2013

On inferring and interpreting genetic population structure - applications to conservation, and the estimation of pairwise genetic relatedness

Arun Sethuraman
Iowa State University

Follow this and additional works at: https://lib.dr.iastate.edu/etd

Part of the Bioinformatics Commons, and the Genetics Commons

Recommended Citation
https://lib.dr.iastate.edu/etd/13332

This Dissertation is brought to you for free and open access by the Iowa State University Capstones, Theses and Dissertations at Iowa State University Digital Repository. It has been accepted for inclusion in Graduate Theses and Dissertations by an authorized administrator of Iowa State University Digital Repository. For more information, please contact digirep@iastate.edu.
On inferring and interpreting genetic population structure - applications to conservation, and the estimation of pairwise genetic relatedness

by

Arun Sethuraman

A dissertation submitted to the graduate faculty
in partial fulfillment of the requirements for the degree of
DOCTOR OF PHILOSOPHY

Major: Bioinformatics and Computational Biology

Program of Study Committee:
Fredric J Janzen, Co-major Professor
Karin S Dorman, Co-major Professor
Anne M Bronikowski
John D Nason
David Fernandez-Baca

Iowa State University
Ames, Iowa
2013

Copyright © Arun Sethuraman, 2013. All rights reserved.
DEDICATION

I would like to dedicate this dissertation to my loving parents, Lalitha Sethuraman and V. Sethuraman, without whose support, none of this work would have been possible.
TABLE OF CONTENTS

LIST OF TABLES ................................................................. vi

LIST OF FIGURES ............................................................... x

ABSTRACT ............................................................... xviii

CHAPTER 1. ON POPULATION DIFFERENTIATION, AND METHODS

TO INFER GENETIC SUBDIVISION ........................................ 1

1.1 Introduction ......................................................... 1

1.2 Methods .............................................................. 4

   1.2.1 F-statistics and AMOVA ..................................... 5

   1.2.2 Model-based Clustering Methods ............................ 8

   1.2.3 Other Clustering Methods .................................... 11

1.3 Discussion ............................................................ 12

1.4 Tables and Figures .................................................. 17

CHAPTER 2. POPULATION GENETICS OF BLANDING’S TURTLE IN

THE MIDWESTERN UNITED STATES ....................................... 25

2.1 Abstract ............................................................. 25

2.2 Introduction ......................................................... 25

2.3 Materials and Methods .............................................. 28

   2.3.1 Fieldwork ......................................................... 28

   2.3.2 Genetic data generation ....................................... 28

   2.3.3 Genetic data analyses .......................................... 29

2.4 Results ............................................................... 33

   2.4.1 Microsatellite characteristics and deviations from HWE .... 33
2.4.2 Population structure .............................................. 34
2.4.3 Historical population parameters .............................. 35
2.5 Discussion ............................................................... 36
2.6 Acknowledgements .................................................... 41
2.7 Tables and Figures ..................................................... 42

CHAPTER 3. MULTICLUST - FAST MULTINOMIAL CLUSTERING OF
MULTILOCUS GENOTYPES TO INFERENCE GENETIC POPULATION
STRUCTURE ............................................................... 48
3.1 Abstract ................................................................. 48
3.2 Introduction ............................................................ 48
3.3 Methods ................................................................. 51
  3.3.1 Model ............................................................... 51
  3.3.2 Model Selection .................................................. 54
  3.3.3 Simulations and Datasets ....................................... 55
  3.3.4 STRUCTURE Runs ............................................... 57
3.4 Results ................................................................. 57
  3.4.1 Simulation .......................................................... 57
  3.4.2 Empirical Datasets from Gilbert et al. 2012 ................. 58
  3.4.3 Blanding’s Turtle ............................................... 58
3.5 Discussion ............................................................. 59
3.6 Tables and Figures .................................................... 62

CHAPTER 4. ESTIMATING RELATEDNESS USING ADMIXTURE PRO-
PORTIONS IN STRUCTURED POPULATIONS ...................... 69
4.1 Abstract ................................................................. 69
4.2 Introduction ........................................................... 69
4.3 Theory ................................................................. 74
  4.3.1 Relatedness under the Admixture Model ...................... 74
  4.3.2 Other Relatedness Estimators ................................. 76
4.3.3 Simulations .................................................. 77
4.4 Results .......................................................... 80
  4.4.1 Effect of Varying $K - 1$ ................................ 80
  4.4.2 Effect of Varying $K - 2$ ............................... 81
  4.4.3 Effect of Number of Loci .............................. 81
4.5 Discussion ..................................................... 82
4.6 Figures and Tables .......................................... 84

CHAPTER 5. CONCLUSION .............................................. 107

APPENDIX A. APPENDIX TO POPULATION GENETICS OF BLANDING’S TURTLE IN THE MIDWESTERN UNITED STATES .......... 110

APPENDIX B. APPENDIX TO MULTICLUST - FAST MULTINOMIAL CLUSTERING OF MULTILOCUS GENOTYPES TO INFER GENETIC SUBPOPULATION STRUCTURE ................................. 129

APPENDIX C. APPENDIX TO ESTIMATING RELATEDNESS USING ADMIXTURE PROPORTIONS IN STRUCTURED POPULATIONS .... 136

BIBLIOGRAPHY ...................................................... 139

ACKNOWLEDGEMENTS ............................................. 157
LIST OF TABLES

1.1 Mean $F$ statistics estimated from the human microsatellite data. Of note is the variability in estimates of ‘similar’, and analogous measures of differentiation. .............................................................. 17

1.2 AMOVA results using 7 geographical groups as subpopulations. ................ 17

2.1 Summary statistics across the 12 populations with sample size $> 5$. ‘N’ is the total number of individuals sampled for the population, ‘n’ is the mean sample genotyped over all the loci, ‘P’ is the total number of polymorphic loci in that population, ‘A’ is the average number of alleles per locus, ‘Ap’ is the mean number of alleles per polymorphic locus, ‘He’ is the expected heterozygosity, and ‘Ho’ is the observed heterozygosity. The statistics were estimated using GDA v.1.1. ‘P-val’ indicates P-values from a global test of heterozygote deficiency, performed in Genepop v. 4.1, under the null hypothesis that the populations are at Hardy-Weinberg Equilibrium, with the alternate hypothesis that the populations have significant heterozygote deficiency ......................................................... 42

2.2 Lower triangle contains pair-wise Fst values between populations; the significant ones before sequential Bonferroni correction are shown in italics, non-significant Fst values regardless of Bonferroni correction are shown in grey, and the significant Fst values post correction are shown in boldface. Fifty-five of 66 pairwise comparisons remained significant after sequential Bonferroni correction. The upper triangle contains linear geographic distance in kilometers between population pairs ........................................ 43
3.1 Simulation 1 settings ................................................. 62
3.2 Simulation results .................................................... 62
3.3 Inference of Ancestral Subpopulations, $K$ from empirical data reported in Gilbert et al. (2012). $P$ indicates models picked using the method of Pritchard et al. (2000), $P+E$ shows models picked using the method of Evanno et al. (2005). .................................................. 67

4.1 Conditional Probabilities $P(S_p \mid D_q)$ ................................ 102
4.2 List of estimators tested and their references .......................... 105
4.3 Results of tests of Hardy-Weinberg Equilibrium ($H_a$=heterozygote deficiency) across all 300 unlinked loci simulated under the continent-island model for $K=3,5$ and 10. .............................................. 106

A.1 Groups of populations of *Emys blandingii* within a radius of 100km of linear geographical distance (Model 1) from each other, as estimated by Geographical Distance Estimator. Populations that are underlined were excluded from analyses of population $F$-statistics, Mantel Tests of Isolation by Distance, population differentiation, and heterozygositities to prevent bias in allele frequency calculations, but they were included in analyses of population structure using STRUCTURE and in estimates of ancestral migration using IM ................................................ 115

A.2 Groups of populations of *Emys blandingii* based on watershed distribution (Model 2) ........................................ 115

A.6 Estimates of genotype and allele counts, as well as expected and observed heterozygosities, across loci in all populations of *Emys blandingii* sampled. Significant deviations from HWE after sequential Bonferroni correction are shown in gray, although the $P$-values given in the table are uncorrected. These tests were performed using 100,000 iterations in Arlequin v.3.0 .............................................. 119
A.3 Groups of populations of *Emys blandingii* based on watershed and Laurentide Ice Sheet distribution (Model 3) .................................................. 120

A.4 Groups of populations of *Emys blandingii* identified by STRUCTURE upon exclusion of the *GmuD95* locus (Fig 3). Populations in parentheses were excluded from IM and BayesAss analyses, owing to small sample sizes. Group 3 was also split into Grant-NE and (Carroll-IL, Will-IL) to resolve ancestral splits between these populations for IM, but not BayesAss, analyses ................................................................. 120

A.5 Expected (*H*<sub>e</sub>) and observed heterozygosities (*H*<sub>o</sub>) across 201 genotyped individuals sampled from 12 populations of *Emys blandingii* with a sample size of greater than 5, estimated using GDA v.1.1. *N* denotes the total number individuals genotyped at that locus, *H*<sub>e</sub> is the expected heterozygosity, and *H*<sub>o</sub> is the observed heterozygosity ........................................... 121

A.7 Pair-wise tests of linkage disequilibrium, performed with 100000 permutations in Genepop v.4.1, on populations of *Emys blandingii*. The first two columns show the pairs of microsatellite loci tested, the third column shows the *χ*<sup>2</sup> values from Fisher’s Exact Test, the fourth column indicates the number of degrees of freedom, and the fifth column contains the P-values. No significant linkage disequilibrium was detected at the genotyped loci, hence no corrections were performed for multiple comparisons ................................................................. 122

A.8 Estimates of Weir and Cockerham *F*-statistics: *F*<sub>is</sub> (population differentiation between individuals among sampled locations), *F*<sub>it</sub> (population differentiation among individuals), and *F*<sub>st</sub> (population differentiation between populations), using GDA v.1.1. All analyses were performed with and without the *GmuD95* locus for 202 individuals (from 12 well-sampled populations of *Emys blandingii*) using only a priori information on their sampling locations (no ‘grouping’). 95% confidence intervals were established by bootstrapping with 10,000 replicates ........................................... 123
A.9 AMOVA results showing genetic variance for *Emys blandingii* populations under the hypothesized models specified in Tables S1, S2, and S3.

A.10 Maximum likelihood estimates of effective population sizes (θ) for *Emys blandingii*. The values shown are HiPt values or the values with the most number of counts (median). The values in parentheses represent 95% confidence intervals around the mean. Values with a * failed to converge in our IM analyses.

A.11 Maximum likelihood estimates of migration rates per generation (m) and estimated time since splitting in years (t) for *Emys blandingii*. The values shown are HiPt values or the values with the most number of counts (median). The values within parentheses represent 95% confidence intervals around the mean. Values with a * failed to converge in our IM analyses.

A.12 Estimates of recent migration rates, with 95% confidence intervals around the means in parentheses, for groups of *Emys blandingii* populations. These estimates were obtained using 10,000,000 iterations, with a burn in of 100,000 iterations, and sampling every 1000th value from the distribution, as suggested by Rannala (2007). These groups are those obtained from STRUCTURE, as listed in Table S4. Group 5 was excluded from these analyses owing to small sample sizes. Group 3 was not split into Grant-NE and (Carroll-IL, Will-IL) as with IM analyses, since BayesAss estimates recent migrations.

A.13 Pair-wise differentiation estimates between microsatellite flanking region sequences of *Emys blandingii* from Carroll-IL, Will-IL, Grant-NE, and McHenry-IL. These estimates were determined using DnaSP 5.10 (Librado et al. 2009) by using all the sequences from these populations.
LIST OF FIGURES

1.1 Distributions of Nei’s $G_{ST}$ (mean = 0.055), Jost’s $D_{ST}$ (mean = 0.207), Hamrick and Godt’s $G_{ST}$ (mean = 0.129), and Weir and Cockerham’s $\phi_{ST}$ (mean = 0.053) over 377 loci for the human microsatellite data.

1.2 Distributions of Nei’s $G_{ST}$ (mean = 0.055), Jost’s $D_{ST}$ (mean = 0.207), Hamrick and Godt’s $G_{ST}$ (mean = 0.129), and Weir and Cockerham’s $\phi_{ST}$ (mean = 0.053) over number of unique alleles across the 377 loci for the human microsatellite data.

1.3 Plot of $\Delta K$ values from four runs of STRUCTURE on the human dataset, varying $K$ from 1 to 10 are shown in the top panel. The most likely value for $K$ was identified at $K = 4$. Plot of BIC estimates using DAPC to infer the number of clusters in the same dataset are shown on the bottom. The results are inconclusive, with the BIC still falling post $K = 50$. These plots illustrate a standing issue with inference of the true $K$ in a given dataset.

1.4 Plot of admixture proportions at the population level, at $K = 4$, identified using STRUCTURE is shown in the top panel. Plot of membership probabilities, which can be equated to admixture proportions, as identified by DAPC at $K = 4$ are shown in the bottom panel. For a detailed list of legends, see Rosenberg et al. 2002. Note the similarity in patterns of admixture, as indicated by the two column graphs.

1.5 DAPC plot of Principal Axes 1 and 2, showing the most variation in the microsatellite genotype data. The first two PC’s show most variation explained in three clusters.
1.6 DAPC plot of Principal Axes 1 and 2, at a chosen $K = 4$.

1.7 Admixture proportion plot at $K = 4$ for most significantly admixed populations. These were identified as Uygur (1306), and Hazara (129).

2.1 Localities and putative groupings of the 18 sites sampled for *Emys blandingii* in the midwestern United States. (top left) Gray areas are water bodies, the more southern dashed line indicates the extent of the pre-Wisconsinan glacial limit (Kansan/Nebraskan glaciations), and the more northern dashed line indicates the extent of the Laurentide Ice Sheet in the more recent Wisconsinan glaciation. (top right) Representation of Model 1 (Table S1), with sites within 100km linear distance from each other clustered to form groups. (middle left) Representation of Model 2 (Table S2), with sites in the same watershed clustered to form groups. (middle right) Representation of Model 3 (Table S3), with sites in the same watershed and location with respect to the limit of the Laurentide Ice Sheet clustered to form groups.

2.2 Representation of sample sizes (indicated by diameter of the pies) and of relative admixture distribution (indicated by shade of gray) inferred by STRUCTURE at $K=5$.

2.3 Plot of linear geographic distance between sampled populations of *Emys blandingii* vs. Slatkin’s linearized Fst genetic distance between these same sampled populations to estimate the presence of isolation by distance. This plot was derived from a Mantel test performed in GenAlex v.6.2 with 1000 permutations.
2.4 Estimates of admixture proportions in sampled populations of *Emys blandingii*. Twenty runs of STRUCTURE were performed for each value of K, under the admixture model, and the Dirichlet parameter ‘alpha’ was inferred from the data. Each run was performed using the 209 genotyped individuals from all 18 localities with a burn-in period of 10000 and 10000 MCMC reps. Left) All loci. Right) all loci except *GmuD95*.

3.1 Plots of AIC versus $K$ using datasets from Gilbert et al. (2012) . . . .

3.2 Plots of BIC versus $K$ using datasets from Gilbert et al. 2012. . . .

3.3 Plots of mean Ln Probabilities estimated by STRUCTURE (Pritchard et al. 2000) on all datasets from Gilbert et al. 2012 . . . .

3.4 Plots of $\Delta K$ estimated by the method of Evanno et al. 2005 on three datasets from Gilbert et al. 2012 . . . .

3.5 Results of mixture and admixture models for $K = 1, \ldots, 9$. . . .

3.6 Plot of mixture proportions of Blanding’s turtles at $K = 5$. Size of the pies indicates the sample sizes, colors indicate mixture proportions .

4.1 Jacquard’s (1977) IBD states, $D_1, \ldots, D_9$ for two diploid individuals. The top row shows the diploid genotype of individual 1, and the bottom row indicates the diploid genotype of individual 2. The alleles are connected by a line if they are Identical By Descent (IBD).

4.2 Bias in estimates of genetic relatedness between 1000 Full Sib (FS) dyads, with increasing degree of subpopulation structure. The top panel shows the bias estimates when parents of the FS dyads were chosen from the same subpopulation. I also performed another set of simulations, where the parents were picked from different subpopulations, bias estimates of which are shown in the bottom panel.
4.3 Bias in estimates of genetic relatedness between 1000 Half Sib (HS) dyads, with increasing degree of subpopulation structure. The top panel shows the bias estimates when parents of the HS dyads were chosen from the same subpopulation. I also performed another set of simulations, where the parents were picked from different subpopulations, bias estimates of which are shown in the bottom panel.

4.4 Bias in estimates of genetic relatedness between 1000 Parent-Offspring (PO) dyads, with increasing degree of subpopulation structure. The top panel shows the bias estimates when both parents of the PO dyads were chosen from the same subpopulation. I also performed another set of simulations, where the parents were picked from different subpopulations, bias estimates of which are shown in the bottom panel.

4.5 Bias in estimates of genetic relatedness between 1000 First Cousin (FC) dyads, with increasing degree of subpopulation structure. The top panel shows the bias estimates when parents of the FC dyads were chosen from the same subpopulation. I also performed another set of simulations, where the parents were picked from different subpopulations, bias estimates of which are shown in the bottom panel.

4.6 Bias in estimates of genetic relatedness between 1000 Unrelated Individual (UR) dyads, with increasing degree of subpopulation structure. The top panel shows the bias estimates when both unrelated individuals were chosen from the same subpopulation. I also performed another set of simulations, where both individuals were picked from different subpopulations, bias estimates of which are shown in the bottom panel.

4.7 Mean Squared Error in estimates of genetic relatedness between 1000 Full Sib (FS) dyads, with increasing degree of subpopulation structure. Top panel shows MSE when parents of FS were picked from the same subpopulation, bottom panel when they were picked from different subpopulations.
4.8 Mean Squared Error in estimates of genetic relatedness between 1000 Half Sib (HS) dyads, with increasing degree of subpopulation structure. Top panel shows MSE when parents of HS were picked from the same subpopulation, bottom panel when they were picked from different subpopulations. ................................. 92

4.9 Mean Squared Error in estimates of genetic relatedness between 1000 Parent-Offspring (PO) dyads, with increasing degree of subpopulation structure. Top panel shows MSE when both parents in the PO dyads were picked from the same subpopulation, bottom panel when they were picked from different subpopulations. ................................. 93

4.10 Mean Squared Error in estimates of genetic relatedness between 1000 First Cousin (FC) dyads, with increasing degree of subpopulation structure. Top panel shows MSE when parents of FC were picked from the same subpopulation, bottom panel when they were picked from different subpopulations. ................................. 94

4.11 Mean Squared Error in estimates of genetic relatedness between 1000 Unrelated Individual (UR) dyads, with increasing degree of subpopulation structure. Top panel shows MSE when both UR individuals were picked from the same subpopulation, bottom panel when they were picked from different subpopulations. ................................. 95

4.12 Bias and Mean Squared Error in estimates of genetic relatedness between 1000 Full Sib (FS) dyads sampled from $K = 3$ subpopulations, with increasing number of genotyped loci ................................. 96

4.13 Bias and Mean Squared Error in estimates of genetic relatedness between 1000 Full Sib (FS) dyads sampled from $K = 5$ subpopulations, with increasing number of genotyped loci ................................. 97

4.14 Bias and Mean Squared Error in estimates of genetic relatedness between 1000 Full Sib (FS) dyads sampled from $K = 10$ subpopulations, with increasing number of genotyped loci ................................. 98
<table>
<thead>
<tr>
<th>Section</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>4.15</td>
<td>Bias and MSE estimates for 1000 full-sib dyads, simulated under the Continent-Island Model. X axis indicates the number of subpopulations (or islands), K.</td>
</tr>
<tr>
<td>4.16</td>
<td>Bias and MSE estimates for 1000 half-sib dyads, simulated under the Continent-Island Model. X axis indicates the number of subpopulations (or islands), K.</td>
</tr>
<tr>
<td>4.17</td>
<td>Bias and MSE estimates for 1000 first cousin dyads, simulated under the Continent-Island Model. X axis indicates the number of subpopulations (or islands), K.</td>
</tr>
<tr>
<td>4.18</td>
<td>Bias and MSE estimates for 1000 parent-offspring dyads, simulated under the Continent-Island Model. X axis indicates the number of subpopulations (or islands), K.</td>
</tr>
<tr>
<td>4.19</td>
<td>Bias and MSE estimates for 1000 unrelated individual dyads, simulated under the Continent-Island Model. X axis indicates the number of subpopulations (or islands), K.</td>
</tr>
<tr>
<td>A.1</td>
<td>Plots of mean logarithmic probability (estimated from STRUCTURE) versus number of putative populations (K) tested for Emys blandingii. (Top) Results from STRUCTURE using all loci; the most likely value of ‘K’ was identified as 4 (Bottom) Results from STRUCTURE without the $GmuD95$ locus; the most likely value of ‘K’ was identified as 5</td>
</tr>
</tbody>
</table>
A.2 The evolutionary history at the *GmuD121* locus for *Emys blandingii* was inferred by using the Maximum Likelihood method based on the Kimura 2-parameter model [6]. The bootstrap consensus tree inferred from 1000 replicates [2] is taken to represent the evolutionary history of the taxa analyzed [2]. Branches corresponding to partitions reproduced in less than 50% bootstrap replicates are collapsed. The percentages of replicate trees in which the associated taxa clustered together in the bootstrap test (1000 replicates) are shown next to the branches [2]. Initial tree(s) for the heuristic search were obtained automatically as follows. When the number of common sites was \( < 100 \) or \( < 1/4 \) of the total number of sites, maximum parsimony method was used; otherwise the BIONJ method with MCL distance matrix was used. A discrete Gamma distribution was used to model evolutionary rate differences among sites (5 categories (+G, parameter = 200.0000)). The rate variation model allowed for some sites to be evolutionarily invariable ([+I], 31.6049% sites). The tree is drawn to scale, with branch lengths measured in the number of substitutions per site. The analysis involved 7 nucleotide sequences (= individuals). All positions containing gaps and missing data were eliminated, yielding 83 nucleotide positions in the final dataset. Evolutionary analyses were conducted in MEGA5 [10].

A.3 The evolutionary history at the *GmuD95* locus for *Emys blandingii* was inferred by using the Maximum Likelihood method based on the Jukes-Cantor model [6]. A discrete Gamma distribution was used to model evolutionary rate differences among sites (5 categories (+G, parameter = 4.1286)). The rate variation model allowed for some sites to be evolutionarily invariable ([+I], 68.6479% sites). Other methods are identical to those described for Fig. S3. The analysis involved 9 nucleotide sequences (= individuals). The final dataset contained 64 nucleotide positions.
A.4 The evolutionary history at the GmuD21 locus for *Emys blandingii* was inferred by using the Maximum Likelihood method based on the Jukes-Cantor model [6]. A discrete Gamma distribution was used to model evolutionary rate differences among sites (5 categories (+G, parameter = 5.3515)). The rate variation model allowed for some sites to be evolutionarily invariable ([+I], 65.3268% sites). Other methods are identical to those described for Fig. S3. The analysis involved 6 nucleotide sequences (= individuals). The final dataset contained 94 nucleotide positions.

A.5 Phylogenetic consensus tree constructed using a Neighbor Joining method and a bootstrap of 1000 replicates in PHYLIP v.3.69, placed alongside the population genetic structure of *Emys blandingii* inferred by STRUCTURE analysis. Diameter of pies on the map indicates sample sizes at those populations, while slices indicate admixture proportions, as estimated in STRUCTURE.
The presence of population structure is ubiquitous in most wild populations of species. Detecting genetic population structure and understanding its consequences for the evolutionary trajectories of species has shaped a lot of our understanding of the process of evolution. This delineation of subdivision within a population plays an important role in several allied fields, including conservation genetics, association studies, phylogeography, and quantitative genetics.

This dissertation addresses methods to infer and interpret subpopulation structure. In this regards, I discuss the standing motivation for developing new analytic tools, a classic population genetics study of the imperiled freshwater turtle, *Emys blandingii*, the development of a fast, likelihood based estimator of subpopulation structure, MULTICLUST, and a likelihood based method to infer pairwise genetic relatedness in the presence of subpopulation structure.

Our analyses of population structure in midwestern populations of *Emys blandingii* detected considerable genetic structure within and among the sampled localities, and revealed ancestral gene flow of *E. blandingii* in this region north and east from an ancient refugium in the central Great Plains, concordant with post-glacial recolonization timescales. The data further implied unexpected ‘links’ between geographically disparate populations in Nebraska and Illinois. Our study encourages conservation decisions to be mindful of the genetic uniqueness of populations of *E. blandingii* across its primary range.

