.424	0.009	0.083	0.352	0.014	0.071	0.073	0.033
ZF	43.3	0.441	0.012	0.075	0.323	0.017	0.069	0.072	0.032
<i>S. gigantea</i>									
BB	47.9	0.333	0.012	0.067	0.337	0.016	0.074	0.073	0.026
CL	47.5	0.323	0.012	0.060	0.325	0.019	0.064	0.073	0.026
CM	93.1	0.374	0.008	0.082	0.330	0.012	0.077	0.072	0.023
ZF	46.2	0.296	0.012	0.070	0.283	0.015	0.074	0.073	0.026

obtained with a random effects model only in the cases where the diversities or heterozygosities can be reasonably modelled as having the same expected values and having no dependencies between loci. It was suggested in Weir (1989) that the composite linkage disequilibrium coefficients  $\tilde{\Delta}_{ilul'v}$  be used as an indicator of non-independence between gene diversity estimates at different loci, since these coefficients are expected to be non-zero if dependencies exist.

Two empirical data sets were analyzed in our study, one that showed evidence for composite linkage disequilibrium and one without it, in order to test the validity of this suggestion. More specifically, for the Nason data presented in Table 3.1, 452  $\chi^2$  tests were performed between the pairs of loci for the null hypothesis  $H_o : \Delta_{ilul'v} = 0$ . Of these, by chance 23 of these tests are expected to be significant at the 5% level, and there were 25 significant tests in this case. The covariance between gene diversities at different loci sampled from the same population should not be a major contributor to the total variance of gene diversity for this data set. As noted in Weir et al. (1990), the use of the single-locus approximation assumes that loci are random effects. For the Nason data set, it is likely that the random effects model gives very similar results to the single-locus approximation because there is no significant linkage disequilibrium present (Table 3.1).

A much greater disparity between the variances of sample gene diversity and heterozygosity as estimated by the single-locus approximations and those obtained with variance component methods or exact expressions can be seen with the Olsen data set, supporting the idea of testing for composite linkage disequilibrium as an indicator of covariances between sample gene diversities or heterozygosities. For this data 18 out of 154 tests for composite linkage disequilibrium were significant

at the 5% level, when 8 would be expected to be significant by chance (Table 3.2). The presence of disequilibrium is consistent with the microsatellite markers being located within a 962-base-pair sequence (Olsen, 2002), as opposed to the markers more evenly distributed throughout the genome used by Nason.

### **Approximation of the total variance of sample diversity with the total variance of sample heterozygosity**

Gene diversity and heterozygosity are two measures that should have very similar sources of variation in their observed values since they both measure levels of genetic variation in populations. Due to the complexity of the exact expression for the total variance of gene diversity and this perceived similarity between the variance of these measures, the utility of approximating  $\text{Var}_T(\tilde{d}_i)$  with  $\text{Var}_T(\tilde{H}_i)$  was studied through simulation. If this approximation were found to be reasonable, the total variance estimates could be obtained with the use of variance components methods and would then have well-studied sampling properties due to the extensive statistical theory developed with these statistical methods.

Unfortunately, the relationship between the variances of sample heterozygosity and gene diversity appears to be complex and one is not well approximated by the other (Table 3.3). As heterozygosity decreases relative to a given level of gene diversity, the associated variance of sample heterozygosity increases, while the variance of sample gene diversity is instead decreasing (Table 3.3). This relationship was determined by comparing the within-population variances for a set of simulated data, since direct comparison of the total variances of sample gene diversity and heterozygosity was not possible due to the computational complexity

Table 3.2: Relationships between different expressions for the variances of  $\tilde{d}_i$  and  $\tilde{H}_i$  for the Olsen data set. \* indicates that one or more pairs of loci genotyped in this population had significant composite linkage disequilibrium tests at the  $\alpha = 0.05$  level.

Population	$n_i$	$\tilde{d}_i$	$SD_W(\tilde{d}_i)$	$\sqrt{s_{\tilde{d}_i}^2}$	$\tilde{H}_i$	$SD_W(\tilde{H}_i)$	$\sqrt{s_{\tilde{H}_i}^2}$	$SD_T(\tilde{H}_i)$	
								Random	Mixed
Axixá	6	0.430	0.037	0.145	0.433	0.087	0.128	0.141	0.136
Luzinópolis	6	0.342	0.038	0.133	0.300	0.078	0.141	0.141	0.136
Miranorte*	8	0.755	0.017	0.085	0.550	0.059	0.040	0.134	0.128
Dueré*	6	0.297	0.062	0.053	0.167	0.119	0.119	0.141	0.136
Campos Belos	6	0.270	0.039	0.113	0.267	0.077	0.111	0.141	0.136
Campinorte	6	0.121	0.045	0.067	0.067	0.038	0.074	0.141	0.136
Rialma	6	0.191	0.017	0.145	0.233	0.073	0.119	0.141	0.136
Corumbá	6	0.221	0.018	0.163	0.233	0.030	0.137	0.141	0.136
Nerópolis*	6	0.285	0.050	0.085	0.233	0.099	0.095	0.141	0.136
Goiás Velho*	6	0.361	0.057	0.082	0.367	0.110	0.087	0.141	0.136
Iporá	6	0.315	0.030	0.113	0.233	0.087	0.162	0.141	0.136
Caiapônia*	6	0.506	0.020	0.145	0.433	0.087	0.133	0.141	0.136
Xavatina*	6	0.245	0.056	0.091	0.167	0.056	0.099	0.141	0.136
Petrovina	5	0.000	0.000	0.000	0.000	0.000	0.000	0.147	0.142
Elvira	2	0.300	0.000	0.245	0.400	0.000	0.186	0.191	0.187
Vincente	4	0.450	0.039	0.146	0.300	0.087	0.118	0.155	0.151
d'Oeste	6	0.548	0.036	0.158	0.500	0.041	0.117	0.141	0.136
Lacerda-A*	6	0.506	0.045	0.111	0.300	0.062	0.076	0.141	0.136
Lacerda-B	6	0.506	0.036	0.082	0.467	0.038	0.111	0.141	0.136
Vilhena*	6	0.372	0.023	0.145	0.433	0.087	0.103	0.141	0.136
Bueno	6	0.654	0.032	0.062	0.700	0.078	0.061	0.141	0.136
Jarú	6	0.230	0.026	0.133	0.133	0.038	0.144	0.141	0.136
Ariquemes	6	0.415	0.044	0.113	0.267	0.077	0.134	0.141	0.136
Teotônio	6	0.570	0.044	0.111	0.533	0.038	0.102	0.141	0.136
Taquaras	6	0.464	0.042	0.113	0.400	0.067	0.128	0.141	0.136
Rio Branco	6	0.658	0.036	0.085	0.567	0.073	0.049	0.141	0.136
Madureira	6	0.412	0.048	0.200	0.533	0.077	0.141	0.141	0.136

detailed previously.

### **Models and variance component methods for the total variance of sample heterozygosity**

The choice between the mixed and random models discussed here should depend on the scope of the inferences to be made in a particular study and whether it is appropriate to treat the loci as a random effect given the reality of how the loci were selected for the study. This choice can be important in that there can be a large difference between the random and mixed models if the variance component due to loci is large. This appears to be the case for the Nason data set (Table 3.4). In contrast, the Olsen data set does not appear to have much of a difference in variances between the two models (Table 3.5), possibly resulting from higher covariance between gene diversities at different loci observed in this data set.