Analyses of both simulated and empirical data suggests that MULTICLUST infers structure consistently (reproducible results), and is time efficient, compared to the popular Bayesian MCMC tool, STRUCTURE (Pritchard et al. (2000b)). The new likelihood estimator of pairwise genetic relatedness also has the least bias, and mean squared error in estimating relatedness in full-sibling, half-sibling, parent-offspring, and a variety of other related dyads, compared to the methods of Anderson and Weir (2007), Queller and Goodnight (1989), Lynch and Ritland (1999).
Overall, this dissertation lays the grounds for several interesting biological and statistical questions that can be addressed with a robust framework for identification of subpopulation structure.
CHAPTER 1. ON POPULATION DIFFERENTIATION, AND
METHODS TO INFER GENETIC SUBDIVISION

1.1 Introduction

Population genetic structure is the presence of genetic differentiation among subpopulations within a global population of a species, where some individuals are more genetically similar to other individuals, than to others. Population genetic structure is created in a global population by the presence of physical, or behavioral barriers to breeding between subpopulations. Such structure ubiquitously contributes towards the process of evolution, owing to decreasing the effective population size of a population, hence making the population susceptible to random genetic drift. Subpopulation structure also leads to localized fluctuations in allele frequencies, which may lead to subpopulation specific selection effects, and environmental interactions. Our understanding of what population genetic structure means for the generalized process of evolution, and the spread of advantageous and deleterious mutations, has had a huge impact on a multitude of fields in biology (see Pritchard et al. (2000b) for a review). For instance, studies of microsatellites and Single Nucleotide Polymorphisms (SNPs) in humans, sampled across the world have brought forth concrete evidence of what was only previously hypothesized - a phylogeographic history of the human species since its origin in the continent of Africa (Rosenberg et al. (2002), Templeton (2002)). Despite the many illuminating studies in this area, of what is termed genetic admixture, or the degrees of mixture between subpopulations, the field is wrought with conflicting definitions of these pseudo boundaries, called subpopulations. Still, Sewall Wright (Wright, 1951) was among the first to recognize the importance of subdivision - individuals within a subpopulation are inherently more similar genetically to each other than to individuals from other subpopulations. This pattern exposes subpopulations (with localized
presence of alleles across genetic loci) to the same sources of environmental variation and opportunities, and in turn, perhaps predicts similar evolutionary trajectories for phenotypes of species-level importance. In his seminal work on the genetic structure of populations, Wright emphasizes the ecological importance of subpopulation structure - on how it could have direct consequences for the spread of mutations, and adaptation.

Evolution of human populations aside, there have been a multitude of other studies on the evolution of populations of species. These studies (see Avise (2009) for a review) range from species of extreme conservation concern with scattered populations (or even individuals) over a geographical scape, with little or no hybridization and propagation, to invasive species that are uncannily successful in novel environments, constantly undergoing selection for sustenance and increased viability in these environments. These studies have hypothesized and tested various simplified models of population evolution – the simplest assuming panmixia, wherein all individuals in the population randomly mate and pass on their genes to further generations, to more complex models of population evolution, which assume panmixia at some point in the distant past, and divergence since into subpopulations, with sporadic intervals of gene flow between subpopulations, and so on. For instance, researchers have identified post-glacial distributions of species from an ancestral panmixia, into current subpopulations (Nason et al. (2002), Starkey et al. (2003)), traced the ancestry of populations that have been isolated by geographical barriers to gene flow (Nason et al. (2002)), attempted to understand patterns of ancestral migration and its lack thereof with applications to species of conservation concern (Mockford et al. (2007), Mockford et al. (1999), Rubin et al. (2001)), delineated species histories and the distribution of levels of genomic diversity across geographical locations, and ancestral migrations using global evidence of population differentiation (Pemberton et al. (2008), Reich et al. (2009), Cruaud et al. (2011), Templeton (2002)), to name a few (see Avise (2009) for a review).

The presence of genetic subpopulation structure, as detected in numerous studies, has direct consequences for downstream genetic analyses. For instance, population structure in genome wide association studies (GWAS) leads to excessive false positives in identifying associations between loci. Several studies have identified this issue and attempted to address it in a variety
of ways (see Xu and Shete (2005), Price et al. (2006), Marchini et al. (2004), etc), but central to all these methods is the identification of subpopulation structure.

Models of evolution are confounded not just by demography, but also by varying levels of selection, mutation, non-random allelic association, and recurrent migration. Simultaneously estimating all these parameters is statistically difficult, and often intractable, even under the simplest of models. Researchers were quick to realize the enormity of this problem and focusing on building approximate models, with few summary statistics that subsume several parameters. Hence a reduced definition of population genetic structure can be explained as quantifying the overall genetic variation in individuals of the same species primarily in terms of their allele or genotype frequency distribution across one or more genetic loci (see Weir and Hill (2002) for review). But regardless of the issues in inferring subpopulation structure, a very common result of ignoring it is termed the Wahlund Effect, wherein perceiving multiple subpopulations as one, affects the Hardy-Weinberg proportions of genotype and allele frequencies, and artificially creates signals of heterozygote deficiency in the absence of Hardy-Weinberg Equilibrium. Consequences of this over-arching Wahlund Effect include linkage (and association), bottlenecks, and cryptic relatedness.

Subpopulation structure can also create an artificial signal of a population bottleneck (Wakeley (2000), Chikhi et al. (2010), etc). This issue comes to light particularly in conservation genetics, where endangered species are undergoing bottlenecks as it is. Hence inference of the bottlenecks faced by these species is going to be inherently biased if the issue of subpopulation structure is not addressed first.

Subpopulation structure also plays a key role in estimating intraspecies pairwise genetic relatedness (which is thereon used in GWAS and other methods mentioned above). Genetic relatedness is measured in terms of the probability that two randomly sampled genes at a genetic locus are identical by descent (IBD - see Weir et al. (2006) for a review). But descent, as mentioned with respect to bottlenecks, could be recent, or deep. Recent relatedness is relatedness between two individuals owing to the immediately previous few generations. Deep relatedness is relatedness measured owing to an ancient admixture event, with gene flow and subsequent incorporation of genes from different genetic subpopulations, and inbreeding since.
Several methods have been proposed to estimate this pairwise genetic relatedness, but the issue of subpopulation structure is again central. Either estimators assume that alleles were sampled from the same one ancestral subpopulation (thus ignoring the effects of stratification), or account for stratification with summary statistics (such as Wright’s $F_{ST}$). If population structure is ignored estimates of pairwise relatedness are invariably going to be biased (upward or downward, depending on the estimator - see Anderson and Weir (2007) for review). I will address this issue of bias in estimation of relatedness due to population genetic structure in Chapter 3.

The goals of this dissertation are to (1) detail the state of the art in estimating population genetic structure, (2) develop two methods for more efficient estimation of population structure using large-scale genomic data, and (3) to utilize this information in estimating pairwise genetic relatedness between two randomly sampled individuals. I supplement these studies with simulations under a host of evolutionary scenarios, and empirical examples from a species of conservation concern (Emys blandingii – a semi aquatic turtle species that is listed as threatened or endangered across its primary range in the midwestern United States), and from humans.

The remainder of this introductory chapter is organized as follows: I detail methods for estimating population genetic structure, and utilize those methods to infer structure in the human dataset described below. I then discuss the pros and cons of each method, highlighting the need for a more robust statistical architecture for inferring and interpreting structure. I end with an overview of applications and interpretations using each method, building a case for the rest of this dissertation.

For the purpose of this introduction, I used the publicly available microsatellite dataset of Rosenberg et al. (2002), which details genotypes of 1056 individuals sampled from 52 geographic populations across the world at 377 microsatellite loci. All this data was obtained using the HGDP-CEPH Human Genome Diversity Cell Line Panel (Rosenberg lab, Stanford University).

1.2 Methods

Several methods have been developed over the years to understand population genetic structure, each with its own pros and cons - Wright’s $F$-statistics, model-based clustering
methods, non-model-based clustering, to name a few. In this introductory chapter, I detail the most commonly used of these methods, and discuss their applications, and differences between them.

1.2.1 F-statistics and AMOVA

One traditional formulation of the presence of population genetic structure is measured in terms of summary statistics, commonly termed $F$ statistics ((Wright, 1951)). The estimation of $F$ statistics assume that all current subpopulations were derived from an ancestral population that was in Hardy-Weinberg Equilibrium (HWE), and in Linkage Equilibrium (LE). These subpopulations are assumed to have undergone the same process of evolution since divergence. There are several definitions of $F$ – but what was originally called the genetic correlation or inbreeding coefficient, has come to be known more generally as the coefficient of differentiation. Differentiation is a tricky concept though, as there are different levels of differentiation (within an individual, between individuals of the same subpopulation, between individuals of different subpopulations, etc). Three coefficients, or $F$-statistics are at hand - (1) the inbreeding coefficient, or $F_{IT}$ (also referred to as $F$), which is the correlation between genes within individuals in the total population, (2) the coancestry coefficient, $F_{ST}$ (also referred to as $\theta$), or the correlation between genes in individuals in the SAME subpopulation, (3) and $F_{IS}$ (also referred to as $f$), which is the correlation between genes in individuals within subpopulations. These $F$-statistics have analogous definitions (see Nei (1973), Weir and Cockerham (1984), Whitlock (2011) for a review).

The most general definition of $F$ is the proportion of reduction in heterozygosity, compared to a population in Hardy-Weinberg Equilibrium (HWE) (see Hartl and Clark (1997)).

Several variants of this generalization have been proposed (Nei’s $G$ statistic ((Nei, 1973), Weir and Cockerham’s $\theta$ ((Weir and Cockerham, 1984)), Jost’s $D$((Jost, 2008)), Hamrick and Godt’s $G_{ST}$ ((Hamrick and Godt, 1997)), etc., which offer corrections for bias due to sample sizes, equal weighting for alleles (i.e substituting allele frequencies with weighted allele frequencies based on their relative rarity), to averaging across alleles (and/or loci), etc. All these versions of $F$ statistics are affected by effective population sizes, and on population history (see

In general (see Holsinger and Weir (2009) for derivations), $F$ statistics range from 0 to 1, with 0 indicating no differentiation, to 1 indicating complete differentiation. In practice, an $F_{ST}$ of $0.00 - 0.05$ indicates low differentiation, $0.05 - 0.15$ indicates moderate differentiation, while $F_{ST} > 0.15$ indicate high levels of differentiation (see Hartl and Clark (1997) for details of observed $F_{ST}$’s in natural populations).

I estimated Nei’s $G_{ST}$, Hamrick and Godt’s $G_{ST}$, Jost’s $D_{ST}$, and Weir and Cockerham’s $\theta_{ST}$ using the R packages mmod (Winter (2012)) and pegas (Paradis (2010)). A comparison of these differentiation estimates over the 377 loci is shown in 1.1. I performed Fisher’s exact test of differentiation (Raymond and Rousset (1995)) across all the loci with 2000 replicates, and obtained a p-value of 0.0005, which indicates significant levels of genotypic differentiation in this data-set. Means of these differentiation statistics revealed a variety of scales of subdivision across 377 loci, as shown in in Table 1.1, with a mean Nei’s $G_{ST}$ and $F_{ST}$ indicating low levels of differentiation ($\approx 0.05$), and the $G_{ST}$ of Hamrick and Godt ((Hamrick and Godt, 1997)) and Jost’s $D_{ST}$ showing higher differentiation across loci($\approx 0.2$).

The AMOVA (Analysis of MOlecular VAriance - as designated by Excoffier et al. (1992)) framework draws from a rich literature on genetic distances, and $F$ statistics. AMOVA is an alternate method to estimate correlation $F$ (here $\Phi$) statistics as discussed above. AMOVA also provides a statistical framework for hypothesis testing of different patterns of subpopulation structure, and is hence discussed here. The focus of the original seminal work on AMOVA was to derive a framework for partitioning total variance in allele frequencies (across multiple loci) within and among different strata (within populations, among populations, within subpopulations, and among subpopulations) by defining genetic distances between haplotypic data. But this method has been generalized since to derive what are called $\Phi$ statistics, analogous to $F$, $G$ and $\theta$ defined in the previous section.

Consider a general distance metric, $D^2_{ij}$, which is the (squared) genetic distance between two individuals (or genotypes) $i$ and $j$. The Sum of Squared Deviations (SSDs) with respect to the ‘average’ genotype can hence be written as a function of the distances, $\delta^2_{ic}$ and $\delta^2_{jc}$, where
c is the mean genotype (or centroid of Euclidian space - Excoffier et al. (1992)):

$$SSD_{total} = \frac{1}{2N} \sum_{i=1}^{N} \sum_{j=1}^{N} \delta_{ij}^2$$ (1.1)

Correspondingly, this $SSD_{total}$ can be partitioned among different strata (within an individual, individuals within a subpopulation, individuals between subpopulations) as

$$SSD_{total} = SSD_{ST} + SSD_{IS} + SSD_{IT}$$ (1.2)

where $SSD_{ST}$ is the $SSD$ within subpopulations, $SSD_{IS}$ is the $SSD$ within an individual relative to a subpopulation, and $SSD_{IT}$ is the $SSD$ among individuals in the total population.

Corresponding Mean Square Deviations (and variance components) can be derived by dividing these $SSD$ terms by the degrees of freedom, and generate equations for correlation coefficients, or $F$-statistics (as described above). Only, these are termed $\phi$ statistics, and are defined in terms of the additive variances ($\sigma^2$, for $a$ defining variance between subpopulations, $b$ defining the variance component between individuals within the subpopulations, and $c$ defining variance within an individual in the total population) as:

$$\sigma^2 = \sigma_a^2 + \sigma_b^2 + \sigma_c^2$$ (1.3)

$$\phi_{ST} = \frac{\sigma_a^2}{\sigma^2}, \phi_{IS} = \frac{\sigma_b^2}{\sigma_b^2 + \sigma_c^2}, \phi_{IT} = \frac{\sigma_a^2 + \sigma_b^2}{\sigma^2}$$ (1.4)

where $\phi_{ST}$ is the correlation between genotypes within a subpopulation relative to the total population, $\phi_{IS}$ is the correlation between genotypes within subpopulations, and $\phi_{IT}$ is the correlation between genotypes of individuals relative to the total population, which are all analogous to the $F_{ST}$, $F_{IS}$ and $F_{IT}$ described before.

Excoffier et al. (1992) extended this method of Weir and Cockerham (1984) to obtain the same $SSD_{total}$, but partitioned at different strata instead (subpopulations, populations and total population). This yields estimates for $\phi_{ST}$, $\phi_{CT}$, and $\phi_{SC}$ in a similar fashion, with $\phi_{ST}$ measuring the correlation between randomly sampled genotypes within subpopulations, $\phi_{SC}$ measuring the correlation among subpopulations within a population, and $\phi_{CT}$ measuring the correlation between populations in the total population.
Global AMOVA estimates (calculated as weighted average over all 377 loci - see Table 1.2) were performed using the geographical location of each sample to indicate the subpopulation structure within the human dataset (America, East Asia, Central South Asia, Europe, Middle East, Oceania, and Africa). This analysis indicated that most variation (94.06%) in the data was captured within subpopulations. Global $\phi_{ST}$ was estimated at 0.0594 (0.05479 – 0.06435 99% CI over 20000 bootstrap reps). $\phi_{CT}$ was estimated at 0.03582 (0.03166 – 0.04032 99% CI), and $\phi_{SC}$ to be 0.02444 (0.02305 – 0.02588 99% CI). Rosenberg et al. (2002) also report the results of other AMOVA’s that were performed by grouping populations of individuals by other epidemiological factors, all of which reflect a similar pattern, with most of the variation explained within geographical populations.

1.2.2 Model-based Clustering Methods

Regardless of the purpose of summary statistics, the methods described above share a common problem - how do we ‘know’ that a group of individuals form a genetic subpopulation? Or a population? These methods assume that individuals sampled from a relatively smaller geographical regime share greater common ancestry, and hence more likelihood of being derived from the same genetic subpopulation. This may be true if populations are known to have diverged from a known source, and the current sampling of individuals involves direct descendants of this source. The use of summary statistics or AMOVA as described above also does not tell us intuitively how to subdivide a total population into genetic subpopulations. The use of $F$ statistics, for instance, make an unreasonable assumption about the population history - that all subpopulations were derived from the same ancestral source in HWE and LE, and have undergone the same evolutionary processes since divergence. More likely than not, population histories of species are more complicated - ancestral divergences (see Edwards and Beerli (2000) for a perspective on within species ancestral divergence times), ample ancestral migration and subsequent incorporation into the current gene pool (see follow up literature on human ancestral divergences and introgression, including inter-species gene flow - Henn et al. (2012), Eriksson and Manica (2012), Innan and Watanabe (2006), Ramachandran et al. (2005), etc), sporadic migration events between long intervals of fixation (see Vuilleumier et al. (2008),
or Takahata (1991) for a mathematical formulation), complete divergence with no migration since (see Wakeley (2000), or Hey and Nielsen (2004) for alternate perspectives on population divergences with and without migration), to name a few. In essence, many wild populations are NOT homogeneous, and are comprised of individuals with mixed ancestries. Their genotypes are in turn a mosaic of alleles derived from multiple ancestral subpopulations. Genetic sampling of individuals, largely seen as random, need not pick individuals which are necessarily directly related by common decent - they may have shared ancestry at some point in the distant past, but their genotypes may indeed be derived from multiple ancestral subpopulations. In the absence of this knowledge, assuming that individuals derive from the same subpopulation leads to incorrect inferences of allele frequencies and estimates of genetic variation and diversity in general. These two issues, (1) unreasonable assumptions on the source population and divergence, and (2) unsupported \textit{a priori} assumptions on genetic population structure, are the standing motivations for the many clustering methods, including the methods introduced in this dissertation’s Chapter 2 - how do we statistically assign individuals sampled from a geographical population to ancestral genetic subpopulations?

In this context, I introduce a commonly used analytical method for inferring subpopulation structure - the admixture model, and its proponents, STRUCTURE (Pritchard et al. (2000b), Falush et al. (2003), Falush et al. (2007)), ADMIXTURE (Alexander et al. (2009)), FRAPPE (Tang et al. (2005)), STRUCTURAMA (Huelsenbeck and Andolfatto (2007)) and the methods of Smouse et al. (1990). Assume that \( I \) individuals have been sampled from a total population comprised of \( K \) subpopulations, and genotyped at \( L \) loci. Let \( A = \{a_1, a_2, \ldots, a_L\} \) be the set of allelic variants at a locus \( l \in L = \{1, 2, \ldots, L\} \). We define the frequency \( p_{kla} \) as the frequency of an allele \( a \in A \) at locus \( l \in L \) in subpopulation \( k \in K = \{1, 2, \ldots, K\} \).

Let \( \eta_{ik} \) be the proportion of an individual \( i \)'s genotype that is derived from subpopulation \( k \). Hence each individual, \( i \) has an associated vector of subpopulation admixture proportions, \( Z_i = [\eta_{i1}, \eta_{i2}, \ldots, \eta_{iK}] \).

Assuming that these admixture proportions (for an individual \( i \)) are sampled from a Dirichlet distribution (with parameter, \( \alpha \)), STRUCTURE iteratively performs Markov-Chain-Monte-Carlo (MCMC) repetitions to update estimates for the (1) subpopulation allele frequencies, (2)
ancestral population of origin of each individual, and (3) admixture proportions. Pritchard et al. (2000b) suggest an ad hoc approach to infer the best value of $K$, or the total number of ancestral subpopulations, by computing the posterior distribution on $K$ using Bayes’ Rule as

$$Pr(K = k) = \frac{Pr(X \mid K = k)}{Pr(X \mid K = 1) + Pr(X \mid K = 2) + \ldots + Pr(X \mid K = k)}$$

(1.5)

where $X$ is the total observed dataset, comprised of the observed genotypes and unobserved structure. Alternately, $K$ can also be inferred using Bayesian deviance, measured as

$$D(X, K) = -2 \log Pr(X \mid K)$$

(1.6)

Both methods rely on the marginal likelihood, $Pr(X \mid K)$, which is provided by STRUCTURE ((Pritchard et al., 2000b)). The greater the value of this marginal likelihood, the better ‘fit’ of the data to the model with a chosen $K$. Researchers have since developed other ad hoc methods to infer the best fitting $K$ - for instance, the method of Evanno et al. (2005) computes the second order rate of change of this logarithmic marginal likelihood. The best fit for $K$ to the data is identified by computing the $K$ which provided the largest second order rate of change in the logarithmic marginal likelihood of the data between successive values of $K$ (i.e. $\Delta K = |Pr'(X \mid K + 1) - Pr'(X \mid K)|$). The best $K$ is then inferred at the step $K$ to $K + 1$ which showed a maximal increase in the marginal likelihood of the data. Evanno et al. (2005) also suggest performing multiple runs of STRUCTURE using several initializations, and computing a mean value for this $\Delta K$, to pick the best fit model.

Alternate solutions to the same admixture model include the method of Tang et al. (2005), implemented in the program, FRAPPE. FRAPPE implements an Expectation Maximization algorithm (similar to that of MULTICLUST, described in Chapter 2), to obtain the best parameter set that fits the data, conditioned on $K$. Alexander et al. (2009) accelerated the process of estimating parameters using quadratic approximations by applying Newton’s method to the likelihood equation (see Chapter 2 for a detailed description). They also proposed a further acceleration by using Quasi-Newton approximations (sensu Zhou et al. (2011)). A known caveat of this method is that it is restricted to SNP datasets, which theoretically only possess two states per allele. I extend this method to the theoretically infinite model of allelic variation in MULTICLUST (see Chapter 2).
To illustrate the umbrella of these methods of model-based clustering, I performed 4 replicate runs of STRUCTURE on the human microsatellite dataset by varying the number of sub-populations between $K = 1$ and $K = 10$, under the admixture model, and with uncorrelated allele frequencies (i.e. the allele frequencies in each subpopulation are drawn independently from a Dirichlet distribution). The Dirichlet parameter, $\alpha$ was allowed to be inferred from the data, and the best model ($K$) was picked using the method of Evanno et al. (2005) described above. Fig1.3 shows a plot of the $\Delta K$ values from all four runs and indicates the putative number of subpopulations present in the human dataset is $K = 4$. Interestingly, this is different from $K = 5$ subpopulations as obtained by Pritchard et al. (2000b).

Several interesting patterns of human ancestral admixture emerge from analyses of admixture proportions, $\eta_{ik}$’s inferred from this dataset (see 1.4). For instance, while a large amount of concordance exists between population of sampling and population genetic structure, several introgression events are clearly noticeable. Mayan Mexican-American populations, for instance, have considerable amount of admixture with populations in Eurasia (Central-South Asia, and Europe). There seems to exist a geographical gradation in admixture as we traverse from Africa into the Middle East and into Europe, with populations in northern Africa (closer to the Middle East) having considerable admixture with European populations, while the remaining African populations stand out as a separate cluster. Oceanic populations also exhibit considerable admixture, with ancestries derived from Africa, East Asia, and the Middle East. Of note is that the admixtures in Oceania and American populations were not reported in Rosenberg et al. (2002) (except for being seen in Fig.1). The analysis of population structure using model-based clustering hence offers what was not previously inferred using summary $F$-statistics - insights into the ancestry of subpopulations, which are otherwise assumed geographical while calculating $F$-statistics.

1.2.3 Other Clustering Methods

Principal Components Analyses (PCA) have been utilized in the context of identifying population genetic structure (eg. see PCagen, Goudet), but the goal of a PCA is to identify orthogonal directions of maximal variance in the multilocus genotype data. An allied semi-
model based solution (DAPC) to the problem of inferring subpopulation structure and cluster assignment was proposed by Jombart et al. (2010), and implemented in the program suite, adegenet (Jombart 2008)) for R. DAPC performs a PCA first, identifying directions of maximal variance. Then the most informative directions (PC’s) are picked, and a K-means clustering is performed on the data, to maximize the variation between \( K \) groups, by incrementally increasing \( K \). The most likely \( K \) is then identified by using a BIC by comparing different K-means clustering solutions. Then a discriminant analysis is performed on this data (retained PC’s which explain the most variance, and clusters which have maximum between cluster variance, but minimum within cluster variance). The theory behind the workings of DAPC is similar to the ANOVA model introduced before (see Jombart (2008)). DAPC also provides membership probabilities for each individual to each identified group (or subpopulation), which can be equated to admixture proportions provided by STRUCTURE ((Pritchard et al., 2000b)).

DAPC analyses of the human microsatellite dataset resulted in inconclusive value for \( K \) by K-means clustering along the first 4 PC’s that were retained (see Fig.1.3,Fig.1.5). In order to obtain a comparable estimate of population structure, as determined by STRUCTURE above, I used an \( a \) priori \( K = 4 \) for all further downstream analyses. A plot of inferred admixture proportions at \( K = 4 \) is shown in Fig.1.6. The most admixed individuals detected at \( K = 4 \) are derived from populations in Central South Asia (from parts of Uygur (China), and Hazara (Pakistan),Fig. 1.7), which was also reported by Rosenberg et al. (2002). None of the other admixture events were significantly detected at \( K = 4 \), which could be a consequence of ignoring higher values of \( K \), and higher principal coordinates in these analyses. But for the most part, population structure identified by DAPC agreed with the same patterns identified by STRUCTURE (Fig.1.4).

1.3 Discussion

Population differentiation and the identification of genetic subpopulations from multi-locus genotype data is an active field of research, with numerous authors adding to the many statistical methods, each with their own pros and cons. The purpose of this chapter was to summarize these methods and apply them to an empirical dataset to highlight the major issues with
performing statistical inference with each method. Some of these issues are: (1) Existing methods for identifying genetic population structure make assumptions about divergences from an ancestral source population, which are not necessarily true, (2) Existing methods for inferring subpopulation structure do not offer reliable estimates of admixture proportions across multiple runs of the same method, and (3) These methods are conditioned on a priori knowledge of subpopulation structure, which are not necessarily known. Incorrect subpopulation assignments, and inference of genetic population structure using any of the above methods could potentially lead to biased downstream genetic analyses, several of which depend on the delineation of population genetic structure in the sampled data. Several seminal works have attempted to identify the magnitude of this problem in building summary statistics to model-based and non-model based inference of subpopulation structure (see Holsinger and Weir (2009), and Lawson and Falush (2012) for review of methods).

A major caveat of estimating $F$-statistics is that it inherently assumes that genetic subpopulations are equivalent to the geographical locale of sampling, unless ‘grouped’ otherwise (by hypotheses, or alternate methods that infer the genetic subpopulations first). Hence, regardless of the presence of ancestral divergence and/or admixture, current allele frequencies (and estimates of heterozygosities, and differentiation thereon) are all inherently biased by this very key assumption. This is a common issue, which is also seen in utilizing non-model methods, and the ANOVA (or AMOVA) framework. To subvert this issue, alternate methods have been developed to parametrically estimate the ancestry (and the current subpopulation structure) within individuals of a species. Another standing issue which I have demonstrated (see Table1.1, Fig.1.1), has to do with interpretation of inferred $F$ statistics. While an ad hoc method of adjudging $F$ statistics was mentioned in a previous section, there is plenty of variability in estimates of ‘similar’ or analogous $F$ statistics, even within the same dataset. The issue of interpretation has been previously dealt with (Holsinger and Weir (2009)), but problems with interpretation remain.

Another frequent criticism with estimating $F$ statistics (eg. Wright’s $F$ (Wright, 1951), Nei’s $G$ (Nei, 1973), Weir and Cockerham’s $\Phi$ (Weir and Cockerham, 1984)) is that they are not sensitive to the allelic diversity within subpopulations (Jost (2008)). To subvert this issue,
I also estimated Hamrick and Godt’s $G$ (Hamrick and Godt, 1997) and Jost’s $D$ (Jost, 2008), and plotted all the above statistics against the number of alleles at each locus. As noted by Jost (2008), all the afore mentioned statistics ((Wright, 1951), (Weir and Cockerham, 1984), (Nei, 1973)) are difficult to interpret with increasing allelic diversity, with low differentiation indicated with higher allelic diversity within subpopulations. On the other hand, both the differentiation statistics of Hamrick and Godt (1997) and Jost (2008) are sensitive to allelic diversity, with increased differentiation indicated with increased diversity. These plots are shown in Fig.1.2.

Model-based approaches come with their own pros and cons. The pros include the ability to infer admixture and subpopulation allele frequencies, which aid in building hypotheses for testing models of evolutionary history of a species. But simplified models come with assumptions, such as assuming HWE, and the near neutrality of genetic markers under study. While it is reasonable to assume that repeat markers (SSRs, or microsatellites) in intronic regions are essentially neutral, a lot of recent studies have indicated selection on microsatellites (see Selkoe and Toonen (2006) for a review on choice of markers). Another common assumption is the genetic independent segregation of markers under study - or linkage equilibrium (LE), which allows these models to make multiplicative assumptions on the probability distribution of subpopulation allele frequencies (equilibrium assumptions are also true of several downstream analyses from using $F$-statistics). This assumption on the admixture model was relaxed by Falush et al. (2003), using a Markov Chain This model requires the independence assumption between individuals, which could be relaxed at the risk of increasing the number of parameters, which would correspondingly increase computational time and space requirements. Similar adaptations to the stochastic sampling process and subsequent Bayesian estimation (BAPS) have also been implemented by Corander and Marttinen (2006).

Another issue with model-based techniques that utilize MCMC methods arises from not allowing the entire stationary distribution to be reached before sampling from it to perform subsequent updates. This issue is commonly referred to as that of ‘poor mixing’ of the MCMC, wherein the sampler is not traversing the sample space sufficiently. The sampler either gets stuck at local maxima, or just isn’t ‘mixed’ or sampled for long enough to obtain a good
distribution over all the maxima (Pritchard et al. (2000b), Falush et al. (2003), Falush et al. (2007), Corander and Marttinen (2006), Corander et al. (2006), etc). Working along the guidelines of Pritchard et al. (2000b), I performed 4 separate runs of 100,000 burn-in’s and MCMC reps, for all the analyses. Gilbert et al. (2012) suggest at least 20 runs of STRUCTURE prior to using the method of Evanno et al. (2005), which would increase the time requirements for this problem substantially (potentially adding weeks of run time on a server). This issue of failure to approach stationarity is conveniently subverted under maximum likelihood frameworks.

A more dire issue with utilizing these methods arises with respect to inconsistency in the estimation of $K$, a problem which also arises in my own re-analyses of the human microsatellite dataset here (see Fig.1.3). While this issue could be approached with multiple initializations and iterations, performing statistical inference on the most likely value of $K$ is still up to the researcher’s choice of ad hoc methods (eg. Evanno et al. (2005)). And oftentimes, this inference on $K$ is not repeatable. A recent meta-study by Gilbert et al. (2012) on inconsistencies in the inferred choice of $K$ in 30 other published studies revealed repeated incorrect/inconsistent classifications in 30% of these studies. In particular, inconsistent analyses of population structure in species of conservation concern raises questions for restoration programs, which largely rely on quantification of genetic differentiation.

A common but important problem that affects genetic analyses, is the presence of erroneous and/or missing data. STRUCTURE ((Pritchard et al., 2000b)) and BAPS ((Corander and Marttinen, 2006), (Corander et al., 2006)) ignore missing data in estimating admixture proportions and allele frequencies, as do several other methods described above. This approach is legitimate, if and only if the presence of the missing alleles is independent of what allele was present at that locus. A common prescription for handling missing data is to remove individuals with a great degree of missing data from the overall analyses to prevent unwanted biases. An alternate, but much more robust solution to this problem is to statistically infer missing data as parameters in the model, which is inherently easy to implement as part of the Expectation Maximization (EM) algorithm (Dempster et al. (1977)).

The presence of ‘higher’ degrees of population genetic structure (for eg. Isolation By Distance) also distorts the identification of distinct subpopulations in a total population. This
issue has been analyzed in detail particularly with the program STRUCTURE ((Pritchard et al., 2000b)) in Schwartz and McKelvey (2009).

This dissertation is organized as follows: A standing motivation for inferring population genetic structure comes from a conservation genetics project on the imperiled Blanding’s turtle (*Emys blandingii* - Chapter 1). Given all the caveats and issues which I identify in this introduction chapter, need a robust, fast, and consistent tool to infer subpopulation structure is of great concern, a major component of which is addressed in Chapter 2. As a glimpse into key inferences that can be made based on estimates of subpopulation allele frequencies and admixture proportions, Chapter 3 explores the issue of estimating pairwise genetic relatedness in structured populations.
1.4 Tables and Figures

<table>
<thead>
<tr>
<th>Estimator Used</th>
<th>$F$-Statistic</th>
</tr>
</thead>
<tbody>
<tr>
<td>Nei (1973)</td>
<td>0.055</td>
</tr>
<tr>
<td>Jost (2008)</td>
<td>0.207</td>
</tr>
<tr>
<td>Hamrick and Godt (1997)</td>
<td>0.129</td>
</tr>
<tr>
<td>Weir and Cockerham (1984)</td>
<td>0.053</td>
</tr>
</tbody>
</table>

Table 1.1 Mean $F$ statistics estimated from the human microsatellite data. Of note is the variability in estimates of ‘similar’, and analogous measures of differentiation.

<table>
<thead>
<tr>
<th>Source</th>
<th>Sum of Squares</th>
<th>Variance</th>
<th>% Variation</th>
</tr>
</thead>
<tbody>
<tr>
<td>Among Populations</td>
<td>10683.355</td>
<td>5.243</td>
<td>3.582</td>
</tr>
<tr>
<td>Among subpopulations within populations</td>
<td>12016.476</td>
<td>3.451</td>
<td>2.358</td>
</tr>
<tr>
<td>Within subpopulations</td>
<td>272468.480</td>
<td>137.679</td>
<td>94.060</td>
</tr>
</tbody>
</table>

Table 1.2 AMOVA results using 7 geographical groups as subpopulations.
Figure 1.1 Distributions of Nei’s $G_{ST}$ (mean = 0.055), Jost’s $D_{ST}$ (mean = 0.207), Hamrick and Godt’s $G_{ST}$ (mean = 0.129), and Weir and Cockerham’s $\phi_{ST}$ (mean = 0.053) over 377 loci for the human microsatellite data.
Figure 1.2  Distributions of Nei’s $G_{ST}$ (mean = 0.055), Jost’s $D_{ST}$ (mean = 0.207), Hamrick and Godt’s $G_{ST}$ (mean = 0.129), and Weir and Cockerham’s $\phi_{ST}$ (mean = 0.053) over number of unique alleles across the 377 loci for the human microsatellite data.
Figure 1.3  Plot of $\Delta K$ values from four runs of STRUCTURE on the human dataset, varying $K$ from 1 to 10 are shown in the top panel. The most likely value for $K$ was identified at $K = 4$. Plot of BIC estimates using DAPC to infer the number of clusters in the same dataset are shown on the bottom. The results are inconclusive, with the BIC still falling post $K = 50$. These plots illustrate a standing issue with inference of the true $K$ in a given dataset.
Figure 1.4  Plot of admixture proportions at the population level, at $K = 4$, identified using STRUCTURE is shown in the top panel. Plot of membership probabilities, which can be equated to admixture proportions, as identified by DAPC at $K = 4$ are shown in the bottom panel. For a detailed list of legends, see Rosenberg et al. 2002. Note the similarity in patterns of admixture, as indicated by the two column graphs.
Figure 1.5  DAPC plot of Principal Axes 1 and 2, showing the most variation in the microsatellite genotype data. The first two PC’s show most variation explained in three clusters.
Figure 1.6 DAPC plot of Principal Axes 1 and 2, at a chosen $K = 4$. 
Figure 1.7  Admixture proportion plot at $K = 4$ for most significantly admixed populations. These were identified as Uygur (1306), and Hazara (129).
CHAPTER 2. POPULATION GENETICS OF BLANDING’S TURTLE IN THE MIDWESTERN UNITED STATES

2.1 Abstract

Blanding’s turtle (*Emys blandingii*) has declined substantially in North America due to anthropogenic activities, leaving populations smaller and increasingly fragmented spatially. We sampled 212 turtles to evaluate variation at eight microsatellite loci within and among 18 populations of *E. blandingii* across its primary range in the midwestern United States (Illinois, Iowa, Minnesota, and Nebraska). All loci and populations were highly polymorphic. Our analyses also detected considerable genetic structure within and among the sampled localities, and revealed ancestral gene flow of *E. blandingii* in this region north and east from an ancient refugium in the central Great Plains, concordant with post-glacial recolonization timescales. The data further implied unexpected ‘links’ between geographically disparate populations in Nebraska and Illinois. Our study encourages conservation decisions to be mindful of the genetic uniqueness of populations of *E. blandingii* across its primary range.

2.2 Introduction

Barriers to migration of individuals can isolate populations over evolutionary time and even elicit speciation (Byrne et al. (2008)). In the interim, displaced populations can establish distinct clades within species, as indicated by fossil and genetic evidence (Amato et al. (2008), Sommer et al. (2009), Ursenbacher et al. (2006)). Range size changes, eliciting diversification and population bottlenecks, are greatly influenced by abiotic, often geological, events and yield distinct genetic patterns. More recently, anthropogenic forces have disturbed habitats and displaced taxa (Hoffmann et al. (2010), Stuart et al. (2004), Sutherland et al. (2012)). This
contemporary uptick of anthropogenic pressures and recognition of the relatively sharp decline in species numbers has elicited an increase in population genetic studies of organisms of conservation concern to illuminate evolutionarily significant units and provide added guidance for management to preserve unique genetic lineages (reviewed in Avise (2010), Frankham et al. (2002)).

Fossil records of relict vertebrate populations in the Great Plains of North America suggest that repeated glacial recession and drastic climate change after glacial advance displaced various species (Smith (1957)). Termed ‘xerothermic’ or interglacial periods, these ages, the last one from 9000-4000 YBP, were marked by aridity and warming that Smith (1957) conjectured should have forced heat-intolerant species to alter their geographic ranges. Smith (1957) also hypothesized, based on the presence of current disjunct relicts of vertebrate species and evidence from skeletal remains, that additional eastern disjunct populations of several species once existed as post-glacial relicts during the latest xerothermic period, but are now extinct. Since Smith’s proposals, advances in molecular phylogeography and paleogeography have yielded more concrete evidence of the establishment of these relict populations due to glacial activity in the midwestern United States (e.g., Amato et al. (2008), Janzen et al. (2002), Weisrock and Janzen (2000), Wooding and Ward (1997)). Such work is particularly important for vulnerable taxa like turtles, as recent and drastic habitat changes have extirpated multiple chelonian species over the last century (Shaffer (2009)). Understanding the current genetic diversity and historical distribution patterns of turtles is, thus, essential to making more informed conservation and management decisions; indeed, identifying genetic discontinuities across the landscape is one of three crucial needs in studies of turtle conservation genetics (Alacs et al. (2007)).

Blanding’s turtle (*Emys* [formerly *Emydoidea*] *blandingii* – Fritz et al. (2011)) is semi-aquatic, with populations in Nebraska, Iowa, Minnesota, Illinois, Wisconsin, Michigan, and Ontario, along with geographically disjunct populations in Nova Scotia and the eastern seaboard (Ernst and Lovich (2009)). In the Pleistocene, *E. blandingii* occupied a much wider range across North America (Jackson and Kaye (1974), Mockford et al. (1999), Smith (1957), Van Devender and King (1975)). Archaeological records reveal post-glacial fossils of *E. blandingii* from Michigan, Maine, New York, and Ontario, as well as Pliocene fossils in Nebraska, Kansas, Oklahoma,
and Pennsylvania (Ernst and Lovich (2009)), which coincides with the idea of disjunct mid-western and eastern populations 8000–4000 YBP (Smith (1957)). Subsequent glacial recession, establishment of waterways connecting to the Great Lakes, and, more recently, anthropogenic factors, have contributed to a unique spatial distribution (Fig. 2.1; see Congdon and Keinath (2006)). Blanding’s turtle is listed as ‘endangered’ across its range and ‘threatened’ on the IUCN Red List (Rhodin and van Dijk (2011)), with this imperiled status mainly attributable to the combined effects of delayed maturity (Congdon et al. (1993), Congdon and Sels (1993)) and habitat destruction (e.g., Rubin et al. (2001)).

Genetic studies have been conducted on populations of *E. blandingii* primarily in Nova Scotia to New York (Howes et al. (2009), Mockford et al. (2005)) and near Chicago, Illinois (Rubin et al. (2001)). These studies reported little to no genetic structure among populations. However, at a broader scale, another study detected strong population genetic structuring between, but not so much within, the Great Lakes region and eastern North America in these turtles (Mockford et al. (2007), but see Spinks and Shaffer (2005)). Even so, a large-scale population genetics study across the range of Blanding’s turtle in the midwestern United States (hereafter defined as west of Lake Michigan), where extensive anthropogenic landscape alterations have occurred over the past 150 years, has yet to be undertaken. In this study, we examined the population genetics of Blanding’s turtles sampled from Nebraska (NE), Iowa (IA), Minnesota (MN), and Illinois (IL) using microsatellite loci. First, we tested the hypothesis (*sensu* Mockford et al. (2007)) that *E. blandingii* largely lacks spatial genetic structure in this region. Second, we examined the alternative hypothesis (*sensu* Smith (1957)) that current levels of genetic variation reflect post-glacial colonization (along watersheds in the Mississippi-Missouri River basins) of northern locales from refugia in Kansas and Missouri (Holman (1995), Van Devender and King (1975)). Paleo-hydro-geological data indicate that such watersheds (see Table A.3) formed after recession of the Laurentide Ice Sheet. At its peak, the Des Moines Lobe of this glacier separated our sampling locales into populations that are inside and outside this region (Ehlers and Gibbard (2008)). We further assessed these two hypotheses by estimating ancestral gene flow and coalescent times (Hey and Nielsen (2004)) and recent migration rates (Wilson and Rannala (2003)). Finally, we provide some perspective on how our findings might impinge on
conservation and management activities involving Blanding’s turtle in the midwestern United States.

2.3 Materials and Methods

2.3.1 Fieldwork

We caught 212 Blanding’s turtles in 18 locations in the midwestern United States (Fig. 2.1). We trapped several wetlands per location for at least 20 trap-nights. From our 12 most productive localities, we sampled 16 individuals per population on average (range 6 – 61; Table 1). Of the remaining six localities, five yielded only a single individual and one yielded two individuals. These six sites were not included in the heterozygosity, population differentiation, and ancestral and recent migration analyses (see below), being used only for inferring population genetic structure and admixture using STRUCTURE (Pritchard et al. (2000b). Tissue samples were either tail tips taken and stored in 95% ethanol or blood extracted from the cranial sinus or caudal vein using a 28-gauge syringe, placed in lysis buffer or EDTA, and stored at −80 C.

2.3.2 Genetic data generation

We extracted genomic DNA from each sample using either High Pure PCR Template Preparation Kit (Roche Laboratory) following the protocol outlined by the Roche Applied Science Chelex (Walsh et al. (1991)), or phenol-chloroform extraction. For genotyping, we used eight tetra-nucleotide repeat microsatellite markers previously developed for a related turtle (*Glyptemys muhlenbergii*) (King and Julian (2004); *GmuD21, GmuD55, GmuD87, GmuD88, GmuD90, GmuD93, GmuD95*, and *GmuD121*). Detailed amplification procedures can be found elsewhere (Howeth et al. (2008)); we genotyped PCR products on an Applied Biosystems 3100 Genetic Analyzer at the Iowa State University DNA facility using the ROX size standard (FAM and HEX dye sets). We genotyped negative and positive controls for each locus to assess false positives or negatives. We visualized and sized the results using GenoProfiler v. 2.1 (You et al. (2007)) and PeakScanner v.1.0 (Applied Biosystems), and we manually determined the diploid allele sizes to identify unique alleles. We noted indeterminate genotypes, possibly due to am-
plification errors, as missing alleles. We could not resolve 76 diploid genotypes (4.5%), which we then classified as missing data. Of the 212 genotyped individuals, 209 contained adequate information to be included in further analyses.

2.3.3 Genetic data analyses

We concentrated the majority of our population genetic analyses on the 12 well-sampled populations (Table 2.7). For additional analyses that do not require a priori specification of population of origin, we included the six other populations with minimal samples.

2.3.3.1 General population genetic analysis

We used Genepop v.4.0.10 (Raymond and Rousset (1995)) and GDA v.1.1 (P. and D. (2008), Weir (1996)) to estimate allele frequencies, observed and expected heterozygosities, \( N_m \) (average number of migrants per generation) by the private allele method, and pair-wise tests of linkage disequilibrium. We set dememorization numbers at 10,000 and performed 100,000 iterations for all permutation tests (exact tests) in Genepop. We tested deviance from Hardy-Weinberg equilibrium (hereafter, HWE) at each locus using FSTAT v.2.9.3.2 (Goudet (1995)). Because we evaluated HWE per locus, per population (8 loci x 12 populations), we used sequential Bonferroni to correct for multiple comparisons on the expected and observed heterozygosities (Rice (1989)). We also performed a test for null alleles using Microchecker v.2.2.3 (van Oosterhout et al. (2006)) because the observed heterozygosities and deviance from HWE suggested the presence of null alleles that could possibly skew the population genetics results. In so doing, we placed limits on allele sizes (repeat lengths) at each locus based on those reported in King and Julian (2004) for *E. blandingii*. Analyses at the population level detected the presence of excessive homozygosity in some loci (Table A.6), but overall \( GmuD95 \) was the most problematic locus and was identified by Microchecker as likely having null alleles. Therefore, we hereafter conducted all data analyses with and without the most homozygous locus (i.e., \( GmuD95 \)) as an assessment of the potential impact of null alleles.
2.3.3.2 Differentiation and population structure

We estimated $F_{st}$ between all pairs of the 12 well-sampled populations and calculated statistical significance (Weir and Cockerham (1984)) using 1,000,000 genotypic permutations in Arlequin (Excoffier and Lischer (2010)) followed by sequential Bonferroni correction for the 66 pairwise population comparisons (Rice (1989)). We also calculated other population-wide $F$-statistics ($F_{is}$, $F_{it}$, $F_{st}$) with 95% confidence intervals, after bootstrapping across all loci with 10,000 replicates in GDA v.1.1 (P. and D. (2008)). As defined by GDA, $F_{is}$ describes average genetic differentiation between 202 individuals within their sampling locations; $F_{it}$ quantifies genetic correlation within 202 individuals in the total population, and $F_{st}$ measures differentiation between individuals in the same sampled location with respect to the total population. We then analyzed population structure in several ways. First, we analyzed genetic differentiation due to linear geographic distance for the 12 well-sampled populations (Rousset (1997)). This isolation-by-distance analysis regressed estimates of pair-wise $\frac{F_{st}}{1-F_{st}}$ (Slatkin’s linearized $F_{st}$ distance) against the linear distance separating pairs of populations. We calculated this regression using a Mantel Test in GenAlEx v.6.2 (Peakall and Smouse (2006)), with 1,000 permutations to assess statistical significance. Second, linear distance is not always the best predictor of genetic differentiation, as different geographic and historical forces may contribute to large genetic differentiation even over very small spatial scales (reviewed in Avise et al. (1987)). To ascertain a potentially better phylogeographic predictor of genetic variance, we addressed multiple hypotheses using AMOVAs and comparing $AIC_c$ values for each model to assess fit (Excoffier et al. (1992), performed with 16,000 permutations across and within the sampled loci in Arlequin). We constructed three models that reflect putatively different genetic structures across the landscape: 1) linear geographical distance, 2) clustering of populations into groups based on current river basins/watersheds, and 3) clustering of populations into groups based on location inside or outside of the Des Moines Lobe of the Laurentide Ice Sheet. To test these models, we employed several grouping schemes (Fig. 2.1). For the first model, we used Geographic Distance Matrix Generator v.1.2.3 (Ersts (2010)) to calculate linear distances between populations from GPS coordinates for the collected specimens (Table 2.7). We clus-
tered populations within 100 linear km (based on geographic distribution of our sampling) of each other (see Table A.1 - Model1) to assess isolation by distance. Populations that fell into two clusters were resolved by grouping them with other populations that were the closest linear distance to them. For the second model, we grouped populations into watersheds (Midwest Natural Resources Group, www.epa.gov/Region5/mnrg/), yielding four groups spanning the Missouri River Watershed, Upper Mississippi River Watershed, Minnesota River Watershed, and Illinois River Watershed - Southern Lake Michigan Crescent Watershed (see Table A.2 – Model2). We designed this model knowing that these semi-aquatic turtles can migrate several kilometers terrestrially (Ernst and Lovich (2009)). Individuals trapped from locations separated by small terrestrial areas were counted as part of the same watershed. We constructed the third model to assess population structure relative to current watersheds, as in the second model, but also incorporated separation by the Last Glacial Maximum (LGM) limit of the Laurentide Ice Sheet (Ehlers and Gibbard (2008)). This model sorted our sampled locations into five groups, essentially similar to the second model but with the Upper Mississippi River Watershed divided into regions located inside (or north) of the LGM limit and outside (or south) of it (see Table A.3 – Model3). For each model, we calculated pairwise $F$-statistics and made comparisons with an exact test of population differentiation (Goudet (1995), Raymond and Rousset (1995)), where ‘populations’ are the defined groups for each model. This test determines the probability that ‘$k$’ genotypes are distributed among ‘$r$’ populations by using an $r \times k$ contingency table. We explored potential states of the contingency table using a Markov chain with 16,000 permutations of genotypes among populations. We compared AMOVA results from the above population differentiation models to determine the better predictor of genetic variance using $AIC_c$ (Burnham and Anderson (1998), Halverson et al. (2008)). Lastly, we included all 209 individuals with adequate genotype information from all 18 populations to explore population structure with STRUCTURE v.2.2 (Pritchard et al. (2000b); STRUCTURE is able to infer structure without prior information on sampling locations) using the admixture model and specifying no a priori models of subpopulation structure. We allowed the Dirichlet parameter for the degree of admixture ($\alpha$) to be inferred from the data, with an initial value of 1.0 and uniform priors for all populations. To determine correlated allele
frequencies and to compute probability of the data to estimate \( K \) (the most likely number of putative populations), we performed 20 runs for each value of \( K \) (1 – 18) with 10,000 MCMC repetitions. In each case, we allowed a burn-in period of 10,000 for \( K \) from 1 to 18, running models with and without \textit{GmuD95}. We first plotted the mean and variance in likelihood per \( K \) using STRUCTURE HARVESTER v.0.6.92 (Earl and vonHoldt (2012)) and implemented the Evanno et al. (2005) method. We extracted and visualized the \( Q \) value tables from the results of STRUCTURE using Distruct v.1.1 (Rosenberg (2004)).