Within a model, the different variance component methods give very similar values of  $\text{Var}_T(\tilde{H}_i)$  for relatively balanced data (Tables 3.4 and 3.5). The maximum likelihood based methods appear to be robust to violations of the normality assumption in the data for both data sets analyzed, despite the minimal sample sizes of the Olsen data set. A more unbalanced data set with greater differences between sample sizes was obtained by pooling the data from geographically contiguous populations in the Olsen data set in order to evaluate the effect of unbalanced data on the estimates of the variances. This pooling followed the previous results that the populations generally had genetic variation distributed in an isolation-by-distance pattern (Olsen, 2002). Much greater differences between the variances obtained from the four variance component methods arose from the analysis of this more

Table 3.3: Within-population standard deviations of sample gene diversity and heterozygosity averaged over 1000 replicates of simulated data compared to the true values of the standard deviations for the simulations. Associated simulation parameter values are given by the inbreeding within the population ( $f$ ), the single locus gene diversities ( $d_{il}$ ) and heterozygosities ( $H_{il}$ ).

$d_{il}$	$f$	$H_{il}$	True	Mean	True	Mean
			$SD_W(d_{il})$	$SD_W(\tilde{d}_{il})$	$SD_W(H_{il})$	$SD_W(\tilde{H}_{il})$
0.15	0.0	0.15	0.0324	0.0317	0.0357	0.0353
	0.1	0.13	0.0340	0.0333	0.0342	0.0338
	0.3	0.10	0.0369	0.0362	0.0306	0.0301
	0.5	0.07	0.0397	0.0386	0.0263	0.0256
0.30	0.0	0.30	0.0346	0.0342	0.0458	0.0456
	0.1	0.27	0.0363	0.0358	0.0444	0.0441
	0.3	0.21	0.0395	0.0390	0.0407	0.0403
	0.5	0.15	0.0424	0.0418	0.0357	0.0353
0.45	0.0	0.45	0.0212	0.0212	0.0497	0.0495
	0.1	0.41	0.0222	0.0218	0.0491	0.0489
	0.3	0.32	0.0242	0.0239	0.0465	0.0462
	0.5	0.23	0.0260	0.0258	0.0418	0.0415

unbalanced data set. These indicate that the variance component methods do not produce equivalent results for unbalanced data, in accordance with theoretical expectations (Table 3.6).

Table 3.4: Example of the relationships between the total standard deviation of sample heterozygosity ( $SD_T(\tilde{H}_i)$ ), obtained with two linear models and four estimation methods for a relatively balanced data set in two species of moths.

Population	$\bar{n}_i$	Random Effects Model				Mixed Effects Model			
		ANOVA	MIVQ	REML	ML	ANOVA	MIVQ	REML	ML
<i>S. altissima</i>									
BB	47.8	0.071	0.072	0.071	0.069	0.032	0.032	0.032	0.028
CL	47.3	0.071	0.072	0.071	0.069	0.032	0.032	0.032	0.028
CM	43.3	0.073	0.073	0.073	0.070	0.033	0.033	0.033	0.029
ZF	43.3	0.072	0.072	0.072	0.069	0.032	0.033	0.032	0.028
<i>S. gigantea</i>									
BB	47.9	0.073	0.074	0.073	0.070	0.025	0.025	0.026	0.022
CL	47.5	0.073	0.074	0.073	0.070	0.026	0.025	0.026	0.022
CM	93.1	0.072	0.073	0.072	0.069	0.022	0.022	0.023	0.019
ZF	46.2	0.074	0.074	0.073	0.071	0.026	0.025	0.026	0.023

Table 3.5: Example of the relationships between the total standard deviation of sample heterozygosity ( $SD_T(\tilde{H}_i)$ ) obtained with two linear models and four estimation methods for a relatively balanced data set with minimal sample sizes in *flabellifolia*.