2.3.3.3 Historical population parameters

We next traced ancestral patterns of gene flow among our sampled populations. We used coalescent reconstructions with IM (Hey and Nielsen (2004)) to evaluate pairwise maximum likelihood estimates of ancestral gene flow and time since splitting between the five groups identified by STRUCTURE (see below). This method yielded an approximate timeline of historic genetic differentiation events among these groups based on a rate of 0.0005 mutations per locus per generation (Howes et al. (2009)), a generation time of 37 years (Congdon et al. (2003)), and a stepwise mutation model suitable for microsatellites (Kimura and Ohta (1978)). We imposed uninformative prior distributions on the upper-bound values of migration rates and effective population sizes, depending on their convergence in the results (see Hey and Nielsen (2004)). We performed 30-min runs of each parameter set to check for convergence and updated parameters until convergence was achieved. Having excluded \textit{GmuD95}, we performed five separate sets of runs of five pairs of clustered populations each, incrementally changing the priors depending on their convergence and the completion of the posterior distributions. These clustered populations derived from the 12 localities grouped into five populations as identified by STRUCTURE in Table S4, but with the population comprised of Carroll-IL, Will-IL, and Grant-NE (hereafter, we refer to populations with a county-state designation) split into Illinois and Nebraska clusters to determine the time since split between these two groups. We then used the estimates of migration rates and population-scaled mutation rates to calculate demographic parameters (with 95% confidence intervals), such as the effective population sizes, time since splitting in years, and migration rates per generation between pairs of the clustered populations.
(Hey and Nielsen (2004)). We also estimated relatively recent (roughly several generations) bidirectional migration rates between the same clusters (see Table S4) with BayesAss v.1.3 (Wilson and Rannala (2003)). This method uses an MCMC method applied to diploid data to determine recent migration rates and to assign ancestries to individuals. We performed multiple initializations of MCMCs and analyzed the trace files of logarithmic probabilities using Tracer v.1.5.0 (Rambaut and Drummond (2007)) to ensure good mixing and effective sampling from the posterior distribution. We constructed approximate 95% confidence intervals around mean recent migration rates as mean ±1.96× standard deviation. For each initialization, we utilized 10 million iterations of the MCMC, with a burn-in of 1 million iterations, sampling from every 1000 iterations.

2.4 Results

2.4.1 Microsatellite characteristics and deviations from HWE

The number of alleles per locus and size of the alleles for all loci were largely within the ranges reported by King and Julian (2004). The average number of alleles per locus (all eight loci were polymorphic) was 20.1 across all populations. The average observed heterozygosity among the 12 well-sampled populations across all loci was 0.54 (range 0.25−0.82; Table A.5) and 0.52 across all 18 populations, indicating high levels of polymorphism. We detected significant heterozygote deficiency on average across all loci in the Grant-NE, Bremer-IA, Muscatine-IA, Clinton-IA, Scott-MN, Carroll-IL, and Will-IL populations (Table 2.7; but see also Table A.6 for a per locus analysis). We detected significant (P < 0.05) heterozygote deficiency in individuals from multiple populations at various loci (Table A.6). However, after correction for multiple comparisons, only Grant-NE, Muscatine-IA, Bremer-IA, Carroll-IL, McHenry-IL, and Will-IL at GmuD95, Muscatine-IA at GmuD90, and Will-IL at GmuD93 remained out of HWE, primarily due to heterozygote deficiency (Table A.6). Microchecker revealed the possibility of null alleles at GmuD95, and hence many further analyses were performed both with and without this locus. Finally, the average frequency of private alleles in the sampled populations was 0.0777, and we detected no evidence of linkage disequilibrium between any pair of loci (all
P > 0.05), suggesting random assortment among the eight loci (Table A.7).

2.4.2 Population structure

Pairwise \( F_{st} \) values fell between 0.01 and 0.47 (Table 2), indicating low to moderate levels of genetic differentiation between the 12 well-sampled populations (55 of 66 comparisons were significant after Bonferroni correction; Table 2.2). The highest significant \( F_{st} \) was 0.469 between McHenry-IL and Grant-NE, which, not surprisingly, is the second most geographically-distant population-pair sampled (≈ 1102 km). Still, significant pairwise \( F_{st} \) values were detected even over short distances. For instance, the \( F_{st} \) of 0.287 (P < 0.05) between Clinton-IA and Carroll-IL fell in the upper half of our 66 comparisons, but are two of the geographically-closest populations sampled (≈ 15 km, though separated by the Mississippi River). Other comparisons exhibited low \( F_{st} \) values. Notably, Winnebago-IA showed non-significant \( F_{st} \) values with both Jones-IA and Clinton-IA (\( F_{st} = 0.045, P = 0.097; F_{st} = 0.048, P = 0.116, \) respectively) even though these populations are ≈ 253 km and ≈ 294 km linear distance apart. Pairwise comparisons between these three populations from eastern Iowa and two other eastern Iowa populations (Worth-IA and Bremer-IA) all yielded \( F_{st} \) values < 0.058 that are not significantly different from zero, indicating little genetic differentiation within the drainages of the Winnebago/Shell Rock/Cedar, Wapsipinicon, and Maquoketa Rivers. Also of note, the Grant-NE population, located in western Nebraska, had comparatively low (albeit, significantly different from zero) \( F_{st} \) values with Carroll-IL in western Illinois and Will-IL in eastern Illinois (\( F_{st} = 0.187, P < 0.0001; F_{st} = 0.135, P = 0.001, \) respectively), considering that our largest \( F_{st} \) value (0.469) was between Grant-NE and McHenry-IL, which is only 84 km north of Will-IL. Overall, estimates of Weir and Cockerham \( F \)-statistics involving the 12 well-sampled populations revealed signatures of inbreeding (\( F_{is} = 0.136 (0.027 - 0.275, 95\% \) CI) with \( GmuD95 \) and \( F_{is} = 0.075 (0.010 - 0.165, 95\% \) CI) without \( GmuD95 \)). Furthermore, global \( F_{st} \) values suggested considerable genetic differentiation, and ranged from 0.263 (0.184 – 0.357, 95% CI) with \( GmuD95 \) to 0.270 (0.178 – 0.379, 95% CI) without \( GmuD95 \) (see Table A.8). Geographic and genetic distances exhibited a positive correlation (\( R^2 = 0.179; \) Fig. 2.3). Thus, while geographic distance among populations likely contributes to genetic differentiation, other fac-
tors play a role as well. We thus performed AMOVAs to estimate the amount of variance in multilocus genotypes explained by each of three models (Model 1 – populations grouped by linear geographic distance, Model 2 – populations grouped by current watershed distributions, and Model 3 – populations grouped by current watershed distributions plus relative location inside or outside the Laurentide Ice Sheet – see Tables A.1, A.2 and A.3; Fig. 2.1). AMOVA revealed that most of the genetic variation occurred within populations in all three models (68.8 – 70.8%), with much smaller amounts occurring among populations (24.1 – 26.3%) or among clusters of populations (4.8 – 5.1%) as identified above (Table A.9). A smaller number of groups (4, as hypothesized in Model 2 – see Table A.2) better explained the genetic data than did a larger number of groups (6, as hypothesized in Model 1 – Table A.1) (∆AICc > 42).

A comparison of the model of clustering by watersheds alone (i.e., 4 groups – see Table A.2) to one clustering by watersheds and the Laurentide Ice Sheet (5 groups, as hypothesized in Model 3 – see Table A.3) yielded a ∆AICc = 2.22, indicating that both models have substantial support (see Table A.9). Population structure was estimated for all 209 individuals from all 18 populations without using any prior geographic information in STRUCTURE. Across all eight loci, the most likely population structure was $K = 4$, but excluding GmuD95 from the STRUCTURE analyses yielded $K = 5$ regions (Figs. 2.2 and 2.4, Fig. A.1).

### 2.4.3 Historical population parameters

We detected considerable diversity in $N_e$ among five clusters of populations (the four from STRUCTURE (with adequate sample sizes) split into five clusters to resolve divergence time between Grant-NE and Carroll-IL, Will-IL) (Table S10.1). Median $N_e$ estimates ranged from 750 (95% CI = 466–1326; Grant-NE vs. Carroll-IL, Will-IL) to 1,681,177 (95% CI = 1,682,883–1,684,589; IA populations vs. Carroll-IL, Will-IL). Although Ne in this latter case and for Scott-MN, Muscatine-IA (Group 2, Table A.3) vs. Grant-NE derived from analyses that failed to converge, all other estimates came from analyses that converged within five runs. The oldest ‘split’ events were estimated to have occurred well into the Pleistocene, while the youngest probably transpired in the recent past (Table A.10.2). In the former case, the IA populations (Group 1, Table A.3) and McHenry-IL (Group 4, Table A.3) apparently split $\approx 353,250$ YBP
36

(95% CI = 185, 250–410, 750 YBP), with strong subsequent unidirectional gene flow from east to west (median $m_2 = 10.65$ individuals/generation, 95% CI = 6.75–23.35; median $m_1 = 1.05$, 95% CI = 0.85–2.65). Also, the IA populations and the Grant-NE population were estimated to have split $\approx 231,750$ YBP (95% CI = 187, 250–495, 250 YBP), with little subsequent gene flow between the two localities (median $m_1 = 0.15$, 95% CI = 0.15–1.35; median $m_2 = 2.95$, 95% CI = 1.95–6.85). At the other extreme, the most recent split occurred around a median of 850 YBP (95% CI around mean = 950–1150 YBP) between Carroll-IL, Will-IL and the IA populations, with strong bi-directional gene flow of 13.45 individuals/generation (95% CI = 11.75–16.95) from IA into Carroll-IL, Will-IL and 22.65 individuals/generation (95% CI = 17.05–26.25) from Carroll-IL, Will-IL into IA. These analyses also accorded with a puzzling result obtained from the Fst and STRUCTURE analyses. That is, despite considerable geographic distance, splits between Grant-NE and Carroll-IL, Will-IL, and Grant-NE and McHenry-IL, were estimated to have occurred relatively recently ($\approx 22,550$ YBP, 95% CI = 16, 550–90, 650 YBP and $\approx 22,250$ YBP, 95% CI = 20, 550–95, 450 YBP, respectively). Subsequent gene flow from Grant-NE to Carroll-IL, Will-IL also appears to be non-trivial (median $m_2 = 8.25$, 95% CI = 3.65–21.35).

In contrast to these analyses of long-term gene flow, BayesAss detected no substantial recent migration between any of the four clusters defined by STRUCTURE (Table A.11, Group 5 was excluded due to small sample size, as with IM analyses above). The highest unidirectional migration rate was estimated for Muscatine-IA, Scott-MN into Grant-NE, Carroll-IL, Will-IL at only 0.03 individuals/generation. By comparison, much higher migration rates were detected by BayesAss relative to IM, but were restricted within their respective STRUCTURE groups ($m > 0.94$, 95% CI = 0.896–0.993).

2.5 Discussion

Overall, our genetic results accord with a classic biogeographic scenario (sensu Smith (1957); see below) that populations of Blanding’s turtle (Emys blandingii) across the midwestern United States (i.e., west of Lake Michigan) are significantly differentiated from each other. We identify 4 – 5 unique genetic groups of Blanding’s turtles in this region, which do not necessarily conform to their present geography. Indeed, although separated by $> 1000$ km, western
Nebraska and eastern Illinois populations exhibit unexpectedly close population genetic structure. Regardless, our results also indicate strong support for the post-Pleistocene distribution of these turtles along watersheds in the Mississippi-Missouri River basins and along aquatic corridors established after the Last Glacial Maximum, with limited gene flow more recently (Fig. 2.1).

Post-Pleistocene distribution of herpetofauna, including Blanding’s turtle, in the Great Lakes Region is thought to have occurred in two main phases. Species first migrated south and west during glacial advances, then colonized northern and eastern regions (and re-adjusted ranges in the south and west) during recession of the Laurentide Ice Sheet, with subsequent declines in population size and connectivity in the new locales (Smith (1957)). Our molecular analyses of *E. blandingii* populations in the midwestern United States are consistent with this scenario, which invokes classic conditions of bottlenecks, reduced population sizes, and limited gene flow that promote among-population differentiation via random genetic drift at neutral genetic loci. Other molecular studies of post-glacial colonization and phylogeography of amphibians, snakes, and other turtle taxa in this general region comport with our findings (e.g., Austin et al. (2002), Fontanella et al. (2008), Janzen et al. (2002), Lee-Yaw et al. (2008), Placyk et al. (2007), Starkey et al. (2003), Weisrock and Janzen (2000)). Unlike in our system, however, most nuclear and mitochondrial phylogeographic studies of herpetofauna in the Great Plains report little genetic variation among populations, particularly among more northerly populations. These authors typically attribute this result to combined effects of population bottlenecks during glacial displacement and subsequent rapid northward colonization (e.g., Amato et al. (2008)), but this pattern of genetic depauperation also may derive from slower rates of molecular evolution in those loci compared to the hypervariable microsatellite loci used in our study.

In general, Blanding’s turtle populations sampled in our study exhibit low to moderate genetic differentiation and considerable molecular phylogeographic structure. Linear distance among localities explains some of the genetic differentiation, but the signals of watershed and last glacial maximum distribution are also detectable and significantly explain a larger fraction of the genetic data. These geographic groupings of populations (Missouri River watershed,
Minnesota River watershed, Mississippi River watershed inside the Des Moines Lobe of the Laurentide Ice Sheet, Mississippi River watershed outside the Des Moines Lobe of the Laurentide Ice Sheet, and Southern Lake Michigan Crescent watershed) notably, although not completely, correspond with the 4–5 unique genetic groups independently identified by the STRUCTURE analyses (Figs. 2.1 and 2.4, Fig. A.5).

These phylogeographic results further accord, in general, with the spatiotemporal dynamics of Pleistocene glacial advances and retreats in the midwestern United States. The peak of the Illinoian glacial period occurred \( \approx 300,000–150,000 \) YBP (Mickelson and Colgan (2003)), during which glaciers extended into Kansas, Missouri, and southern Illinois (Ehlers and Gibbard (2008)). Our genetic data, through multiple phylogeographic and migration analyses, suggest that it is around this time that the NE, IA, and IL groups began to differentiate, potentially from an ancestral source population in the south-central Great Plains, which would accord with fossil evidence (summarized in Ernst and Lovich (2009)). Subsequent glacial advances and retreats included those involving the Laurentide Ice Sheet during the Wisconsinan glacial period (\( \approx 100,000–4000 \) years ago) (Ehlers and Gibbard (2008)), which reached as far south as south-central Iowa (i.e., the Des Moines Lobe). These non-uniform advance-retreat dynamics by glaciers presumably further created isolated aquatic corridors for northward and eastward colonization of regions by *E. blandingii* and other water-linked herpetofauna during our current Holocene interglacial period (e.g., Amato et al. (2008), Austin et al. (2002), Lee-Yaw et al. (2008), Starkey et al. (2003); see Fig. 2.1), which are reflected in the phylogeographic and gene flow relationships among population clusters.

Our phylogeographic findings are intriguing in light of other studies that had previously detected relatively little genetic differentiation among *E. blandingii* populations west of Lake Michigan (Mockford et al. (2007), Rubin et al. (2001), Spinks and Shaffer (2005)). These three studies employed various molecular (nuclear RAPDs and microsatellites, nDNA and mtDNA sequences) and analytical (ANOVA, Fst, STRUCTURE, phylogenetic, etc.) approaches. Within this region, Mockford et al. (2007) used five nuclear microsatellite loci to examine one IL, one WI, and three MN populations, finding low \( F_{st} \) values (< 0.10 in all 10 pairwise comparisons) yet significant differentiation between IL, WI, and two of the MN populations. Rubin et al.
(2001) studied the same IL and WI populations (and two others outside the region) using nuclear RAPD markers and detected negligible differentiation among populations. Finally, in the course of a larger study with an interspecific phylogenetic focus, Spinks and Shaffer (2005) included four midwestern populations (WI, two from MN, and NE) from which one individual each was sequenced at one mtDNA locus and three nDNA loci, and noted that "intraspecific branch lengths were relatively short..." In our case, we targeted more populations (18 vs. 5, 3, and 4, respectively) over a larger fraction of this geographic region. We also chose microsatellite loci that were likely to be hypervariable (average number of alleles per locus across all 18 midwestern populations was ≈ 20 vs. < 10 across 5 midwestern populations, < 3 across 3 midwestern populations, and < 4 across 9 populations throughout the range, respectively). Both of these methodological considerations may have enhanced our capability to detect genetic structure in this geographic region of the United States compared to the three previous studies.

As discussed above, while our findings are generally congruent with those obtained in studies of other aquatic herpetofauna, spatiotemporal dynamics of glacial advances and retreats, and multiple analytical approaches, we nonetheless obtained some unexpected results. Of particular note is the apparent genetic similarity of microsatellites between western Nebraska and eastern Illinois populations despite a linear distance between these localities of > 1000 km. Had we not sampled intervening Minnesota and Iowa populations of E. blandingii, we might have inferred incorrectly that this species possesses negligible genetic variation and structure in the midwestern United States. As it is, we are left without a geologically plausible explanation for this puzzling similarity. Possible hypotheses include (1) humans transported Blanding’s turtles between the two sites (e.g., Mormons during their 1846 forced exodus from Illinois to Utah), (2) Blanding’s turtles at both sites retain similar ancestral genetic polymorphisms (i.e., incomplete lineage sorting), and (3) alleles at the microsatellite loci for Blanding’s turtles at these two sites have converged independently (i.e., homoplasy). In the first case, we cannot rule out the possibility of translocations by humans, but this would seem an improbable explanation for our findings considering our abundant sampling in these regions. Consequently, we attempted to resolve this conundrum by analyzing variation in DNA sequences of flanking regions of the most
homozygous microsatellite locus \((GmuD95)\) and two of the most variable microsatellite loci \((GmuD121\) and \(GmuD21)\). Results (see Figs. A.2, A.3, A.4, Tables A.12 and A.13) from this analysis suggest two patterns of molecular evolution at these three loci – (a) possible allele size homoplasy and flanking region SNP variation at \(GmuD95\), and (b) no repeat size variation and little SNP variation in flanking regions at \(GmuD121\) and \(GmuD21\), indicative of microsatellite saturation at these two loci (longest recorded allele size was 154 bases for \(GmuD121\) and 152 bases for \(GmuD21\) in these populations). Regardless, analysis without \(GmuD121\), \(GmuD21\), and \(GmuD95\) still shows low genetic differentiation \((F_{st} < 0.184, P < 0.0001)\) between Grant-NE and eastern Illinois, and higher genetic differentiation between McHenry-IL and Will-IL \((F_{st} = 0.325, P < 0.0001)\) and between Grant-NE and Iowa populations \((F_{st} > 0.23, P < 0.0001)\).

Regardless of the explanation for this unusual pattern, our extensive genetic study of Blanding’s turtle in the midwestern United States has significant conservation and management implications for this imperiled species. We identified a considerable pool of genetic variation across populations and substantial geographic structuring of this genetic variation with relatively little recent gene flow, possibly because of colossal loss of hospitable environments between essential terrestrial and aquatic habitats and between population localities (e.g., Beaudry et al. (2008)). This current situation could be catastrophic for \(E.\ blandingii\) and other taxa with movement-heavy, biphasic natural histories. For example, Blanding’s turtle has not reproduced in any known western Iowa populations for at least 20 years and those populations are now limited to a few very old (possibly 50 – 100 years old) turtles (Christiansen (1998)). Moreover, evidence consistent with excessive inbreeding (e.g., % inviable eggs) has been detected in other Iowa localities (53%; JLC, unpublished) and in the McHenry-IL population (48%; SH, unpublished), in contrast to Grant-NE where the population size remains large (21%; FJJ et al., unpublished). Beyond the need for detailed demographic studies, our genetic evaluation of midwestern Blanding’s turtles makes clear that any management action, such as assisted translocation of \(E.\ blandingii\) between localities, would benefit from being conducted with an eye toward accounting for the genetically structured groups that we detected (reviewed in Alacs et al. (2007)). Although we unfortunately have no evidence of local phenotypic adapta-
tion in *E. blandingii*, as noted above most molecular phylogeographic studies of herpetofauna in this region have reported little genetic variation. Consequently, midwestern Blanding’s turtles exemplify a relatively unique outcome and should accordingly evince a vigilant management approach to ensure retention of genetic diversity. Still, intensification of changes to the regional landscape is further restricting natural gene flow and population size for *E. blandingii*, thus balancing genetic and demographic concerns, among other issues, will require challenging management decisions.

2.6 Acknowledgements

Thanks to Steve Dinkelacker, Robb Goldsberry, Chris Phillips, and Sara Ruane for assistance with collecting and to two anonymous reviewers for constructive criticisms. Turtles and tissues for this project were obtained under Iowa DNR Scientific Collecting permits to JLC and TJV, Illinois Endangered Species permits to FJJ and SH, Minnesota DNR permits to JMR, and a Nebraska Game and Parks Commission permit to Steve Dinkelacker. Funding for this research was provided in part by a contract from the McHenry County Conservation Foundation and National Science Foundation grants IBN-0080194, DEB-0089680, and DEB-0640932 to FJJ, and by a National Aeronautics and Space Administration Iowa Space Grant Consortium grant to FJJ and JLC. Additional funding for sequencing and bioinformatics support was provided by National Science Foundation DEB-0508665 to FJJ and EMM, as well as a grant from the EEOB Department at ISU and a James Cornette Fellowship to AS.
2.7 Tables and Figures

<table>
<thead>
<tr>
<th>Population</th>
<th>Latitude</th>
<th>Longitude</th>
<th>N</th>
<th>n</th>
<th>P</th>
<th>A</th>
<th>$A_p$</th>
<th>$H_e$</th>
<th>$H_o$</th>
<th>P val</th>
</tr>
</thead>
<tbody>
<tr>
<td>Grant-NE</td>
<td>42.00</td>
<td>-101.73</td>
<td>24</td>
<td>23.12</td>
<td>1.00</td>
<td>4.25</td>
<td>4.25</td>
<td>0.50</td>
<td>0.38</td>
<td>&lt;0.0001</td>
</tr>
<tr>
<td>Winnebago-IA</td>
<td>43.46</td>
<td>-93.56</td>
<td>7</td>
<td>6.88</td>
<td>1.00</td>
<td>4.75</td>
<td>4.75</td>
<td>0.63</td>
<td>0.61</td>
<td>0.4482</td>
</tr>
<tr>
<td>Wright-IA</td>
<td>42.63</td>
<td>-93.65</td>
<td>9</td>
<td>9.00</td>
<td>0.88</td>
<td>3.50</td>
<td>3.86</td>
<td>0.50</td>
<td>0.54</td>
<td>0.6649</td>
</tr>
<tr>
<td>Scott-MN</td>
<td>44.70</td>
<td>-93.34</td>
<td>9</td>
<td>8.50</td>
<td>1.00</td>
<td>4.50</td>
<td>4.50</td>
<td>0.70</td>
<td>0.54</td>
<td>0.0042</td>
</tr>
<tr>
<td>Bremer-IA</td>
<td>42.84</td>
<td>-92.24</td>
<td>13</td>
<td>13.00</td>
<td>1.00</td>
<td>7.12</td>
<td>7.12</td>
<td>0.71</td>
<td>0.68</td>
<td>0.0677</td>
</tr>
<tr>
<td>Jones-IA</td>
<td>41.99</td>
<td>-91.20</td>
<td>6</td>
<td>5.38</td>
<td>1.00</td>
<td>3.62</td>
<td>3.62</td>
<td>0.58</td>
<td>0.53</td>
<td>0.1</td>
</tr>
<tr>
<td>Muscatine-IA</td>
<td>41.56</td>
<td>-91.14</td>
<td>14</td>
<td>14.00</td>
<td>1.00</td>
<td>8.00</td>
<td>8.00</td>
<td>0.79</td>
<td>0.59</td>
<td>&lt;0.0001</td>
</tr>
<tr>
<td>Worth-IA</td>
<td>43.41</td>
<td>-93.44</td>
<td>9</td>
<td>8.75</td>
<td>1.00</td>
<td>4.25</td>
<td>4.25</td>
<td>0.60</td>
<td>0.59</td>
<td>0.164</td>
</tr>
<tr>
<td>Clinton-IA</td>
<td>41.78</td>
<td>-90.80</td>
<td>7</td>
<td>7.00</td>
<td>1.00</td>
<td>4.00</td>
<td>4.00</td>
<td>0.59</td>
<td>0.50</td>
<td>0.028</td>
</tr>
<tr>
<td>Carroll-IL</td>
<td>41.95</td>
<td>-90.12</td>
<td>22</td>
<td>21.25</td>
<td>1.00</td>
<td>5.62</td>
<td>5.62</td>
<td>0.64</td>
<td>0.52</td>
<td>&lt;0.0001</td>
</tr>
<tr>
<td>McHenry-IL</td>
<td>42.33</td>
<td>-88.22</td>
<td>61</td>
<td>60.62</td>
<td>1.00</td>
<td>6.00</td>
<td>6.00</td>
<td>0.56</td>
<td>0.52</td>
<td>&lt;0.0001</td>
</tr>
<tr>
<td>Will-IL</td>
<td>41.58</td>
<td>-88.07</td>
<td>21</td>
<td>20.12</td>
<td>1.00</td>
<td>5.00</td>
<td>5.00</td>
<td>0.58</td>
<td>0.43</td>
<td>&lt;0.0001</td>
</tr>
<tr>
<td>Mean or Total</td>
<td>202</td>
<td>16.47</td>
<td>0.99</td>
<td>5.05</td>
<td>5.08</td>
<td>0.61</td>
<td>0.54</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Table 2.1 Summary statistics across the 12 populations with sample size > 5. ‘N’ is the total number of individuals sampled for the population, ‘n’ is the mean sample genotyped over all the loci, ‘P’ is the total number of polymorphic loci in that population, ‘A’ is the average number of alleles per locus, ‘$A_p$’ is the mean number of alleles per polymorphic locus, ‘$H_e$’ is the expected heterozygosity, and ‘$H_o$’ is the observed heterozygosity. The statistics were estimated using GDA v.1.1. ‘P-val’ indicates P-values from a global test of heterozygote deficiency, performed in Genepop v. 4.1, under the null hypothesis that the populations are at Hardy-Weinberg Equilibrium, with the alternate hypothesis that the populations have significant heterozygote deficiency.
Table 2.2: Lower triangle contains pair-wise Fst values between populations; the significant ones before sequential Bonferroni correction are shown in italics, non-significant Fst values regardless of Bonferroni correction are shown in grey, and the significant Fst values post correction are shown in boldface. Fifty-five of 66 pairwise comparisons remained significant after sequential Bonferroni correction. The upper triangle contains linear geographic distance in kilometers between population pairs.

<table>
<thead>
<tr>
<th></th>
<th>Bremer-IA</th>
<th>Carroll-IL</th>
<th>Clinton-IA</th>
<th>Grant-NE</th>
<th>Jones-IA</th>
<th>McHenry-IL</th>
<th>Muscatine-IA</th>
<th>Will-IL</th>
<th>Winnebago-IA</th>
<th>Worth-IA</th>
<th>Wright-IA</th>
<th>Scott-MN</th>
</tr>
</thead>
<tbody>
<tr>
<td>Bremer-IA</td>
<td>0</td>
<td>200.45</td>
<td>167.05</td>
<td>784.92</td>
<td>126.44</td>
<td>334.26</td>
<td>168.84</td>
<td>370.55</td>
<td>128.06</td>
<td>116.79</td>
<td>117.62</td>
<td>225.64</td>
</tr>
<tr>
<td>Carroll-IL</td>
<td>0.2085</td>
<td>0</td>
<td>59.39</td>
<td>900.21</td>
<td>90.22</td>
<td>162.44</td>
<td>95.41</td>
<td>174.4</td>
<td>328.24</td>
<td>316.97</td>
<td>300.49</td>
<td>402.25</td>
</tr>
<tr>
<td>Clinton-IA</td>
<td>0.0575</td>
<td>0.2867</td>
<td>0</td>
<td>905.59</td>
<td>41.43</td>
<td>221.79</td>
<td>37.78</td>
<td>227.38</td>
<td>293.85</td>
<td>282.69</td>
<td>253.39</td>
<td>384.88</td>
</tr>
<tr>
<td>Grant-NE</td>
<td>0.2666</td>
<td>0.1870</td>
<td>0.3288</td>
<td>0</td>
<td>870</td>
<td>1114.15</td>
<td>879.9</td>
<td>1132.96</td>
<td>687.13</td>
<td>695.66</td>
<td>668.43</td>
<td>742.4</td>
</tr>
<tr>
<td>Jones-IA</td>
<td>0.0070</td>
<td>0.2705</td>
<td>0.0105</td>
<td>0.3752</td>
<td>0</td>
<td>249.33</td>
<td>49.04</td>
<td>263.86</td>
<td>252.62</td>
<td>241.5</td>
<td>213.13</td>
<td>347.19</td>
</tr>
<tr>
<td>McHenry-IL</td>
<td>0.3008</td>
<td>0.3806</td>
<td>0.3589</td>
<td>0.4687</td>
<td>0.3158</td>
<td>0</td>
<td>256.74</td>
<td>84.17</td>
<td>453.39</td>
<td>442.56</td>
<td>447.01</td>
<td>490.03</td>
</tr>
<tr>
<td>Muscatine-IA</td>
<td>0.1647</td>
<td>0.1532</td>
<td>0.2610</td>
<td>0.2798</td>
<td>0.2144</td>
<td>0.2635</td>
<td>0</td>
<td>255.21</td>
<td>290.74</td>
<td>279.91</td>
<td>239.16</td>
<td>392.93</td>
</tr>
<tr>
<td>Will-IL</td>
<td>0.1750</td>
<td>0.0511</td>
<td>0.2327</td>
<td>0.1345</td>
<td>0.2412</td>
<td>0.3714</td>
<td>0.1714</td>
<td>0</td>
<td>496.22</td>
<td>485.04</td>
<td>474.75</td>
<td>550.38</td>
</tr>
<tr>
<td>Winnebago-IA</td>
<td>0.0211</td>
<td>0.2605</td>
<td>0.0475</td>
<td>0.2775</td>
<td>0.0448</td>
<td>0.2959</td>
<td>0.1618</td>
<td>0.1608</td>
<td>0</td>
<td>11.28</td>
<td>93.4</td>
<td>138.96</td>
</tr>
<tr>
<td>Worth-IA</td>
<td>0.0250</td>
<td>0.2642</td>
<td>0.0202</td>
<td>0.3210</td>
<td>0.0244</td>
<td>0.3324</td>
<td>0.2357</td>
<td>0.2172</td>
<td>0.0433</td>
<td>0</td>
<td>89.07</td>
<td>143.73</td>
</tr>
<tr>
<td>Wright-IA</td>
<td>0.0923</td>
<td>0.3188</td>
<td>0.0983</td>
<td>0.3811</td>
<td>0.1158</td>
<td>0.3709</td>
<td>0.2971</td>
<td>0.2787</td>
<td>0.1305</td>
<td>0.0817</td>
<td>0</td>
<td>232.29</td>
</tr>
<tr>
<td>Scott-MN</td>
<td>0.2163</td>
<td>0.2752</td>
<td>0.2951</td>
<td>0.3506</td>
<td>0.2787</td>
<td>0.3111</td>
<td>0.0847</td>
<td>0.2648</td>
<td>0.2172</td>
<td>0.2866</td>
<td>0.351</td>
<td>0</td>
</tr>
</tbody>
</table>
Figure 2.1 Localities and putative groupings of the 18 sites sampled for *Emys blandingii* in the midwestern United States. (top left) Gray areas are water bodies, the more southern dashed line indicates the extent of the pre-Wisconsinan glacial limit (Kansan/Nebraskan glaciations), and the more northern dashed line indicates the extent of the Laurentide Ice Sheet in the more recent Wisconsinan glaciation. (top right) Representation of Model 1 (Table S1), with sites within 100km linear distance from each other clustered to form groups. (middle left) Representation of Model 2 (Table S2), with sites in the same watershed clustered to form groups. (middle right) Representation of Model 3 (Table S3), with sites in the same watershed and location with respect to the limit of the Laurentide Ice Sheet clustered to form groups.
Figure 2.2  Representation of sample sizes (indicated by diameter of the pies) and of relative admixture distribution (indicated by shade of gray) inferred by STRUCTURE at K=5
Figure 2.3  Plot of linear geographic distance between sampled populations of Emys blandingii vs. Slatkin’s linearized Fst genetic distance between these same sampled populations to estimate the presence of isolation by distance. This plot was derived from a Mantel test performed in GenAlex v.6.2 with 1000 permutations.
Figure 2.4  Estimates of admixture proportions in sampled populations of *Emys* *blandingii*. Twenty runs of STRUCTURE were performed for each value of K, under the admixture model, and the Dirichlet parameter ‘alpha’ was inferred from the data. Each run was performed using the 209 genotyped individuals from all 18 localities with a burn-in period of 10000 and 10000 MCMC reps. Left) All loci. Right) all loci except *GmuD95*
CHAPTER 3. MULTICLUST - FAST MULTINOMIAL CLUSTERING
OF MULTILOCUS GENOTYPES TO INFER GENETIC POPULATION
STRUCTURE

3.1 Abstract

Identifying population structure from multilocus genotype data is key to downstream population genetic analyses in a variety of fields, including conservation, evolutionary genetics, Genome Wide Association Studies (GWAS), and pedigree reconstruction for quantitative genetics. Several methods have been put forth to estimate population structure, but issues with consistency, efficiency, and reliability of estimation remain. One of the most popular methods has been the Bayesian approach of STRUCTURE (Pritchard et al. (2000b)). Originally, the same model was estimated by maximum likelihood approaches, and recently the benefits of a ML approach are being re-examined in the context of reliability, consistency, and efficiency issues with Bayesian methods. Here we extend the method of Alexander et al. (2009) to handle more than bi-allelic loci. Comparative analyses with both simulated and empirical data with MULTICLUST and STRUCTURE (Pritchard et al. (2000b)) indicate consistency of inference, and reliability of inference (addressed using information criteria).

3.2 Introduction

The concept of a ‘population’ plays a key role in the field of ecological and evolutionary population genetics. Recently, Waples and Gaggiotti (2006) defined the idea of a ‘population’ under ecological and evolutionary paradigms. A common feature of both definitions is that within a population, members of the same species can randomly mate with each other. Random mating has two main implications. Biologically, it helps to maintain genetic diversity in
a population, and statistically it means allele counts follow a simple multinomial distribution. When a population is subdivided into subpopulations, within which Waples and Gaggiotti (2006)’s definitions apply, but between which mating is restricted, the population is said to have structure. Many generations of reproduction in a populations with structure leads to genetic patterns or genetic structure. Identifying intra-species genetic structure has played an important role in the fields of conservation, phylogeography, and genetics of contagious diseases, among others ((Avise, 2009),(Cruaud et al., 2011),(Dutech et al., 2005),(Hohenlohe et al., 2010),(Mockford et al., 2007),(Mockford et al., 1999),(Nason et al., 2002),(Pemberton et al., 2008),(Reich et al., 2009),(Rosenberg et al., 2002),(Rubin et al., 2001),(Starkey et al., 2003),(Templeton, 2002), etc). One important objective is to estimate population structure from genetic structure reflected in multilocus genotype data. Specifically, we might be interested in estimating the number of, size, and membership of subpopulations within which random mating holds. A very important, and difficult question, is then - how do we identify these genetic subpopulations from multilocus genotype data? Several methods have been proposed to address this question, largely classified into distance-based methods, such as K-means (Hartigan and Wong (1979)) or haplotype network analysis (Templeton (1998)), and PCA/PCoA [Principal Coordinates/Components Analyses](Jombart et al. (2010),Li and Yu (2008)), and model-based methods (STRUCTURE((Pritchard et al., 2000b), BAPS((Corander et al., 2003),(?)), FASTSTRUCT((Chen et al., 2006)), ADMIXTURE((Alexander et al., 2009)), PSMIX((Wu et al., 2006)), FRAPPE((Tang et al., 2005)), STRUCTURAMA((Huelsenbeck and Andolfatto, 2007)), etc. Both distance and model-based methods work with the same information - multilocus genotypes (microsatellite, AFLP’s, SNP’s, RFLP’s, allozymes, etc), obtained from randomly sampled individuals of the same species in a population. Both types of methods attempt to group individuals with similar genotypes, but distance-based methods reduce the data to pairwise distances, while model-based methods use a generative model (to randomly sample observations from). Several distance-based methods have been summarized by Lawson and Falush (2012), and we shall not deal with them in this manuscript. The focus of this manuscript is on the utility of model-based clustering methods, and their pros and cons. Several maximum likelihood based ((Chen et al., 2006),(Tang et al., 2005),(Smouse
et al., 1990), (Wu et al., 2006)), and Bayesian ((Pritchard et al., 2000b), (?), (Huelsenbeck and Andolfatto, 2007)) frameworks have been developed for model-based clustering of multilocus genotype data into subpopulations.

In recent years, Bayesian approaches, particularly STRUCTURE ((Pritchard et al., 2000b)), have been more popular among molecular ecologists at large, owing to its ease of use, intuitive visualizations of genetic admixture, and robustness to large genomic datasets (Gilbert et al. (2012)). Despite these advantages, STRUCTURE ((Pritchard et al., 2000b)) requires the specification of priors, and has been noted to have issues with converging to the most likely subpopulation structure, despite other intuitive procedures (Evanno et al., 2005) to infer this. Results so obtained are often ambiguous, especially since there is no statistical ‘test’ or method that evaluates the validity of the models (Gilbert et al. (2012)). Oftentimes, multiple runs of multiple initializations are required to obtain approximate convergence. More recently, several issues have been raised with respect to (1) reliability, (2) consistency, and (3) efficiency of inferring subpopulation structure, especially in the context of ecological and evolutionary questions (ref). Issues with convergence and inference using Bayesian clustering methods were addressed by Latch et al. (2006), and Francois and Durand (2010).

Likelihood frameworks on the other hand, while statistically robust, suffer a disadvantage in efficiency (see Alexander et al. (2009)). This is due to relatively slow convergence of likelihood estimation algorithms. Several algorithms, primarily under the framework of Expectation-Maximization (EM - Dempster et al. (1977)) have been implemented in the methods listed above, but share the same issues of efficiency. A fast alternative, which utilizes quadratic approximations to explore the likelihood surface was developed by Alexander et al. (2009), for bi-allelic (eg. SNP) genetic datasets, which report considerable efficiency in computing the ‘true’ subpopulation structure, compared to STRUCTURE((Pritchard et al., 2000b)).

Our goal was to address all these issues by extending the maximum-likelihood based algorithm of Tang et al. (2005) to perform quadratic approximations (sensu Alexander et al. (2009), Zhou et al. (2011)) for reliably and consistently estimating the most likely number of subpopulations, and do this efficiently with little computation and space requirements. The algorithm statistically infers the most likely number of subpopulations, admixture proportions
for each individual (or subpopulation), and subpopulation allele frequencies. Since MULTICLUST is likelihood-based, estimates obtained from multiple models are readily amenable in a hypothesis testing or model selection framework (AIC, BIC, bootstrap, etc). In this way, MULTICLUST provides a robust and reliable method for making informed inferences about the model that best fits the sampled genotypic data. MULTICLUST is also consistent in converging at the global maximum likelihood, after a certain number of initializations, and all convergences are achieved in a fraction of the time required by STRUCTURE ((Pritchard et al., 2000b). MULTICLUST also offers methods to handle missing or erroneous genotypic data (sensu Wu et al. (2006)).

This paper is organized as follows. In Section 2, we formulate mixture and admixture models in detail. Most of the detailed derivations for the Expectation-Maximization (EM) and Block Relaxation Quasi Newton (BR) algorithms with missing data are handled in the Appendix. In Section 3, we compare the relative performance of MULTICLUST and STRUCTURE((Pritchard et al., 2000b)) on simulated data, designed to exhibit different levels of subpopulation structure over evolutionary time. We then compare the consistency of MULTICLUST with STRUCTURE, across multiple identical iterations, by assessing four datasets reported as being inconsistently inferred (Gilbert et al. (2012)). We also demonstrate an analysis using MULTICLUST for another empirical dataset, involving conservation genetics of the endangered Blanding’s Turtle (Emys blandingii - Sethuraman et al. (2011)). All the algorithms were implemented in C((Kernighan and Ritchie, 1988)) for efficiency, and scripts for data simulation and result visualization were developed in R((R Development Core Team, 2010)).

3.3 Methods

3.3.1 Model

Both the mixture and the admixture models contain two fundamental assumptions: Hardy-Weinberg equilibrium (HWE) and linkage equilibrium (LE). HWE results when an infinite, panmictic population mates randomly without mutation, selection, migration, genetic drift, or meiotic drive. Under these conditions, each genotype is a random combination of alleles, and
allele frequencies are unchanging across generations (Stern, 1943). If we let integers represent alleles, then diploid genotype \{33, 2\}, consisting of allele 33 and allele 2, has likelihood \(2p_{33}p_2\), where \(p_m\) is the population allele frequency of allele \(m\) in all generations. LE posits that alleles present at different loci are also independent draws from the population. Thus, diploid, dual-locus genotype \{33, 2\}; \{4, 17\} has probability \(p_{1,33}p_{1,2}p_{2,4}p_{2,17}\), for allele frequencies now distinguishing loci 1 and 2.

If we further assume codominance of all alleles at each locus, then the 'observed' data are comprised of all observed alleles. Suppose we sample \(I\) individuals from a population with \(K\) possible subpopulations. We observe the genotype of each individual at \(L\) loci, which consists of \(M\) (the ploidy level), individual alleles. Let \(A_l\) be the number of distinct alleles at locus \(l\). For DNA or SNP data, \(A_l = 4\), and is constant across loci. The \(m\)th allele observed in individual \(i\) at locus \(l\) is denoted by \(x_{ilm}\), and all data are subsumed into the \(I \times L \times M\) matrix, \(X\). Because the \(M\) alleles observed at locus \(l\) are random replicates of the same process, we might restructure the data into the sufficient statistics, an \(I \times L \times \{A_1, \ldots, A_L\}\) jagged array \(N\), where \(n_{ila}\) is the number of times allele \(a\) is observed in individual \(i\) at locus \(l\). We caution that it is only computationally advantageous to work with the data in this form when ploidy \(M\) is routinely larger than the number of distinct alleles \(A_l\). We also assume that alleles at a locus are modeled by a multinomial distribution, if data were sampled from a homogeneous population; otherwise a 'mixture' or 'admixture' structure should be imposed (see below).

Subpopulations within the sampled population are distinguishable to the extent that they have different ancestral allele frequencies. Let \(p_{kla}\) be the frequency of allele \(a\) at locus \(l\) in ancestral subpopulation \(k\). Clearly, \(\sum_a A_l p_{kla} = 1\) for all \(k\) and \(l\), and \(0 \leq p_{kla} \leq 1\) for all \(a\). Evidence that \(p_{kla}\) varies with \(k\) is the key to detecting population substructure. The problem is then reduced to inferring \(p_{kla}\) when we do not know if individual \(i\) came from ancestral subpopulation \(k\), a problem solved by the Expectation Maximization (EM) algorithm (Dempster et al., 1977),(McLachlan and Krishnan, 1996). Since the EM algorithm is a prescription more than an algorithm (Lange, 1999), we will derive the exact equations for the EM algorithm in the appendix, where we also consider the possibility that some alleles are not observed, or “missing.” Considering the computational requirements of the EM algorithm, we have also
implemented a quadratic programming method (Block Relaxation - see Alexander et al 2009) to boost performance of convergence on the maximum data likelihood. Details of this algorithm are also derived in the appendix.

3.3.1.1 Mixture Model

In this model, each individual is sampled independently from one of $K$ ancestral subpopulations. In other words, this model assumes that there has been no interbreeding between the subpopulations (case of complete independence between subpopulations - see Waples and Gaggiotti(Waples and Gaggiotti, 2006)). We define $\eta = (\eta_1, \ldots, \eta_K)$, with $\eta_k > 0$ for all $k$ and $\sum_{k=1}^{K} \eta_k = 1$ to be the mixing proportions for $K$ ancestral subpopulations. For parameter vector $\Theta = (\eta, p)$, the observed data likelihood is

$$L(\Theta | X) = \prod_{i=1}^{I} \prod_{k=1}^{K} \eta_{ik} \prod_{l=1}^{L} \prod_{m=1}^{M} p_{klx_{ilm}} = \prod_{i=1}^{I} \prod_{k=1}^{K} \eta_{ik} \prod_{l=1}^{L} \prod_{a=1}^{A_i} p_{nila}^{nila}_a, \quad (3.1)$$

written in terms of the data matrix $X$ or sufficient statistics $N$.

This mixture model was first fully estimated using the EM by (Smouse et al., 1990), who also considered incomplete data with genotypes missing at random. It has subsequently been rederived without missing data in (Chen et al., 2006). The details of the EM with and without missing data are handled in the appendix.

3.3.1.2 Admixture Model

The admixture model (see (Pritchard et al., 2000b), (Tang et al., 2005)) allows some level of interbreeding between the $K$ ancestral subpopulations so that alleles in a single individual can be derived from more than one ancestral subpopulation (case of ‘modest’ or ‘substantial’ connectivity, as mentioned in Waples and Gaggiotti(Waples and Gaggiotti, 2006)). It is important to note that the LE assumption implies that the interbreeding has continued for enough generations to eliminate the linkage disequilibrium induced by mixing populations. The admixture model has been defined with varying forms for the mixing proportions. Pritchard et al. (2000b) define the model with distinct mixing proportions $\eta_i = (\eta_{i1}, \eta_{i2}, \ldots, \eta_{iK})$ for each individual $i$. Wu et al. (2006), and the accompanying technical report (Liu et al., 2006), propose two
variations: a *constrained admixture model* where individuals share mixing proportions, \( \eta = \eta_i \) and an *overparameterized admixture model* with distinct mixing proportions for each individual and locus, \( \eta_{il} = (\eta_{il1}, \eta_{il2}, \ldots, \eta_{ilK}) \). Perhaps the most biologically realistic of these models is *constrained admixture*, which arises by founding a population from \( K \) sources (subpopulations) and then breeding it according to the HWE assumptions until LE is reestablished. The other models are misspecifications of more complex realities and hence only approximate the true structure in the data. Probably the most serious of these approximations is LE, which was first relaxed by Falush et al. (2003), but will not be explored here.

Under the admixture model, all observed alleles are independent and the likelihood becomes

\[
L(\Theta | X) = \prod_{i=1}^{I} \prod_{l=1}^{L} \prod_{m=1}^{M} \prod_{k=1}^{K} \sum_{il} \eta_{kllim} = \prod_{i=1}^{I} \prod_{l=1}^{L} \prod_{a=1}^{A_l} \left( \sum_{k=1}^{K} \eta_{kla} \right)^{n_{ila}},
\]

again displayed for the data matrix \( X \) or sufficient statistics \( N \).

### 3.3.2 Model Selection

#### 3.3.2.1 Information Criteria

Choice of \( K \) is a model selection problem. Two information criteria, AIC = \(-2 \log L + 2p\) (Akaike, 1974), and BIC = \(-2 \log L + p \log n\) (Schwarz, 1978) are popularly used for model selection, where \( p \) is the total number of parameters and \( n \) is the number of observations. Intuitively, we require that \( \log L \) be as large as possible, and the AIC and BIC as small as possible, and choose \( K \) to minimize AIC or BIC. For our models, we set \( n = I \) and let \( A = \sum_{l=1}^{L} (A_l - 1) \), then the number of parameters is \( p = (K - 1) + KA \) for the mixture model and \( p = I(K - 1) + KA \) for the admixture model. Generally, model-based clustering in simulation studies tends to overestimate \( K \) by AIC, but underestimate it by BIC (Maitra and Melnykov, 2010). Congruence between clusterings identified by AIC and BIC, with the ‘true’ \( K \) are then computed by using a Rand index ((Rand, 1971)). The Rand index is calculated as a ratio of the number of agreements in clustering to the total number of comparisons, and can range from 0 to 1, with a value of 1 indicating agreements in all replicates.
3.3.2.2 Parameteric Bootstrap

If the null is nested within the alternative, then it might be reasonable to reject $H_0$ if the likelihood ratio $T = -2[l_0(\theta_0 \mid X) - l_A(\theta_A \mid X)]$ is in the right tail of the asymptotic $\chi^2$-distribution. However, the theory supporting this asymptotic distribution breaks down for mixture models (McLachlan and Krishnan (1996)). Instead, bootstrapping (Efron, 1979) can be used to test the null hypothesis $H_0 : K = k$ vs. the alternate hypothesis $H_a : K = k^*$ where $k \neq k^*$. The overall best model can be chosen by repeatedly testing $K = k$ against $K = k + 1$ until the null is no longer rejected, while controlling error rates due to multiple testing (Maitra et al. (2012)). The parametric bootstrap generates new datasets (of multilocus genotypes, of the same size as the observed data) from the fitted $K = k$ model. The data is refit to both $H_0$ and $H_a$ and the likelihood ratio test statistic $T_b$ is recorded. The $T_b$ form an empirical sampling distribution for $T$ under $H_0$. We then compare the statistic $T$ to the bootstrap empirical distribution, and obtain the $p$-value as $\frac{1}{B} \sum_{b=1}^{B} I(T_b > T)$. If the $p$-value is less than 0.05, then we reject $H_0$, otherwise we take $k$ as the better solution. Simply, we use $k^* = k + 1$ and sequentially test for $k = 1, \ldots, K_{\text{max}}$. For each pair, $k$ and $k + 1$, we generate $B = 100$ (default) bootstrap samples from a fitted model with $K = k$, and perform the test given above. As the tests stop at some $k$ where $H_0$ is not rejected or $k$ reaches $K_{\text{max}}$, then we claim this $k$ is the $K$ that is statistically supported by the data. We use $K_{\text{max}} = 6$ for illustrations using simulations in this paper.