Population	$n_i$	Random Effects Model				Mixed Effects Model			
		ANOVA	MIVQ	REML	ML	ANOVA	MIVQ	REML	ML
Axixá	6	0.141	0.141	0.141	0.140	0.137	0.136	0.136	0.134
Luzinópolis	6	0.141	0.141	0.141	0.140	0.137	0.136	0.136	0.134
Miranorte	8	0.134	0.133	0.134	0.132	0.129	0.129	0.128	0.126
Dueré	6	0.141	0.141	0.141	0.140	0.137	0.136	0.136	0.134
Campos Belos	6	0.141	0.141	0.141	0.140	0.137	0.136	0.136	0.134
Campinorte	6	0.141	0.141	0.141	0.140	0.137	0.136	0.136	0.134
Rialma	6	0.141	0.141	0.141	0.140	0.137	0.136	0.136	0.134
Corumbá	6	0.141	0.141	0.141	0.140	0.137	0.136	0.136	0.134
Nerópolis	6	0.141	0.141	0.141	0.140	0.137	0.136	0.136	0.134
Goiás Velho	6	0.141	0.141	0.141	0.140	0.137	0.136	0.136	0.134
Iporá	6	0.141	0.141	0.141	0.140	0.137	0.136	0.136	0.134
Caiapônia	6	0.141	0.141	0.141	0.140	0.137	0.136	0.136	0.134
Xavatina	6	0.141	0.141	0.141	0.140	0.137	0.136	0.136	0.134
Petrovina	5	0.147	0.146	0.147	0.146	0.142	0.142	0.142	0.140
Elvira	2	0.191	0.191	0.191	0.189	0.188	0.188	0.187	0.185
Vincente	4	0.155	0.155	0.155	0.154	0.151	0.151	0.151	0.148
d'Oeste	6	0.141	0.141	0.141	0.140	0.137	0.136	0.136	0.134
Lacerda-A	6	0.141	0.141	0.141	0.140	0.137	0.136	0.136	0.134
Lacerda-B	6	0.141	0.141	0.141	0.140	0.137	0.136	0.136	0.134
Vilhena	6	0.141	0.141	0.141	0.140	0.137	0.136	0.136	0.134
Bueno	6	0.141	0.141	0.141	0.140	0.137	0.136	0.136	0.134
Jarú	6	0.141	0.141	0.141	0.140	0.137	0.136	0.136	0.134
Ariquemes	6	0.141	0.141	0.141	0.140	0.137	0.136	0.136	0.134
Teotônio	6	0.141	0.141	0.141	0.140	0.137	0.136	0.136	0.134
Taquaras	6	0.141	0.141	0.141	0.140	0.137	0.136	0.136	0.134
Rio Branco	6	0.141	0.141	0.141	0.140	0.137	0.136	0.136	0.134
Madureira	6	0.141	0.141	0.141	0.140	0.137	0.136	0.136	0.134

Table 3.6: Example of the relationships between the total standard deviation of sample heterozygosity ( $SD_T(\tilde{H}_i)$ ) for two linear models and four estimation methods made by pooling geographically contiguous populations in the *flabellifolia* data in order to get a more unbalanced data set.

Pop.	$n_i$	Random Effects Model				Mixed Effects Model			
		ANOVA	MIVQ	REML	ML	ANOVA	MIVQ	REML	ML
Toc.	26	0.075	0.075	0.077	0.073	0.046	0.063	0.067	0.061
Goiás	48	0.067	0.067	0.070	0.066	0.039	0.054	0.059	0.051
Mato	35	0.071	0.071	0.073	0.069	0.042	0.058	0.063	0.056
Rond.	36	0.070	0.070	0.073	0.069	0.042	0.058	0.062	0.055
Acre	12	0.091	0.091	0.093	0.089	0.060	0.081	0.085	0.079

## Discussion

A complete analysis of data in surveys of natural population variation includes calculating appropriate sample variances. We have advocated this practice by illustration with analyses of empirical data sets, and by attempting to clarify the underlying statistical theory motivating the methods and models for obtaining variance estimates. This illustration has shown that the range of heterozygosity and gene diversity estimates can be large and demonstrated that this can be well summarized by including associated variances with point estimators. The results presented here should encourage investigators estimating gene diversity or heterozygosity to consider the sampling properties of their estimators in their analyses in order to increase the quality of their inferences.