3.3.3 Simulations and Datasets

We simulated microsatellite genotype datasets (hereafter Simulation 1) from the coalescent using SIMCOAL (Excoffier et al. (2000)) under three different models. Model 1 comprised 50 individuals in each of two populations, with 10 unlinked loci on the same chromosome. The two populations diverged 1000 generations ago and have evolved independently with no migration since, and a mutation rate of 0.005 per generation. Model 2 comprised 50 individuals in each of two populations, at 10 loci on the same chromosome, which split 1000 generations ago, and have since evolved with a bidirectional migration rate of 0.005 per generation, with the same
mutation rate as in Model 1. Model 3 comprised 50 individuals in a single population, with 10 unlinked loci on the same chromosome, which has been continually evolving for 1000 generations with a mutation rate of 0.005 mutations per site per generation. Models 5, and 6 were generated under the multinomial distribution, sampling from the allele frequency distributions of Model 1, and Model 2, respectively, with 3 loci, and 5 alleles per locus. Twenty different replicate datasets were generated under each model.

Further, we used four datasets identified by Gilbert et al. (2012) as not offering consistent results using the software STRUCTURE. We obtained genotype data (microsatellites, AFLP’s, SNP’s) from the Dryad database (https://datadryad.org/), or other author resources. We tested for consistency of MULTICLUST by using multiple initializations (with different seeds under ‘random initialization’ - see Appendix), and counting the number of times that MULTICLUST achieved the same value for $K$ (ancestral subpopulations) using $AIC$ or $BIC$, compared to STRUCTURE (see Section 2.4).

We also tested the methods on a multilocus dataset from 212 Blanding’s turtles (Emys blandingii) sampled from 18 locales across the midwestern United States and genotyped over 8 microsatellite loci. This semi-aquatic turtle is imperiled across most of its range ((Sethuraman et al., 2011)) and conservation efforts are currently underway. Life history and genetic studies are being undertaken in order to understand their phylogeography and apparent decline in numbers. The dataset has $I = 212$ turtles, each with $L = 8$ loci, with $M = 2$ since the chromosomes are diploid, and about 4.5% of the data is missing. Based on the derivations and implementation of the EM/QN algorithms above, we analyzed the dataset using the ‘mixture’ and ‘admixture’ models for putative values of $K = 1, 2, \ldots, 9$. The goals were threefold: (1) to identify reasonable values of $K$ for these data, (2) provide a partition of the data into subpopulations using the fitted models, and (3) compare models to determine evidence of ‘admixture’. For the mixture model and a given value of $K$, we assign each individual $i$ to the population, that is most likely \textit{a posteriori},

$$
\arg\max_{1 \leq k \leq K} P(V_{ik} = 1 \mid \Theta, X),
$$

where $V_{ik}$ indicates of individual $i$ is in subpopulation $k$ and is computed by Eq.(B.2) in the
final iteration $T$ of the EM algorithm. For the admixture model, alleles of a single individual may be derive from multiple ancestral subpopulations. We choose to assign individual $i$ to the population from whence most of its alleles putatively derived,

$$\arg \max_{1 \leq k \leq K} \sum_{l=1}^{L} \sum_{m=1}^{M} P(D_{ilmk} = 1 \mid \Theta, X),$$  \hspace{1cm} (3.4)

where this conditional probability $D_{ilmk}$ indicates if the $m^{th}$ allele at locus $l$ in individual $i$ is from subpopulation $k$ and is computed by Eq.(B.3) at EM convergence.

3.3.4 STRUCTURE Runs

To obtain results from STRUCTURE, we executed 10 runs (with different initializations) of each of the 20 replicate datasets under all six simulation conditions 3.1) using both the admixture model and the mixture model in STRUCTURE. Under all models, we used a burn-in period of 10000, and 100000 MCMC repetitions after burn-in, assuming no linkage. The parameter $\alpha$, which is the Dirichlet parameter that determines the degree of admixture, was inferred from the data. All other parameters were set to their default values for all runs. The method of Evanno et al. (2005) was used to obtain the most likely number of putative genetic subpopulations with the greatest second order rate of change in logarithmic probability of the data. (For details, see (Evanno et al., 2005)). For the empirical datasets compiled from Gilbert et al. (2012), we ran 10 iterations of STRUCTURE, under the admixture model (as suggested by Gilbert et al. (2012), and the original authors). These results were compiled and the best $K$ was chosen by using the method of Evanno et al. (2005) (referred to as method $E$ in Table 3.3), as well as using the method suggested by Pritchard et al. (2000b) (referred to as method $P$ in Table 3.3).

3.4 Results

3.4.1 Simulation

Results of comparable runs of MULTICLUST and STRUCTURE for the first set of simulations are shown in Table 3.2. We performed 10 runs each of both STRUCTURE and
MULTICLUST in order to obtain comparable results. The ‘true’ $K$ was estimated using the method of Evanno et al. (2005) and the method proposed by Pritchard et al. (2000b). MULTICLUST used $AIC$, and $BIC$ to choose $K$. The reported $K$’s inferred are the rounded median values over 20 replicate datasets, while the Rand indices (Rand (1971)) are calculated medians in comparison with the true structure specified in the simulation from Table 3.1. In general, STRUCTURE over-estimates the number of subpopulations in both Models 3 and 4, where there is no population structure ($K = 1$) but MULTICLUST finds $K = 1$ when using the BIC under both the mixture and the admixture models. MULTICLUST’s AIC also consistently overestimates the number of subpopulations, except for the Models 5 and 6, which were both simulated under the multinomial distribution. Both AIC and BIC using MULTICLUST under the mixture and the admixture model seem to infer the correct subpopulation structure in these Models (5 and 6), as does STRUCTURE. In general, bootstrap offers correct solutions under Models 4, 5, and 6, while it always over-estimates structure under Models 1, 2, and 3.

3.4.2 Empirical Datasets from Gilbert et al. 2012

Over all the four datasets that were inconsistent in their inference of $K$ from Gilbert et al. (2012), multiple iterations of MULTICLUST were consistent in identifying $K$, using both AIC and BIC over multiple initializations. On the other hand, our runs of STRUCTURE yielded different results for $K$, in two of the cases, compared to those reported by Gilbert et al. (2012) (see Table 3.3, Figs.3.1,3.2,3.3,3.4).

3.4.3 Blanding’s Turtle

We analyzed our datasets using AIC and BIC (Fig.3.5). The log likelihoods increase with $K$ and plateau after $K = 6$, supporting convergence to a global maximum likelihood peak. Due to the extra sets of parameters $\eta_{ik}$, the admixture models have larger log likelihood increases than the mixture models. While the AIC and BIC do not give clear solutions, AIC picks $K = 5, 4$ and BIC picks $K = 2, 1$ for the mixture and admixture models, respectively. An adjusted Rand index (Hubert and Arabie, 1985) was also obtained (close to 1 for perfect match) to determine the consistency of classifications between the mixture and admixture models that inferred the
same $K$. Except for $K = 1$, when $K = 5$ the maximum adjusted Rand index is reached. These lines of evidence provide a reasonable range, one to five. Fig. 3.6 shows a sample plot of mixture proportions laid out on a geographical scale, with the colors representing degrees of mixture in these populations, at a $K = 5$ level, under the mixture model (as chosen above by AIC - see Fig. 3.5). Interestingly, an unusual pattern of mixing is apparent (between widely-separated populations in Nebraska and Illinois), which was also reported using STRUCTURE and other analyses in Sethuraman et al. (2011).

### 3.5 Discussion

Several methods have been developed to infer genetic population mixtures using multilocus genotype data ((Pritchard et al., 2000b), (Tang et al., 2005), (Alexander et al., 2009)). Unfortunately, these methods can be slow, especially when used to estimate the number of subpopulations ($K$) for larger $K$, or for larger datasets, which often requires repeated runs. Perhaps as a consequence of insufficient replication, these methods often make unreliable inference on the ‘true’ value of $K$. Three major issues arise in utilizing these tools - (a) Inferring the true value of $K$ ancestral subpopulations, (b) Inconsistencies in inferred true values of $K$ over multiple iterations (and initializations) of these tools, and (c) Speed of computations, leading to limitations on testing greater values of $K$, and for larger datasets (with more individuals or loci or alleles per locus). MULTICLUST extends the acceleration strategy of Alexander et al. (2009) to multiallele traits and outperforms STRUCTURE((Pritchard et al., 2000b)) for estimation of $K$ with respect to speed of computation.

Over all the tests we performed with simulated and empirical data, we obtained comparable or better (more congruence with ‘true’ structure) results, in comparison with STRUCTURE.

Inference of both model parameters and $K$ is stochastic for all methods, and as a result multiple runs of the software, even with the same settings, may produce different estimates. Bayesian inference by MCMC is random by design and it is well known that improper run settings can lead to lack of reproducibility. Traditionally, estimation process is allowed to proceed for a certain ‘burn-in’ period, all estimates of which are discarded prior to the actual MCMC iterations that are retained and corresponding parameter estimates are obtained from. There
has been considerable debate on how long these burn-in periods and how many MCMC repetitions are needed for inference (see Gilbert et al 2012, Latch et al, Tutorial for STRUCTURE, BAPS, etc), with the general consensus that the longer, the better.

Meanwhile maximum likelihood methods find local maxima, and may produce different results across initializations, which requires multiple initializations to be performed to obtain better estimates. In light of the importance of repetition, the speed of inference becomes critical.

One key utility of MULTICLUST is that we often also obtained these results in a fraction of the time required to obtain convergence from STRUCTURE - a major hurdle for ‘problematic’ datasets, where inference of $K$ is difficult, either owing to great degrees of admixture or inflated parameter sets due to excessive heterozygosity and/or a large number of observed alleles at multiple loci. For instance, inference of the true number of subpopulations is challenging under evolutionary scenarios such as the Stepping Stone Model (a discrete approximation of the Isolation By Distance). In these scenarios, even though adjacent populations are more likely to be structured together into the same ancestral subpopulation, inference of true $K$ becomes muddled owing to the distribution of allele frequencies over a continuum of (sub)populations. On the other hand, inference of true $K$ becomes easier in populations that are significantly differentiated, with several alleles being fixed to localized subpopulations (eg. the Hierarchical Island Model).

But regardless of speed and efficiency of computation, the issues of difficulty in inferring $K$ persist with STRUCTURE and MULTICLUST for the same reasons of model assumptions on equilibria and sampling distribution of allele counts. Equilibrium assumptions are also yet to be relaxed for model-based clustering methods to identify subpopulation structure. One assumption, that of linkage equilibrium (LE) was relaxed and implemented into the admixture model by Falush et al. (2003), and subsequent versions of STRUCTURE((Pritchard et al., 2000b)). Of future research would be to incorporate chromosomal linkage into the likelihood framework, described in this manuscript. MULTICLUST also assumes that alleles are sampled from a randomly mating population in HWE, which follows a simple multinomial distribution. Natural populations need not necessarily be randomly mating, which could give rise to exceptions to
the multinomial sampling of allele counts. This issue is addressed in STRUCTURE((Pritchard et al., 2000b)), where allele counts are sampled according to probabilities drawn from a Dirichlet distribution instead, with a parameter, $\alpha$. Of potential interest is a Dirichlet extension of MULTICLUST, to account for exceptions to model assumptions on random mating.
3.6 Tables and Figures

<table>
<thead>
<tr>
<th>No.</th>
<th>Model</th>
<th>$I$</th>
<th>$K$</th>
<th>$L$</th>
<th>$M$</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>coalescent, no migration</td>
<td>100</td>
<td>2</td>
<td>10</td>
<td>≈ 10</td>
</tr>
<tr>
<td>2</td>
<td>coalescent, migration</td>
<td>100</td>
<td>2</td>
<td>10</td>
<td>≈ 10</td>
</tr>
<tr>
<td>3</td>
<td>coalescent</td>
<td>50</td>
<td>1</td>
<td>10</td>
<td>≈ 10</td>
</tr>
<tr>
<td>4</td>
<td>multinomial mixture</td>
<td>100</td>
<td>1</td>
<td>10</td>
<td>≈ 10</td>
</tr>
<tr>
<td>5</td>
<td>multinomial mixture</td>
<td>100</td>
<td>2</td>
<td>3</td>
<td>5</td>
</tr>
<tr>
<td>6</td>
<td>multinomial mixture</td>
<td>500</td>
<td>2</td>
<td>3</td>
<td>5</td>
</tr>
</tbody>
</table>

Table 3.1 Simulation 1 settings.

<table>
<thead>
<tr>
<th>Sim. No.</th>
<th>Fitted Model</th>
<th>$K$</th>
<th>$L$</th>
<th>$M$</th>
<th>Rand</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>Struct.</td>
<td>AIC</td>
<td>BIC</td>
<td>Boot</td>
</tr>
<tr>
<td>1</td>
<td>mixture</td>
<td>2</td>
<td>4</td>
<td>2</td>
<td>1.0</td>
</tr>
<tr>
<td>2</td>
<td>mixture</td>
<td>2</td>
<td>4</td>
<td>2</td>
<td>2.5</td>
</tr>
<tr>
<td>3</td>
<td>mixture</td>
<td>2</td>
<td>3</td>
<td>2</td>
<td>4</td>
</tr>
<tr>
<td>4</td>
<td>mixture</td>
<td>2</td>
<td>1</td>
<td>1</td>
<td>4</td>
</tr>
<tr>
<td>5</td>
<td>mixture</td>
<td>2</td>
<td>2</td>
<td>1</td>
<td>2</td>
</tr>
<tr>
<td>6</td>
<td>mixture</td>
<td>2</td>
<td>2</td>
<td>2</td>
<td>2</td>
</tr>
<tr>
<td>1</td>
<td>admixture</td>
<td>2</td>
<td>3.5</td>
<td>2</td>
<td>4</td>
</tr>
<tr>
<td>2</td>
<td>admixture</td>
<td>2</td>
<td>4</td>
<td>4</td>
<td>4</td>
</tr>
<tr>
<td>3</td>
<td>admixture</td>
<td>2</td>
<td>3</td>
<td>1</td>
<td>4</td>
</tr>
<tr>
<td>4</td>
<td>admixture</td>
<td>2</td>
<td>1</td>
<td>1</td>
<td>4</td>
</tr>
<tr>
<td>5</td>
<td>admixture</td>
<td>2</td>
<td>2</td>
<td>2</td>
<td>2</td>
</tr>
<tr>
<td>6</td>
<td>admixture</td>
<td>2</td>
<td>2</td>
<td>2</td>
<td>2</td>
</tr>
</tbody>
</table>

Table 3.2 Simulation results.
Figure 3.1 Plots of AIC versus $K$ using datasets from Gilbert et al. (2012)
Figure 3.2 Plots of BIC versus $K$ using datasets from Gilbert et al.2012.
Figure 3.3  Plots of mean Ln Probabilities estimated by STRUCTURE (Pritchard et al. 2000) on all datasets from Gilbert et al. 2012
Figure 3.4  Plots of $\Delta K$ estimated by the method of Evanno et al. 2005 on three datasets from Gilbert et al. 2012
Table 3.3  Inference of Ancestral Subpopulations, $K$ from empirical data reported in Gilbert et al. (2012). $P$ indicates models picked using the method of Pritchard et al. (2000), $P + E$ shows models picked using the method of Evanno et al. (2005).

<table>
<thead>
<tr>
<th>Dataset</th>
<th>$K$ tested</th>
<th>$P$</th>
<th>$P + E$</th>
<th>AIC</th>
<th>BIC</th>
<th>$P + E$ Original</th>
<th>$P + E$ Gilbert</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>1-4</td>
<td>2</td>
<td>2</td>
<td>2(2,2)</td>
<td>2(2,2)</td>
<td>NA</td>
<td>NA</td>
</tr>
<tr>
<td>2</td>
<td>1-10</td>
<td>3</td>
<td>3</td>
<td>2(2,3)</td>
<td>1(1,1)</td>
<td>3,4</td>
<td>3,5</td>
</tr>
<tr>
<td>3</td>
<td>1-3</td>
<td>3</td>
<td>2</td>
<td>3(3,3)</td>
<td>2(2,2)</td>
<td>NA</td>
<td>2</td>
</tr>
<tr>
<td>4</td>
<td>1-6</td>
<td>5</td>
<td>5</td>
<td>3(3,3)</td>
<td>1(1,1)</td>
<td>3</td>
<td>2</td>
</tr>
</tbody>
</table>

Figure 3.5  Results of mixture and admixture models for $K = 1, \ldots, 9$. 
Figure 3.6 Plot of mixture proportions of Blanding’s turtles at $K = 5$. Size of the pies indicates the sample sizes, colors indicate mixture proportions.
4.1 Abstract

The estimation of pairwise genetic relatedness is fundamental to applications in several fields, including quantitative genetics, conservation, genome-wide association studies (GWAS), and population genetics. Genetic relatedness can be classified into ‘recent’ and ‘deep’ descent, based on the choice of reference populations used to estimate it. The presence of population structure in sampled populations distorts relatedness measures owing to issues with computing allele frequencies with respect to a reference population. While several estimators have been described in the literature to obtain recent relatedness between two individuals, there is yet to be a robust framework for estimating both recent and deep relatedness between two individuals in the presence of population genetic structure. Anderson and Weir (2007), and Wang (2011b) develop maximum likelihood estimates of relatedness in the presence of population structure, but require that this population structure be known a priori. Here I propose a novel method to address the presence of population genetic structure in the estimation of genetic relatedness. Bias and mean squared errors in replicated estimation of genetic relatedness between admixed (i.e. genotypes derived from multiple subpopulations) full sib (FS), half sib (HS), first cousin (FC), parent-offspring (PO) and unrelated (UR) dyads shows considerably low bias and error using the new method, compared to previously developed methods.

4.2 Introduction

The genetic relatedness ($r_{XY}$) of two individuals $X$ and $Y$ can be defined in terms of the probability that their genes are Identical By Descent (IBD). Relatedness, $r_{XY}$ is also twice the
the coefficient of coancestry ($\theta_{XY}$) and can be thought of as the inbreeding coefficient of any offspring they may sire (Weir et al. (2006)).

This genetic relatedness can be classified into ‘recent’, and ‘deep’. Recent relatedness describes the relatedness between two individuals relative to a recent reference population, a few generations back. Deep relatedness measures the relatedness between two individuals due to both recent shared ancestry and ancient shared ancestry from pervasive inbreeding due to, for example, small population sizes.

Estimating genetic relatedness has been an important problem in biological statistics and population genetics. For instance, paternity or maternity assignment (see Avise (2001), Pearse et al. (2002), Yue and Chang (2010), Coleman and Jones (2011)), and forensic studies (reviewed in Weir (2004)), require a robust statistical framework to infer relatedness between genotyped individuals. Genetic relatedness also plays an important role in the study of quantitative traits, where the proportion of trait variability explained by shared alleles indicates the strength of the genetic component of the trait (Falconer and Mackay (1996), Visscher et al. (2008)). Association studies and linkage analyses without accounting for the increased relatedness due to population genetic structure could lead to spurious associations (Pritchard et al. (2000a)). Genetic relatedness is also important in fields such as conservation genetics (Oliehock et al., 2006).

Conventional relatedness estimators work in either of three ways – (1) estimating a coefficient of relatedness between two individuals using multilocus genotype data, or (2) assigning sib-ship partitions, reconstructing pedigrees, and using the pedigrees to estimate relatedness, or (3) directly estimating relatedness from known pedigrees (Weir et al., 2006)). All relatedness estimators have high variances, primarily owing to difficulty in parsing out true IBD alleles versus Identity By State (IBS) alleles (Blöin (2003)). This delineation of IBS versus IBD is achieved by estimating the conditional probabilities of observing a genotype in one individual $X$, given the observed genotype at the same locus in individual $Y$.

The presence of population genetic structure though, causes localized pervasive inbreeding. This makes individuals within the same subpopulation more related, than as suggested by their pedigree. The detection of subpopulation structure is complicated by migration. Genetic ad-
mixture, or the exchange of alleles between populations (Waples and Gaggiotti (2006)) is driven by incorporation of migrant genes into the receiving population’s gene pool. Genetic admixture across populations of a species hence acts to maintain homogeneity in genetic variants. Pervasive or specific inbreeding in recent generations past (between two related individuals) can be quantified though, if sufficient information is available on the existing genetic subpopulation structure. The estimated inbreeding coefficients (eg. \( \theta \), Weir (1994)) affect the afore mentioned conditional probabilities ((Weir, 1994)). Alternately, maintenance of advantageous alleles in subpopulations by selection (within a total population) could also yield ‘artificial’ patterns of relatedness between individuals that share alleles, but not by direct descent.

Not accounting for such ‘shared’ allelic ancestry by utilizing subpopulation allele frequencies leads to incorrect estimation of genetic relatedness. Anderson and Weir (2007) subvert this issue of estimating subpopulation allele frequencies by directly quantifying the amount of inbreeding (due to subpopulation structure), conditioned on \textit{a priori} knowledge of the existing subpopulations within a total population. Thus estimates of relatedness using the inbreeding coefficient \( \theta \) in its formulation could be potentially biased.

A lot of prior methods also utilize current population allele frequencies as proxies for ‘ancestral’ (this could mean subpopulation allele frequencies of the current generation, as in Anderson and Weir (2007), or allele frequencies of subpopulations from generations past, equated to current allele frequencies, as in Wang (2002)) subpopulation allele frequencies, under Hardy-Weinberg Equilibrium (HWE), in their estimates of the inbreeding coefficient, \( \theta \). This assumption can be problematic because, (a) we do NOT know the precise number of ancestral subpopulations (but can be approximated by the current subpopulation structure in a reference population, as assumed by the methods of Anderson and Weir (2007) and Wang (2011b)), and (b) the populations may not be in HWE. Most methods for estimating pairwise genetic relatedness assume that individuals whose pairwise relatedness is being estimated are derived from the same single, panmictic subpopulation, which is NOT necessarily true in most real life cases, where population structure is common. The methods of Anderson and Weir (2007), and Wang (2011b) that attempt to relax this assumption by handling individuals sampled from multiple subpopulations, assume that individuals derived from different subpopulations are genetically
unrelated. However, in the presence of genetic admixture and migration, alleles are shared between subpopulations.

One could surmise that similar allelic variants sampled from a gene pool (current ecological/geographical population) are also more likely to have been derived from the same recent ancestral genetic meta or sub-population, than by chance. This idea draws from the interplay between genetic structure of populations and relatedness (Anderson and Weir (2007), Wang (2011b)). This issue has to deal with the inflation or underestimation of genetic relatedness owing to the relative abundance or absence of a certain allele at a genetic locus in an individual, with respect to its frequency in the ancestral subpopulation from which it was derived, and the total population. ‘Deep’ descent is difficult to parse from ‘recent’ descent, unless relatedness is measured as relative to ancestral allele frequencies and ancestral subpopulation structure. Using the same example stated by Anderson and Weir (2007), if indeed a researcher is interested in estimating recent genetic relatedness (say estimate paternity or maternity), knowing the current population’s allele frequencies, as well as the current subdivision (or the subpopulation allele frequencies), he or she would be better off utilizing estimators that account for current subdivision allele frequencies, such as those developed by Anderson and Weir (2007), and Wang (2011b). On the other hand, if a researcher is interested in determining the suitability of two individuals for breeding them on a conservation strategy, he or she is likely interested in knowing how deeply inbred, or deeply related the individuals are by descent, in order to ensure creation of a population with maximum heterozygosity and genetic variation. Anderson and Weir suggest that methods that estimate this deep relatedness should ideally utilize current population allele frequencies to estimate genetic relatedness. But by their own argument, this approach assumes that current population allele frequencies are unchanging in time, and hence representative of the ancestral subpopulation or population frequencies. This assumption is surely untrue in most biological systems - most systems at some point in their evolutionary history have undergone population bottlenecks and expansions, which implies that current allele frequency distributions are likely not a good representation of the ancestral allele frequency distributions.