For the two data sets analyzed here, testing for the presence of composite linkage disequilibrium was shown to be adequate to predict whether the total variance of sample heterozygosity would differ in a large way from averaging the single locus variances. Generally however, if data from multiple populations is available, avoiding approximation of the total variance of sample heterozygosity or gene diversity by the single-locus approximation is desirable. Instead, it is best to use the random or mixed model variances, which will be more reliable than the single-locus approximation because they more generally account for all sources of variation due to evolutionary history, and also alleviate the need for testing composite linkage disequilibrium coefficients.

Because of the complexity of the exact expression for the total variance of sample gene diversity, approximating this with the total variance of sample het-

erozygosity has been considered here. Heterozygosity is a measure closely related to gene diversity, and the variance of sample heterozygosity can be obtained relatively easily for the linear models described in this paper with the variance component methods available in many statistical packages. Although a direct comparison of the total variances could not be performed due to the complexity of the calculation, a comparison of the within-population variances of these measures using computer simulation indicates that this is not a reasonable approximation to pursue.

From analysis of the empirical data sets presented here, it is clear that the random effects model can produce very different results than the mixed effects model, if the variance due to loci is large. The variance component for loci may be more likely to be large for loci that are not in linkage disequilibrium. The choice of models for an analysis should be considered carefully because of these large differences in results, and should then be made based on the scope of inferences that the investigator wishes to make, and how reasonable it is to assume that the observed loci were randomly sampled.

It has been shown that the variance component methods studied here can produce different estimates for unbalanced data sets. As expected, the ANOVA method performs differently than REML, ML and MIVQUE for unbalanced data. It is reassuring however that from the results of the analyses given here, the maximum likelihood based methods appear to be robust to the effect of small sample sizes and the assumption of normality made by these approaches. The REML method is perhaps to be preferred because it accounts for fixed effects with respect to degrees of freedom and has guaranteed minimum variance properties, which the ML approach does not, and because it does not allow negative variance estimates,

as can be obtained in some cases with MIVQUE and ANOVA. While for the cases where the populations were of relatively similar sizes, similar results for all variance component methods were obtained, because the use of any of these methods should be just a matter of directing the statistical software package to perform the appropriate analysis, it may be best to rely on MIVQUE or REML methods for all analyses, regardless of the sampling design of the study.

Choosing a variance component method such as MIVQUE or REML that has better properties for unbalanced data, and working with the random model described here are both conservative approaches and should work to increase the confidence that an investigator can place in making inferences comparing the gene diversity or heterozygosity of different populations of interest. Certainly, using a random or mixed effects model to obtain estimates of the total variance of sample heterozygosity, rather than estimating a within-population variance, is likely to be more desirable for most biological studies, since the total variance will then include the effects of both the evolutionary and statistical sampling that has occurred within the data set. This means that inferences could then properly be made about the genetic variation of larger groups such as species, rather than limiting the scope of inferences to the specific populations sampled in a study.

## Literature Cited

- Cockerham, C. C. 1971. Higher order probability functions of identity of alleles by descent. *Genetics* 69:235–246.
- Dodds, K. G. 1986. Resampling Methods in Genetics and the Effect of Family Structure in Genetic Data. PhD thesis, North Carolina State University, Raleigh, NC.
- Grenier, C., Deu, M., Kresovich, S., Bramel-Cox, P. J., and Hamon, P. 2000. Assessment of genetic diversity in three subsets constituted from the ICRISAT sorghum collection using random vs non-random sampling procedures B. Using molecular markers. *Theor. Appl. Genet.* 101:197–202.
- Marshall, D. R. and Allard, R. W. 1970. Isozyme polymorphisms in natural populations of *Avena fatua* and *A. barbata*. *Heredity* 25:373–382.
- Nason, J. D., Heard, S. B., and Williams, F. R. 2002. Host-associated genetic differentiation in the goldenrod elliptical-gall moth, *Gnorimoschema Gallaesolidaginis* (Lepidoptera: Gelechiidae). *Evolution* 56:1475–1488.
- Nei, M. 1973. Analysis of gene diversity in subdivided populations. *Proc. Natl. Acad. Sci. U.S.A.* 70:3321–3323.
- Nei, M. and Roychoudhury, A. K. 1974. Sampling variances of heterozygosity and genetic distance. *Genetics* 76:379–390.