To account for unobserved population structure, I introduce a novel method for estimat-
ing genetic relatedness. This method of inferring genetic relatedness between individuals uses the following information: (a) admixture proportions of alleles at multiple loci in individuals, in most likely genetic subpopulations, as determined by likelihood or Bayesian methods such as those implemented in STRUCTURE ((Falush et al., 2007)), ADMIXTURE (Alexander et al. 2009), and MULTICLUST (Sethuraman et al. (in prep)), and (b) ancestral allele frequencies that are estimated as parameters in the model. The proposed estimator allows for the independent assortment of alleles at multiple genotyped loci, sampled from multiple ancestries. Specifically, the model uses the probability distribution that an allele at a locus in an individual, or a multi-locus genotype of an individual, was derived from a subpopulation in the recent past. I then calculate the IBS probabilities for two individuals, conditioned on their IBD states (Jacquard (1972), Anderson and Weir (2007)). This calculation contributes to a likelihood function (sensu Thomson 1977), which can then be maximized using a non linear programming or expectation maximization algorithm to obtain maximum likelihood estimates for relatedness coefficients. These relatedness coefficients are then utilized in calculating pairwise genetic relatedness, $r_{XY}$, and coancestry coefficients, $\theta_{XY}$. In this paper, I try to address two questions based on the new framework 1) how does this estimator of pairwise genetic relatedness compare with other estimators of relatedness for structured and unstructured populations?, and 2) how does this estimator compare to other estimators with increases in available information (measured in terms of the number of genotyped loci)? I first introduce the admixture model and its assumptions, and then develop the theory behind the likelihood equation for estimating relatedness in admixed (structured) populations. Thereon, I describe my scheme for simulations to test this method, and allied methods, that account for (and don’t) the presence of population structure. Then, I estimate genetic relatedness among individuals in simulations with commonly used relatedness estimators and compare the relative performance of these estimators with my admixture relatedness estimates.
4.3 Theory

4.3.1 Relatedness under the Admixture Model

Assume that each individual \( i \) has been genotyped at \( L \) codominant, unlinked, neutral, diploid loci. I also assume that the loci are unlinked to satisfy the independence assumption in calculating relatedness between alleles (and estimating ancestral allele frequencies and admixture proportions). Assume that each locus has \( M \) alleles (ploidy level). Neutrality is not a necessary assumption (and nor is ploidy level), but I assume neutral loci in a diploid organism for simplicity in describing the mathematical model. Also note that missing (due to erroneous genotyping) or unobserved alleles (due to incomplete sampling of individuals for genotyping) are missing at random. For a SNP marker, \( M = 2 \), for an AFLP marker (presence or absence data), \( M = 2 \), or for a microsatellite marker under the infinite alleles model, \( M = \infty \) (theoretically).

Under the admixture model (see Pritchard et al., Falush et al. (2007)), the proportion of alleles in individual \( i \in I \), derived from subpopulation \( k \in K \) is \( \eta_{ik} \), and the frequency of allele \( a \), at locus \( l \in L \), in ancestral subpopulation \( k \) is \( p_{kla} \). Under the assumptions of the model, for each individual \( i \), \( \sum_{k=1}^{K} \eta_{ik} = 1 \), and for each locus \( l \) and a subpopulation \( k \), \( \sum_{a=1}^{M} p_{kla} = 1 \), where \( M \) is the number of distinct alleles at the locus \( l \). Let \( n_{ila} \) be the number of alleles of type \( a \), at a locus \( l \). The likelihood of the observed data, \( X \), given the parameter set \( \theta = \eta_{ik}, p_{kla} \) under the admixture model then becomes:

\[
L(X \mid \Theta) = \prod_{i=1}^{I} \prod_{l=1}^{L} \prod_{a=1}^{A} \left( \sum_{k=1}^{K} \eta_{ik} p_{kla} \right)^{n_{ila}}, \tag{4.1}
\]

If two individuals were full-sibs from parents from the same subpopulation, genetic relatedness estimated using ancestral subpopulation frequencies would be expected to account for deep descent (and potential inbreeding) of the parents. The relatedness between these full-sibs, estimated using the parameters of the admixture model should be as close to the true estimate, i.e. \( r_{XY} = 0.5 \), as possible. On the other hand, if two individuals are full-sibs from parents derived from two different subpopulations, genetic relatedness estimated using current subpopulation allele frequencies would likely be an over- or under-estimate because the recent
admixture event is not accommodated (between the two parents in the previous generation). This result permits defining conditional probabilities of IBS states, given their IBD state using this new parametrization, sensu Jacquard (1972), Thomson (1979), Anderson and Weir (2007), Wang (2002), Wang (2011b), etc. This method is different from previous methods, since those models assume that the sampling process of alleles is from just ONE ancestral subpopulation. More plausible, however, is that all individuals in a species comprise invariably of a mosaic of genotypes, derived from multiple ancestral subpopulations.

Following the leads of Jacquard (1972), Anderson and Weir (2007), and Wang (2011b), I define the same set of nine IBD conditions (see 4.1), \{D_1, D_2, \ldots, D_9\} given a diploid locus between two individuals, 1 and 2. Each IBD condition could have nine (or more) possible IBS states, \{S_1, S_2, \ldots, S_9\}. Under the above assumptions, the probability that an allele \(a_p\), is observed at a locus \(l\), in individual \(i\) is \(\sum_{k=1}^{K} p_{klp} \eta_{ik} = Z_{pi}\), the probability that an allele \(a_q\), observed at the same locus \(l\), in individual \(j\) is \(\sum_{k=1}^{K} p_{klq} \eta_{jk} = Z_{qj}\), and so on. With these definitions, I define all the conditional probabilities, \(P(S_x | D_y)\) in Table 4.1. The likelihood of the IBD states over a single locus, \(L(X | \Delta)\) can be written (see Thomson 1977, Weir and Anderson 2007) as

\[
L(X | \Delta) = P(S_x | \Delta) = \sum_{y} P(S_x | D_y) \Delta_y
\]

(4.2)

, where \(\Delta\) is the vector of 9 IBD probabilities, \(X\) is the observed data, and \(S_x\) is the observed IBS state of \(x \in X\). Over \(L\) independent loci, this likelihood can be written as a product of individual locus likelihoods as

\[
L(X | \Delta) = \prod_{l}^{L} P(S_x | \Delta) = \prod_{l}^{L} \sum_{y}^{1,2,\ldots,9} P(S_x | D_y) \Delta_y
\]

(4.3)

This likelihood function can be maximized using the constraints that each IBD coefficient, \(\Delta_y, y \in 1, \ldots, 9\) is \(\geq 0\) and \(\leq 1\), and \(\sum_{y=1}^{9} \Delta_y = 1\). I used the solnp function in the Rsolnp package in R (Ghalanos and Theussl, 2012), which implements the augmented Lagrange method of Ye (1987) to solve this nine-dimensional problem with linear constraints. The coancestry coefficient, \(\theta_{XY}\), between two individuals \(X\) and \(Y\) then can be calculated as \(\theta_{XY} = \Delta_1 + \frac{1}{2}(\Delta_3 + \Delta_5 + \Delta_7) + \frac{1}{4}\Delta_8\) and, by definition, the relatedness as \(r_{XY} = 2\theta_{XY}\). Note that \(r_{XY}\) is \(\leq 1\) only if the population is outbred (\(\Delta_j, j = 1, \ldots, 6 = 0\), and \(\Delta_7, \Delta_8, \Delta_9 \neq 0\)).
4.3.2 Other Relatedness Estimators

I also implemented the methods of Anderson and Weir (2007) and Wang (2011b) under the same optimization framework, using Rsolnp. In both cases, subpopulation allele frequencies are modeled under the Dirichlet distribution, with the global parameter, $\theta$, measured as the probability that two randomly sampled individuals from a subpopulation are IBD under a simple island model. Anderson and Weir (2007) do not state explicitly how they estimate $\theta$, but Wang (2011b) indicates using the Weir and Cockerham $\theta$ estimator (1984), which I use as well in the framework of Anderson and Weir (2007) (and Wang (2011b)) to obtain comparable relatedness estimates. Regardless, under the equilibrium assumption that population subdivision is unchanging in time, the probability that two randomly drawn alleles are IBD (or from the same subpopulation) is $p_{la} + (1 - \theta)p_{la}$, where $p_{la}$ is the frequency of allele $a$ at a locus $l$ in the ancestral population. This leads into the same likelihood framework described above (4.2, 4.3), for the estimators of Anderson and Weir (2007), and Wang (2011b). I used the same non-linear programming method in 9 variables ($\Delta_i, i \in 1, 2, \ldots, 9$) to obtain maximum likelihood estimates. Genetic relatedness, $r_{XY}$ and the coancestry coefficient, $\theta_{XY}$ are obtained as before. Wang (2011b) offers another numerical solution by using Powell’s quadratically convergent method (Press et al 1996) to obtain likelihood estimates for all 9 variables above, as well as derived moment estimators under the same population structure framework (accounting for inbreeding using the inbreeding coefficient, $\theta$) for other previously derived estimators (Queller and Goodnight (1989), Lynch and Ritland (1999), Wang (2002)). Other estimators that I utilized in my comparisons include that of Queller and Goodnight (1989), which extends the method of Harpending (1979) to derive the relatedness coefficient as a ratio of the kinship coefficients of an individual to itself:

$$r_{YX} = \frac{\sum_c (p_{cy} - p)}{\sum_d (p_{dx} - p)} = \frac{f_y}{f_x}$$

where $c$ indexes all unique alleles in an individual $Y$ and $d$ indexes all unique alleles in an individual $X$, $p$ is the mean population allele frequency, and $f_x$ and $f_y$ are the kinship coefficients of an individual to itself (see Harpending’s extension to Hamilton’s rule - Harpending (1979)). Note that the Queller and Goodnight ((Queller and Goodnight, 1989)) estimator is asymmetric,
and in practice, genetic relatedness between two individuals is calculated as the average of $r_{XY}$ and $r_{YX}$. The Lynch and Ritland (1999) estimator and the Wang (2002) estimators, on the other hand, derive equations for the probability that a single allele in an individual is IBD with one in another individual ($\phi_{XY}$), and the probability that both alleles in an individual are IBD with both allele in another ($\Delta_{XY}$ which is equivalent to the IBD coefficients derived above).

Thereon, relatedness between two individuals $X$ and $Y$, $r_{XY}$ is derived as $r_{XY} = \phi_{XY}^2 + \Delta_{XY}$.

In the appendix, I derive versions of these three estimators using admixture proportions and ancestral allele frequencies, but were not implemented for the purpose of this manuscript. Instead, I utilized the seven methods implemented in the program, COANCESTRY (Wang (2011a) and see 4.6), which do not account for population genetic structure (and/or inbreeding) to compare relatedness estimated using admixture proportions against the truth. These include the methods of Lynch and Li (ref), Lynch and Ritland (1999), Queller and Goodnight (1989), Ritland (2005), Wang (2002), Wang (2007), which directly utilize current population allele frequencies in estimation of pairwise relatedness between two individuals. Derivations of multilocus (global) estimates of these methods are given elsewhere (see Wang (2011a) for review).

4.3.3 Simulations

I performed two separate sets of simulations to test the performance of genetic relatedness estimated as a function of admixture proportions and subpopulation allele frequencies (MC2013, hereon), against other estimators.

4.3.3.1 Varying $K$ - 1

Allele frequencies were simulated at 50 diploid, codominant, multiallelic (maximum of 50 allelic variants per locus) loci, using Easypop (Balloux (2001)). I utilized the Hierarchical Island Model, wherein each total population (out of 3) is comprised of subpopulations, which are in turn comprised of smaller subpopulations. I varied the number of subpopulations ($K$) to be one of 3, 5, 10, or 15. To allow for genetic admixture, I specified relatively greater levels of gene flow (0.01 migrant females and males per generation) between subpopulations inside each
population, and relatively lower gene flow (0.001 migrant females and males per generation) between populations. Subpopulation sizes (25 males and 25 females per subpopulation) were held constant across generations. I performed a forward-time simulation for 3000 generations and utilized the last generation’s allele frequency distribution for all further simulations. All populations at generation 3000 were tested for HWE, to ensure that a stable allele frequency distribution had been reached. Under this model, one would expect that individuals within the same subpopulation would be more related by deep and recent descent, than individuals from different subpopulations. I also did not control for inbreeding (ie. individuals underwent random mating), which allows one to test for models that incorporate and control for inbreeding versus those that do not.

I then simulated $k = 1000$ replicate dyads each of Parent-Offspring (PO), Full Sibs (FS), Half Sibs (HS), First Cousins (FC), and UnRelated (UR) individuals under different levels of known population subdivision ($K = 3, 5, 10, 15$) and both parents (or individuals for UR) were picked from the same subpopulation. Offspring genotypes were then simulated from these parents (or individuals) by random (Mendelian) segregation at all the 50 loci.

Under this scenario (same subpopulation parents), the population subdivision is ‘known’ while performing the simulations, i.e. source subpopulation $K$, of each individual sibling, offspring, or cousin is known. For the purpose of relatedness estimation, all admixture proportions, $\eta_{ik}$ and the ancestral allele frequencies $p_{kla}$ that assumed to be ‘true’ at the same $K$ as the parental $K$.

**4.3.3.2 Varying $K - 2$**

In order to further test the accuracy of this method in estimating relatedness in the presence of genetic subpopulation structure, I performed another set of simulations using SIMCOAL v.2.1 (Excoffier et al. (2000)). SIMCOAL is a backward time simulator under the coalescent, and allows one to specify historical divergence processes in generations, ancestral migration (per generation), mutation and recombination rates per locus, among other complex processes. For this study, I performed simulations under the ‘Continent-Island’ model (or Source-Sink), where an ancestral population (continent) split 1000 generations ago into either of $K = 3, 5, 10$. 
subpopulations (islands). Thereon, each adjacent island (i.e. islands that are only immediately adjacent to each other are able to directly exchange genes) was set to exchange migrants at the rate of 0.005 of the total population size (set to a constant size of 1000) per generation. I simulated 300 unlinked genetic loci (microsatellites) and a mutation rate of 0.005 per allele per generation. Out of these 1000 individuals, I sampled a total of 500 individuals per island to build a dataset of 1500, 2500, and 5000 individuals respectively ($K = 3, 5, 10$). I performed global exact tests of Hardy-Weinberg Equilibrium (HWE) to ensure that these islands had equilibrated with respect to allele frequencies over the 1000 generations using Genepop v.4.2, with the alternate hypothesis that there exists significant heterozygote deficiency in the islands. But for a few populations (islands), most of the islands were in HWE (see Table 4.6).

These final datasets were thereon utilized to create 1000 pairs each of full-sibling, half-sibling, first cousin, parent-offspring, and unrelated individuals by the same strategy described above. I only picked parents (or individuals) from the same putative ‘island’ of origin. Thereon, all analyses above were repeated.

### 4.3.3.3 Varying Number of Loci

For the third set of simulations, I was interested in observing the effect that increasing the sampling (number of observed loci) had on bias and MSE of estimates of genetic relatedness while accounting for (and not) population genetic structure. To address this question, I simulated 1000 replicate dyads of full-siblings from parents chosen from the same ancestral subpopulation according to the same hierarchical island model described above, under different levels of ancestral subpopulation structure ($K = 3, 5, 10$). I varied the number of observed loci between 10 and 40, to simulate a realistic scenario wherein individuals are genotyped at $< 50$ variant loci. I then performed the same analyses as listed above to obtain bias and MSE.

All admixture proportions and allele frequency estimates were obtained using MULTI-CLUST v.1.1. (Sethuraman et al.(in prep.)). The best model fitting the data (best $K$) was picked by comparing $BIC$ values calculated using the models of Sethuraman et al. (in prep). In order to obtain comparable estimates of relatedness using other relatedness measures, I supplied the program COANCESTRY (Wang (2011a)) with ‘true’ subpopulation allele frequencies.
(estimated assuming the subpopulation of origin of the parents from the simulation is the ‘true’ subpopulation), and utilized those in estimating relatedness (disregarding population structure, since none of the methods implemented in COANCESTRY account for structure). For comparing estimates of relatedness measures that account for population structure (Anderson and Weir (2007), Wang (2011b)), I estimated \( \theta \), using the \textit{geneclust} package in R, and utilized those estimates in the same IBD-IBS framework in R to obtain pairwise relatedness. The package \textit{geneclust} implements the method of Weir and Cockerham (1984) to obtain a normalized multilocus global \( \theta \) estimate.

I evaluated the deviation from true relatedness by calculating the Mean Square Error (MSE), as suggested by Wang (2011b). MSE for replicate \( r \) is measured as

\[
\frac{1}{r} \sum_{i=1}^{R} (\hat{r}_i - r_{true})^2
\]

(4.4)

where \( R \) is the total number of replicate dyads (here 1000), \( \hat{r}_i \) is the relatedness estimated using one of the above methods, and \( r_{true} \) is the true relatedness value, \( r_{xy} \), which is 0.5 for PO and FS dyads, 0.25 for HS dyads, 0.125 for FC dyads, and 0.0 for UR dyads. Bias was calculated as the deviation of the mean for all \( k = 1000 \) replicates under each scenario from the true mean.

\[
\bar{r}_{true} - \bar{r}_i
\]

(4.5)

### 4.4 Results

#### 4.4.1 Effect of Varying \( K - 1 \)

In general, in all scenarios that measured genetic relatedness among FS, PO, and HS dyads (accounting for ancestral subpopulation structure), my estimator (MC2013) performed better, or comparably with the AW (Anderson and Weir (2007)) and W (Wang (2011b)) estimators - see Figures 4.2,4.3,4.4,4.5,4.6 for bias estimates, and Figures 4.7,4.8,4.9,4.10,4.11 for MSE estimates. FS and PO relatedness had the least bias, compared to all other estimators. Interestingly, MC2013 performed worse than the AW and W estimators when estimating relatedness in FC and UR dyads, consistently under-estimating relatedness. The other estimators that did
not account for population structure consistently over-, or under-estimated genetic relatedness between dyads, with large mean squared errors (MSE).

Correspondingly, MC2013 had the lowest MSE in estimation of relatedness in FS, PO, and HS dyads from the same ancestral subpopulation, while the methods of AW and W had the lowest MSE for FC and UR dyads. The Ritland (2005) estimator, and the methods of Anderson and Weir (2007) and Wang (2011b), had the highest MSE for PO dyads, while the Ritland estimator (Ritland (2005)) had the highest MSE in all the cases. The estimators of Queller and Goodnight (1989), Lynch and Ritland (1999), and Wang (2007) performed similarly, with higher bias and MSE, than MC2013. Also, the estimators of Ritland, Queller and Goodnight may have values < 0 or > 1, but I did not truncate these to fall inside this range, as performed by Wang (2011b)) in order to observe the true trend in estimation of relatedness.

When comparing estimators and their biases in genetic relatedness of dyads with parents sampled from different subpopulation, the AW and W estimators had the least bias and MSE, as expected, except for PO dyads, where the TrioML (Wang (2011b)) and MC2013 estimators had the least bias. In general, accounting for inbreeding (MC2013 versus MC2013-WI) also led to smaller biases and MSEs.

4.4.2 Effect of Varying \(K\) - 2

In general, the new estimator (MC13) performed either better than or similar to the estimators of Anderson and Weir (2007), and Wang (2011b) across FS, PO, and HS dyads (see figures 4.15, 4.18, 4.16), while the methods of AW07 ((Anderson and Weir, 2007)) and Wang11 ((Wang, 2011b)) outperformed the new estimator in all FC and UR dyads (see figures 4.17 and 4.19). There was no definite trend with increase in degree of population structure (between K=3,5 and 10), but interestingly, all estimators had the least bias and MSE at \(K = 5\) across all relationship groups.

4.4.3 Effect of Number of Loci

Bias and MSE estimates of pairwise genetic relatedness in FS dyads showed a trend of decrease with increase in the number of loci (see - 4.12,4.13,4.14) across all estimators at \(K = 3\),
5, and 10, indicating the relative better estimation with increased genotypic information. In
general, the new estimator of relatedness had the least bias and least MSE in estimation of FS
dyads across different levels of available information (measured as a function of the number
of loci), with and without accounting for inbreeding (see - 4.12,4.13,4.14). The estimator that
accounted for inbreeding (MC2013-WI) outperformed all other estimators with the least bias
and MSE in estimation of FS relatedness. All other estimators of relatedness which did or did
not did not account for subpopulation structure performed with consistent decrease in bias and
MSE with increase in the number of analyzed loci, as expected. The Ritland estimator was the
least accurate, at $K = 3, 5, 10$, across $L = 10, 20, 30, 40$, followed by the estimators of Anderson
and Weir ((Anderson and Weir, 2007)), and Wang ((Wang, 2011b)).

4.5 Discussion

The presence of subpopulation structure affects estimates of pairwise relatedness between
individuals from the same subpopulation, owing to pervasive inbreeding in recent ancestral
They developed a likelihood-based estimator of IBD coefficients ($\Delta = \Delta_1, \Delta_2, \ldots, \Delta_9$) and
the coancestry coefficient $\theta_{XY}$, which gives rise to an estimate of the genetic relatedness, $r_{XY}$
for subdivided populations with recent admixture. Their estimator assumes that the current
sampling locations (and distribution of allele frequencies at these sampling locations) is repre-
sentative of the subdivision population since divergence from the ancestral population. This
assumption allows the current population allele frequencies and subdivision estimates in the
framework of Anderson and Weir (2007) and Thomson (1979), to obtain likelihood estimates of
recent relatedness. The primary goal of this paper was to develop a maximum-likelihood frame-
work using an alternate parametrization, to estimate pair-wise genetic relatedness between two
individuals $X$ and $Y$, while accounting for the ‘true’ genetic subpopulation structure in the
population. Since the proposal of an admixture model by Pritchard et al. (2000b), several
tools have been developed to estimate subpopulation structure (primarily to infer the number
of subpopulations, $K$, admixture proportions (here $\eta_{ik}$), and subpopulation allele frequencies,$p_{kla}$). These estimates have been applied widely, including to infer ancestral migration pat-
terns (eg. Rosenberg et al. (2002), Eriksson and Manica (2012), etc.), in association studies (eg. Collins-Schramm et al. (2002)), and to inform conservation decisions (see Allendorf et al. (2010) for review). To my knowledge, this is the first application of this approach to infer pairwise genetic relatedness, particularly in an evolutionary context. My new method utilizes inferred information from population structure studies (using methods such as STRUCTURE ((Pritchard et al., 2000b)) or MULTICLUST (Sethuraman et al.) - see Liu et al. (2013) for a review on software for inferring ancestral subpopulation structure that could also be used to infer population structure and subpopulation allele frequencies) to inform the estimation of relatedness.

From my simulations, this new method performs better, in all evolutionary scenarios where genetic admixture between two individuals are indicated. On the other hand, prior methods perform better, or on par, in cases with no deep IBD patterns, but only very recent admixture. In all estimates of FS and PO dyads, regardless of the number of subpopulations or the number of loci, MC2013 and MC2013-WI estimators had the least bias and MSE among all estimators.

As noted by Anderson and Weir ((Anderson and Weir, 2007)), estimates of relatedness in unrelated individuals are upwardly biased by all methods (see Fig. 4.6), except for those of Anderson and Weir ((Anderson and Weir, 2007)), and Wang((Wang, 2011b)). This result is simply an artifact of ignoring subpopulation structure (in the presence of undetected ancient admixture), which results in an upward bias for all estimates. While MC2013 and MC2013-WI account for this by using subpopulation allele frequencies, the other estimators (AW((Anderson and Weir, 2007), W((Wang, 2011b)) approximate it by using current allele frequencies, estimated from the sampled populations. A similar pattern is observed while estimating relatedness between FC dyads, again indicating that MC2013 estimators are accounting for and removing relatedness owing to ancient admixture and inbreeding of parents sampled from the same subpopulation.

Varying the number of loci minimally affects all relatedness estimators. This outcome may derive from variation in allele frequencies being sufficiently explained by the parameters of the admixture model (admixture proportions and subpopulation allele frequencies), as against biasing all estimates using a single non-varying parameter, $\theta$ (sensu Anderson and Weir (2007))
and Wang (2011b)). Several methods can estimate this coefficient $\theta$ and each method has its own biases and efficiencies. I used the method of Weir and Cockerham (1984), conditioned on the ‘true’ population structure, as being the population of origin of the parents. This approach could potentially cause increased bias and MSE in using the estimators of Anderson and Weir (2007) and Wang (2011b), which could be addressed by utilizing a population structuring method to assign individuals to subpopulations, conditioning on that population structure in estimating $\theta$. Regardless, increasing the number of sampled loci decreased bias of all estimators, as expected.

Different methods for inferring ancestral subpopulation structure have different assumptions on the allele frequency sampling process. For instance, STRUCTURE assumes that alleles are drawn according to the Dirichlet process, sampling around local mean allele frequencies (localized according to true sampling population structure), MULTICLUST assumes that alleles are drawn according to the multinomial process since an ancient admixture event.