- Olsen, K. M. 2002. Population history of *Manihot esculenta* (Euphorbiaceae) inferred from nuclear DNA sequences. *Mol. Ecol.* 11:901–911.
- Olsen, K. M. and Schaal, B. A. 2002. Microsatellite variation in cassava (*Manihot esculenta*, Euphorbiaceae) and its wild relatives: further evidence for a southern Amazonian origin of domestication. *Amer. Jo. Bot.* 88:131–142.
- Rao, C. R. 1971. Estimation of variance covariance components - MINQUE theory. *J. Multi. Anal.* 1:257–275.
- Rao, C. R. and Kleffe, J. 1988. Estimation of Variance Components and Applications. North-Holland, Amsterdam.
- Searle, S. R., Casella, G., and McCulloch, C. E. 1992. Variance Components. Wiley, New York.
- Shete, S. 2003. Uniformly minimum variance unbiased estimation of gene diversity. *J. Hered.* 94:421–424.
- Weir, B. S. 1989. Sampling properties of gene diversity, Ch. 2 *in* Brown, A. H. D., Clegg, M. T., Kahler, A. L. and B. S. Weir, eds. *Plant Population Genetics, Breeding and Genetic Resources*. Sinauer, Sunderland, MA.
- Weir, B. S. 1996. *Genetic Data Analysis*. Sinauer, Sunderland, MA.
- Weir, B. S. 2001. Forensics, Pp. 721-739 *in* Balding, D.J., Bishop, M. and C. Cannings, eds. *Handbook of Statistical Genetics*. Wiley, New York.

Weir, B. S., Reynolds, J., and Dodds, K. G. 1990. The variance of sample heterozygosity. *Theor. Popul. Biol.* 37:235–253.

## Appendix

### Using descent measures to approximate the total variance of gene diversity

An expression for the total variance of sample gene diversity for a population at a single locus in terms of descent measures was given by (Weir, 1989) as

$$\begin{aligned}
 \text{Var}_T(\tilde{d}_{il}) = & \left[ (\Delta - \theta^2) + 2(\theta^2 - \delta) \sum_u p_u^2 \right. \\
 & + 4(\theta - 2\gamma - \Delta + 2\delta) \sum_u p_u^3 \\
 & \left. - (4\theta - 8\gamma - 3\Delta + 6\delta + \theta^2) \left( \sum_u p_u^2 \right)^2 \right] \\
 & + \frac{1}{n_i} \left[ (2\gamma - 3\Delta + \theta^2) + 2(\theta - 3\gamma + 3\delta - \theta^2) \sum_u p_u^2 \right. \\
 & + 2(1 - 9\theta + 14\gamma + 6\Delta - 12\delta) \sum_u p_u^3 \\
 & \left. - (2 - 16\theta + 24\gamma + 9\Delta - 18\delta - \theta^2) \left( \sum_u p_u^2 \right)^2 \right]. \quad (3.15)
 \end{aligned}$$

For the sake of simplicity this expression assumes random-mating populations so that allelic relationships do not depend on the arrangement of alleles within genotypes. For this case, four measures are adequate to describe relationships between alleles:  $\theta$ ,  $\gamma$ ,  $\delta$  and  $\Delta$  for pairs, triples quadruples and two-pairs of alleles respectively (Cockerham, 1971). Despite this assumption reducing the potential number of descent measures needed, the expression for the total variance of sample gene diversity remains very complex.

Weir (2001) found that these four measures could further be simplified into terms of  $\theta$ , the relationship between a pair of alleles, by assuming evolutionary

stationarity, a common assumption of work with the coalescent theory. The terms can then be rewritten as

$$\begin{aligned}\gamma &= \frac{2\theta^2}{1+\theta}, \\ \delta &= \frac{6\theta^3}{(1+\theta)(1+2\theta)}, \\ \Delta &= \frac{\theta^2(1+5\theta)}{(1+\theta)(1+2\theta)}.\end{aligned}$$

Substituting these expressions into Eq. (3.15) gives

$$\begin{aligned}\text{Var}_T(\tilde{d}_{il}) &= \frac{2(1-\theta)}{n(1+\theta)(1+2\theta)} \left\{ \theta[\theta(n-1)+1] \left[ \theta + (1-2\theta) \sum_u p_u^2 \right] \right. \\ &\quad + (1-\theta) \left[ [2\theta(n-2)+1] \sum_u p_u^3 \right. \\ &\quad \left. \left. + [\theta^2(n-1) + \theta(2n-3) + 1] \left( \sum_u p_u^2 \right)^2 \right] \right\}.\end{aligned}$$

An example of the use of this approximation is given in Table 3.7. The approximation of  $\text{Var}_T(\tilde{d}_{il})$  appears to be fairly conservative, relative to the results of the single-locus approximation.

Table 3.7: Comparison of the total standard deviation of sample gene diversity ( $SD_T(\tilde{d}_{il})$ ) approximated as a function of  $\theta$  and allele frequencies to the standard deviation of the sample gene diversities at individual loci  $\left(\sqrt{s_{\tilde{d}_i}^2}\right)$  for the 5 microsatellite loci of the Olsen *flabellifolia* data set. For this data set  $\hat{\theta} = 0.42$ .

Population	$SD_T(\tilde{d}_{il})$					$\sqrt{s_{\tilde{d}_i}^2}$
	GAGG5	GA134	GA16	GA12	GA140	
Axixá	0.697	0.531	0.404	0.395	0.304	0.145
Luzinópolis	0.697	0.365	0.414	0.697	0.402	0.133
Miranorte	0.316	0.247	0.347	0.277	0.263	0.085
Dueré	0.610	0.610	0.697	0.359	0.414	0.053
Campos Belos	0.697	0.432	0.697	0.460	0.476	0.113
Campinorte	0.697	0.536	0.697	0.536	0.697	0.067
Rialma	0.697	0.476	0.697	0.395	0.697	0.145
Corumbá	0.697	0.697	0.697	0.359	0.432	0.163
Nerópolis	0.610	0.536	0.476	0.697	0.408	0.085
Goiás Velho	0.476	0.610	0.610	0.365	0.465	0.082
Iporá	0.697	0.303	0.697	0.610	0.338	0.113
Caiapônia	0.432	0.359	0.697	0.314	0.320	0.145
Xavatina	0.697	0.610	0.378	0.610	0.536	0.091
Petrovina	0.692	0.692	0.692	0.692	0.692	0.000
Elvira	0.641	0.641	0.641	0.387	0.336	0.245
Vincente	0.684	0.393	0.413	0.458	0.347	0.146
d'Oeste	0.476	0.338	0.610	0.338	0.263	0.158
Lacerda-A	0.324	0.338	0.536	0.476	0.460	0.111
Lacerda-B	0.610	0.465	0.349	0.465	0.269	0.082
Vilhena	0.395	0.404	0.432	0.536	0.697	0.145
Bueno	0.365	0.331	0.465	0.281	0.309	0.062
Jarú	0.697	0.697	0.432	0.697	0.338	0.133
Ariquemes	0.536	0.697	0.476	0.402	0.279	0.113
Teotônio	0.365	0.610	0.349	0.329	0.320	0.111
Taquaras	0.460	0.298	0.697	0.359	0.432	0.113
Rio Branco	0.408	0.263	0.354	0.378	0.338	0.085
Madureira	0.536	0.536	0.697	0.338	0.281	0.200