The issue of ancestral versus current subpopulation structure is definitely critical. Researchers utilizing prior methods are in fact obtaining estimates of ancestral subpopulation structure and NOT current subpopulation structure. Hence approximating ancestral subpopulation structure using current subpopulation structure embodies significant caveats and assumptions - primarily that the evolutionary process has since driven allele frequencies to equilibrium. This outcome may be far from true in most wild populations. Thus, as shown here, these methods could inherently bias estimation of relatedness, and in turn lead to incorrect estimates of heritability, pedigrees, associations and linkages across the genome, among other fundamental genetic applications.

4.6 Figures and Tables
Figure 4.1  Jacquard’s (1977) IBD states, $D_1, \ldots, D_9$ for two diploid individuals. The top row shows the diploid genotype of individual 1, and the bottom row indicates the diploid genotype of individual 2. The alleles are connected by a line if they are Identical By Descent (IBD).
Figure 4.2 Bias in estimates of genetic relatedness between 1000 Full Sib (FS) dyads, with increasing degree of subpopulation structure. The top panel shows the bias estimates when parents of the FS dyads were chosen from the same subpopulation. I also performed another set of simulations, where the parents were picked from different subpopulations, bias estimates of which are shown in the bottom panel.
Figure 4.3  Bias in estimates of genetic relatedness between 1000 Half Sib (HS) dyads, with increasing degree of subpopulation structure. The top panel shows the bias estimates when parents of the HS dyads were chosen from the same subpopulation. I also performed another set of simulations, where the parents were picked from different subpopulations, bias estimates of which are shown in the bottom panel.
Figure 4.4  Bias in estimates of genetic relatedness between 1000 Parent-Offspring (PO) dyads, with increasing degree of subpopulation structure. The top panel shows the bias estimates when both parents of the PO dyads were chosen from the same subpopulation. I also performed another set of simulations, where the parents were picked from different subpopulations, bias estimates of which are shown in the bottom panel.
Figure 4.5  Bias in estimates of genetic relatedness between 1000 First Cousin (FC) dyads, with increasing degree of subpopulation structure. The top panel shows the bias estimates when parents of the FC dyads were chosen from the same subpopulation. I also performed another set of simulations, where the parents were picked from different subpopulations, bias estimates of which are shown in the bottom panel.
Figure 4.6  Bias in estimates of genetic relatedness between 1000 Unrelated Individual (UR) dyads, with increasing degree of subpopulation structure. The top panel shows the bias estimates when both unrelated individuals were chosen from the same subpopulation. I also performed another set of simulations, where both individuals were picked from different subpopulations, bias estimates of which are shown in the bottom panel.
Figure 4.7 Mean Squared Error in estimates of genetic relatedness between 1000 Full Sib (FS) dyads, with increasing degree of subpopulation structure. Top panel shows MSE when parents of FS were picked from the same subpopulation, bottom panel when they were picked from different subpopulations.
Figure 4.8 Mean Squared Error in estimates of genetic relatedness between 1000 Half Sib (HS) dyads, with increasing degree of subpopulation structure. Top panel shows MSE when parents of HS were picked from the same subpopulation, bottom panel when they were picked from different subpopulations.
Figure 4.9  Mean Squared Error in estimates of genetic relatedness between 1000 Parent-Offspring (PO) dyads, with increasing degree of subpopulation structure. Top panel shows MSE when both parents in the PO dyads were picked from the same subpopulation, bottom panel when they were picked from different subpopulations.
Figure 4.10  Mean Squared Error in estimates of genetic relatedness between 1000 First Cousin (FC) dyads, with increasing degree of subpopulation structure. Top panel shows MSE when parents of FC were picked from the same subpopulation, bottom panel when they were picked from different subpopulations.
Figure 4.11  Mean Squared Error in estimates of genetic relatedness between 1000 Unrelated Individual (UR) dyads, with increasing degree of subpopulation structure. Top panel shows MSE when both UR individuals were picked from the same subpopulation, bottom panel when they were picked from different subpopulations.
Figure 4.12  Bias and Mean Squared Error in estimates of genetic relatedness between 1000 Full Sib (FS) dyads sampled from $K = 3$ subpopulations, with increasing number of genotyped loci.
Figure 4.13 Bias and Mean Squared Error in estimates of genetic relatedness between 1000 Full Sib (FS) dyads sampled from $K = 5$ subpopulations, with increasing number of genotyped loci
Figure 4.14  Bias and Mean Squared Error in estimates of genetic relatedness between 1000 Full Sib (FS) dyads sampled from $K = 10$ subpopulations, with increasing number of genotyped loci
Figure 4.15 Bias and MSE estimates for 1000 full-sib dyads, simulated under the Continent-Island Model. X axis indicates the number of subpopulations (or islands), $K$. 
Figure 4.16  Bias and MSE estimates for 1000 half-sib dyads, simulated under the Continent-Island Model. X axis indicates the number of subpopulations (or islands), $K$. 
Figure 4.17  Bias and MSE estimates for 1000 first cousin dyads, simulated under the Continent-Island Model. X axis indicates the number of subpopulations (or islands), $K$. 
<table>
<thead>
<tr>
<th>IBS Mode</th>
<th>Allelic State</th>
<th>$D_1$</th>
<th>$D_2$</th>
<th>$D_3$</th>
<th>$D_4$</th>
<th>$D_5$</th>
<th>$D_6$</th>
<th>$D_7$</th>
<th>$D_8$</th>
<th>$D_9$</th>
</tr>
</thead>
<tbody>
<tr>
<td>$S_1$</td>
<td>$A_i A_i$, $A_i A_i$</td>
<td>$Z_1 Z_2$</td>
<td>$Z_1 Z_2$</td>
<td>$Z_1 Z_2$</td>
<td>$Z_1 Z_2$</td>
<td>$Z_1 Z_2$</td>
<td>$Z_1 Z_2$</td>
<td>$Z_1 Z_2$</td>
<td>$Z_1 Z_2$</td>
<td>$Z_1 Z_2$</td>
</tr>
<tr>
<td>$S_2$</td>
<td>$A_i A_i$, $A_j A_j$</td>
<td>$Z_1 Z_2$</td>
<td>$Z_1 Z_2$</td>
<td>$Z_1 Z_2$</td>
<td>$Z_1 Z_2$</td>
<td>$Z_1 Z_2$</td>
<td>$Z_1 Z_2$</td>
<td>$Z_1 Z_2$</td>
<td>$Z_1 Z_2$</td>
<td>$Z_1 Z_2$</td>
</tr>
<tr>
<td>$S_3$</td>
<td>$A_i A_i$, $A_i A_i$</td>
<td>$Z_1 Z_2$</td>
<td>$Z_1 Z_2$</td>
<td>$Z_1 Z_2$</td>
<td>$Z_1 Z_2$</td>
<td>$Z_1 Z_2$</td>
<td>$Z_1 Z_2$</td>
<td>$Z_1 Z_2$</td>
<td>$Z_1 Z_2$</td>
<td>$Z_1 Z_2$</td>
</tr>
<tr>
<td>$S_4$</td>
<td>$A_i A_i$, $A_j A_k$</td>
<td>$Z_1 Z_2$</td>
<td>$Z_1 Z_2$</td>
<td>$Z_1 Z_2$</td>
<td>$Z_1 Z_2$</td>
<td>$Z_1 Z_2$</td>
<td>$Z_1 Z_2$</td>
<td>$Z_1 Z_2$</td>
<td>$Z_1 Z_2$</td>
<td>$Z_1 Z_2$</td>
</tr>
<tr>
<td>$S_5$</td>
<td>$A_i A_i$, $A_i A_i$</td>
<td>$Z_1 Z_2$</td>
<td>$Z_1 Z_2$</td>
<td>$Z_1 Z_2$</td>
<td>$Z_1 Z_2$</td>
<td>$Z_1 Z_2$</td>
<td>$Z_1 Z_2$</td>
<td>$Z_1 Z_2$</td>
<td>$Z_1 Z_2$</td>
<td>$Z_1 Z_2$</td>
</tr>
<tr>
<td>$S_6$</td>
<td>$A_j A_k$, $A_i A_i$</td>
<td>$Z_1 Z_2$</td>
<td>$Z_1 Z_2$</td>
<td>$Z_1 Z_2$</td>
<td>$Z_1 Z_2$</td>
<td>$Z_1 Z_2$</td>
<td>$Z_1 Z_2$</td>
<td>$Z_1 Z_2$</td>
<td>$Z_1 Z_2$</td>
<td>$Z_1 Z_2$</td>
</tr>
<tr>
<td>$S_7$</td>
<td>$A_i A_j$, $A_j A_j$</td>
<td>$Z_1 Z_2$</td>
<td>$Z_1 Z_2$</td>
<td>$Z_1 Z_2$</td>
<td>$Z_1 Z_2$</td>
<td>$Z_1 Z_2$</td>
<td>$Z_1 Z_2$</td>
<td>$Z_1 Z_2$</td>
<td>$Z_1 Z_2$</td>
<td>$Z_1 Z_2$</td>
</tr>
<tr>
<td>$S_8$</td>
<td>$A_i A_j$, $A_i A_k$</td>
<td>$Z_1 Z_2$</td>
<td>$Z_1 Z_2$</td>
<td>$Z_1 Z_2$</td>
<td>$Z_1 Z_2$</td>
<td>$Z_1 Z_2$</td>
<td>$Z_1 Z_2$</td>
<td>$Z_1 Z_2$</td>
<td>$Z_1 Z_2$</td>
<td>$Z_1 Z_2$</td>
</tr>
<tr>
<td>$S_9$</td>
<td>$A_i A_j$, $A_i A_i$</td>
<td>$Z_1 Z_2$</td>
<td>$Z_1 Z_2$</td>
<td>$Z_1 Z_2$</td>
<td>$Z_1 Z_2$</td>
<td>$Z_1 Z_2$</td>
<td>$Z_1 Z_2$</td>
<td>$Z_1 Z_2$</td>
<td>$Z_1 Z_2$</td>
<td>$Z_1 Z_2$</td>
</tr>
</tbody>
</table>

Table 4.1  Conditional Probabilities $P(S_p \mid D_q)$
Figure 4.18  Bias and MSE estimates for 1000 parent-offspring dyads, simulated under the Continent-Island Model. X axis indicates the number of subpopulations (or islands), $K$. 

<table>
<thead>
<tr>
<th>$K$</th>
<th>MSE_PO</th>
<th>Bias_PO</th>
</tr>
</thead>
<tbody>
<tr>
<td>3</td>
<td>0.006</td>
<td>0.04</td>
</tr>
<tr>
<td>4</td>
<td>0.008</td>
<td>0.02</td>
</tr>
<tr>
<td>5</td>
<td>0.010</td>
<td>0.00</td>
</tr>
<tr>
<td>6</td>
<td>0.012</td>
<td>-0.02</td>
</tr>
<tr>
<td>7</td>
<td>0.014</td>
<td>-0.04</td>
</tr>
</tbody>
</table>

Estimators:
- AW07
- MC13
- MC13_WI
- Wang11
Figure 4.19  Bias and MSE estimates for 1000 unrelated individual dyads, simulated under the Continent-Island Model. X axis indicates the number of subpopulations (or islands), $K$. 
<table>
<thead>
<tr>
<th>Label</th>
<th>Reference</th>
<th>Accounts for structure?</th>
</tr>
</thead>
<tbody>
<tr>
<td>AW07</td>
<td>Anderson and Weir (2007)</td>
<td>Yes</td>
</tr>
<tr>
<td>Wang11</td>
<td>Wang (2011b)</td>
<td>Yes</td>
</tr>
<tr>
<td>MC13_WI</td>
<td>MULTICLUST with inbreeding</td>
<td>Yes</td>
</tr>
<tr>
<td>MC13</td>
<td>MULTICLUST</td>
<td>Yes</td>
</tr>
<tr>
<td>TrioML</td>
<td>Wang (2007)</td>
<td>No</td>
</tr>
<tr>
<td>LynchLi</td>
<td>Lynch (1988), Li et al. (1993)</td>
<td>No</td>
</tr>
<tr>
<td>LynchRi</td>
<td>Lynch and Ritland (1999)</td>
<td>No</td>
</tr>
<tr>
<td>Ritland</td>
<td>Ritland (1996)</td>
<td>No</td>
</tr>
<tr>
<td>QuellerG</td>
<td>Queller and Goodnight (1989)</td>
<td>No</td>
</tr>
<tr>
<td>DyadML</td>
<td>Milligan (2003)</td>
<td>No</td>
</tr>
</tbody>
</table>

Table 4.2  List of estimators tested and their references.
<table>
<thead>
<tr>
<th>K=3</th>
<th>Population</th>
<th>P-val</th>
<th>SE</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>0.7962</td>
<td>0.0229</td>
<td></td>
</tr>
<tr>
<td>2</td>
<td>&lt; 0.05</td>
<td>0</td>
<td></td>
</tr>
<tr>
<td>3</td>
<td>1</td>
<td>0</td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>K = 5</th>
<th>Population</th>
<th>P-val</th>
<th>SE</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>&lt; 0.05</td>
<td>0</td>
<td></td>
</tr>
<tr>
<td>2</td>
<td>&lt; 0.05</td>
<td>0</td>
<td></td>
</tr>
<tr>
<td>3</td>
<td>0.9935</td>
<td>0.0025</td>
<td></td>
</tr>
<tr>
<td>4</td>
<td>0.7911</td>
<td>0.0277</td>
<td></td>
</tr>
<tr>
<td>5</td>
<td>1</td>
<td>0</td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>K = 10</th>
<th>Population</th>
<th>P-val</th>
<th>SE</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>0.0169</td>
<td>0.0065</td>
<td></td>
</tr>
<tr>
<td>2</td>
<td>1</td>
<td>0</td>
<td></td>
</tr>
<tr>
<td>3</td>
<td>1</td>
<td>0</td>
<td></td>
</tr>
<tr>
<td>4</td>
<td>0.9719</td>
<td>0.0077</td>
<td></td>
</tr>
<tr>
<td>5</td>
<td>1</td>
<td>0</td>
<td></td>
</tr>
<tr>
<td>6</td>
<td>0.517</td>
<td>0.0311</td>
<td></td>
</tr>
<tr>
<td>7</td>
<td>0.0005</td>
<td>0.0005</td>
<td></td>
</tr>
<tr>
<td>8</td>
<td>1</td>
<td>0</td>
<td></td>
</tr>
<tr>
<td>9</td>
<td>&lt; 0.05</td>
<td>0</td>
<td></td>
</tr>
<tr>
<td>10</td>
<td>1</td>
<td>0</td>
<td></td>
</tr>
</tbody>
</table>

Table 4.3 Results of tests of Hardy-Weinberg Equilibrium ($H_a$=heterozygote deficiency) across all 300 unlinked loci simulated under the continent-island model for K=3, 5 and 10.
CHAPTER 5. CONCLUSION

The presence of population structure is ubiquitous in most wild populations of species. Detecting genetic population structure and understanding its consequences for the evolutionary trajectories of species has shaped a lot of our understanding of the process of evolution. This delineation of subdivision within a population plays an important role in several allied fields, including conservation genetics, association studies, phylogeography, and quantitative genetics. The goal of this dissertation was to understand methods to identify and infer subpopulation structure (discussed in Chapter 1), and to address some of its immediate applications in the field of conservation, and estimation of pairwise genetic relatedness. I have supplemented all the methods developed in this dissertation with several genetic datasets, from populations under simple evolutionary scenarios, to complexly admixed wild populations, which stand for the generic applicability of all these methods.

Chapter 2 of this dissertation discusses the standing motivation for the methods developed in the rest of the chapters - a classic population genetics study of the imperiled freshwater turtle, *Emys blandingii* across its primary range in the midwestern United States. We identified several interesting patterns of genetic admixture and subpopulation structure within the 18 genotyped populations of *E. blandingii*. Of particular conservation interest was the genetic structuring of populations from Grant County, Nebraska, with populations in the geographically distant counties of Carroll, and Will, Illinois. We tested several hypotheses of ancestral population divergences of all 18 populations, and showed significant support for the subpopulation structuring of these turtle populations into clusters created during climactic fluctuations in the Xerothermic period, along the recession of the Laurentide Ice Sheet. Besides having consequences for conservation of this threatened turtle, this chapter illustrates the array of genetic analyses that are dependent on the delineation of subpopulation structure in wild pop-
ulations, including AMOVA (Excoffier et al. (1992)), and estimation of ancestral migration and demographic parameters (Hey and Nielsen (2004)).

Chapter 3 focuses on the development of a robust statistical framework for the estimation of subpopulation admixture proportions, and allele frequencies as parameters of mixture, and admixture models (Pritchard et al. (2000b)), MULTICLUST. We have identified some common, but important issues with existing MCMC-based methods for inferring subpopulation structure, particularly with consistency, and efficiency of computation for large datasets. Our analyses of several coalescent simulations, and empirical genetic datasets (that were previously reported to have inconsistencies in estimation of subpopulation structure, quantified by the number of subpopulations, $K$ - Gilbert et al. (2012)), have shown that MULTICLUST offers a (1) time efficient, (2) consistent, and (3) reliable alternative to MCMC methods for identifying subpopulation structure. We are currently in the process of making further improvements to the time efficiency of the proposed algorithm, which should provide a reasonable good framework for hypothesis testing (and bootstrap) with relatively large genetic datasets.

In Chapter 4 of this dissertation, I detail a novel maximum-likelihood estimator of pairwise genetic relatedness that accounts for the presence of subpopulation structure. Prior relatedness estimation methods that account for the presence of subpopulation structure assume a priori that the location of sampling of genotyped individuals is indicative of the existing subpopulation structure in the total population ((Anderson and Weir, 2007),(Wang, 2011b)). This could lead to biased estimation of pairwise genetic relatedness, which I have addressed by offering an alternate parametrization to this problem. The new method (MC2013) offers quantification of subpopulation structure through admixture proportions, and subpopulation allele frequencies estimated using the methods from Chapter 3, and using this information to obtain estimates of pairwise relatedness. Analyses of several simulated genetic datasets of 1000 replicate dyads of full-siblings, half-siblings, parent-offsprings, first cousins, and unrelated individuals has demonstrated unbiased estimation using the new method, compared to prior methods. Of further research is testing new parametrizations of other commonly used estimators (reported in Appendix 2) against the MC2013 estimator.

Overall, this dissertation lays the foundation for several interesting questions that can be
addressed with a robust framework for identification of subpopulation structure. I hope that the methods developed in this dissertation will open the doors to new, unaddressed questions, and solutions, which will better our understanding of the evolution of populations.
APPENDIX A. APPENDIX TO POPULATION GENETICS OF BLANDING’S TURTLE IN THE MIDWESTERN UNITED STATES

Supplementary tables and figures
Figure A.1  Plots of mean logarithmic probability (estimated from STRUCTURE) versus number of putative populations (K) tested for Emys blandingii. (Top) Results from STRUCTURE using all loci; the most likely value of ‘K’ was identified as 4 (Bottom) Results from STRUCTURE without the *GmuD95* locus; the most likely value of ‘K’ was identified as 5
Figure A.2 The evolutionary history at the GmuD121 locus for *Emys blandingii* was inferred by using the Maximum Likelihood method based on the Kimura 2-parameter model [6]. The bootstrap consensus tree inferred from 1000 replicates [2] is taken to represent the evolutionary history of the taxa analyzed [2]. Branches corresponding to partitions reproduced in less than 50% bootstrap replicates are collapsed. The percentages of replicate trees in which the associated taxa clustered together in the bootstrap test (1000 replicates) are shown next to the branches [2]. Initial tree(s) for the heuristic search were obtained automatically as follows. When the number of common sites was \( \leq 100 \) or \( \leq 1/4 \) of the total number of sites, maximum parsimony method was used; otherwise the BIONJ method with MCL distance matrix was used. A discrete Gamma distribution was used to model evolutionary rate differences among sites (5 categories (+G, parameter = 200.0000)). The rate variation model allowed for some sites to be evolutionarily invariable (+[I], 31.6049% sites). The tree is drawn to scale, with branch lengths measured in the number of substitutions per site. The analysis involved 7 nucleotide sequences (= individuals). All positions containing gaps and missing data were eliminated, yielding 83 nucleotide positions in the final dataset. Evolutionary analyses were conducted in MEGA5 [10].
Figure A.3  The evolutionary history at the GmuD95 locus for Emys blandingii was inferred by using the Maximum Likelihood method based on the Jukes-Cantor model [6]. A discrete Gamma distribution was used to model evolutionary rate differences among sites (5 categories (+G, parameter = 4.1286)). The rate variation model allowed for some sites to be evolutionarily invariable ([+I], 68.6479% sites). Other methods are identical to those described for Fig. S3. The analysis involved 9 nucleotide sequences (= individuals). The final dataset contained 64 nucleotide positions.
Figure A.4  The evolutionary history at the GmuD21 locus for *Emys blandingii* was inferred by using the Maximum Likelihood method based on the Jukes-Cantor model [6]. A discrete Gamma distribution was used to model evolutionary rate differences among sites (5 categories (+G, parameter = 5.3515)). The rate variation model allowed for some sites to be evolutionarily invariable ([+I], 65.3268% sites). Other methods are identical to those described for Fig. S3. The analysis involved 6 nucleotide sequences (= individuals). The final dataset contained 94 nucleotide positions.

Figure A.5  Phylogenetic consensus tree constructed using a Neighbor Joining method and a bootstrap of 1000 replicates in PHYLIP v.3.69, placed alongside the population genetic structure of *Emys blandingii* inferred by STRUCTURE analysis. Diameter of pies on the map indicates sample sizes at those populations, while slices indicate admixture proportions, as estimated in STRUCTURE.
Group 1: Bremer-IA, Winnebago-IA, Worth-IA, Wright-IA, Palo Alto-IA, Butler-IA, Linn-IA
Group 2: Carroll-IL, Clinton-IA, Jones-IA, Muscatine-IA, Tama-IA
Group 3: Grant-NE
Group 4: Green-IA, Guthrie-IA (This group was removed from AMOVA calculations due to low sample size)
Group 5: McHenry-IL, Will-IL
Group 6: Scott-MN

Table A.1: Groups of populations of *Emys blandingii* within a radius of 100km of linear geographical distance (Model 1) from each other, as estimated by Geographical Distance Estimator. Populations that are underlined were excluded from analyses of population *F*-statistics, Mantel Tests of Isolation by Distance, population differentiation, and heterozygosities to prevent bias in allele frequency calculations, but they were included in analyses of population structure using STRUCTURE and in estimates of ancestral migration using IM.

Table A.2: Groups of populations of *Emys blandingii* based on watershed distribution (Model 2)
<table>
<thead>
<tr>
<th></th>
<th>County</th>
<th>Genotype</th>
<th>Size</th>
<th>VT</th>
<th>LR</th>
<th>DeltaL</th>
<th>DeltaV</th>
</tr>
</thead>
<tbody>
<tr>
<td>8</td>
<td>Will-IL</td>
<td>GmuD93</td>
<td>21</td>
<td>3</td>
<td>0.29</td>
<td>0.54</td>
<td>0.0006</td>
</tr>
<tr>
<td>9</td>
<td>Will-IL</td>
<td>GmuD87</td>
<td>21</td>
<td>7</td>
<td>0.62</td>
<td>0.65</td>
<td>0.0007</td>
</tr>
<tr>
<td>10</td>
<td>Will-IL</td>
<td>GmuD88</td>
<td>21</td>
<td>9</td>
<td>0.57</td>
<td>0.79</td>
<td>0.0009</td>
</tr>
<tr>
<td>11</td>
<td>Grant-NE</td>
<td>GmuD121</td>
<td>24</td>
<td>5</td>
<td>0.25</td>
<td>0.48</td>
<td>0.0012</td>
</tr>
<tr>
<td>12</td>
<td>Carroll-IL</td>
<td>GmuD121</td>
<td>21</td>
<td>6</td>
<td>0.57</td>
<td>0.69</td>
<td>0.0012</td>
</tr>
<tr>
<td>13</td>
<td>Wright-IA</td>
<td>GmuD95</td>
<td>9</td>
<td>3</td>
<td>0.89</td>
<td>0.63</td>
<td>0.0014</td>
</tr>
<tr>
<td>14</td>
<td>McHenry-IL</td>
<td>GmuD88</td>
<td>60</td>
<td>11</td>
<td>0.97</td>
<td>0.89</td>
<td>0.0018</td>
</tr>
<tr>
<td>15</td>
<td>Muscatine-IA</td>
<td>GmuD93</td>
<td>14</td>
<td>6</td>
<td>0.50</td>
<td>0.76</td>
<td>0.0019</td>
</tr>
<tr>
<td>16</td>
<td>McHenry-IL</td>
<td>GmuD55</td>
<td>60</td>
<td>5</td>
<td>0.70</td>
<td>0.71</td>
<td>0.002</td>
</tr>
<tr>
<td>17</td>
<td>Carroll-IL</td>
<td>GmuD21</td>
<td>22</td>
<td>5</td>
<td>0.41</td>
<td>0.59</td>
<td>0.0025</td>
</tr>
<tr>
<td>18</td>
<td>Muscatine-IA</td>
<td>GmuD21</td>
<td>14</td>
<td>6</td>
<td>0.57</td>
<td>0.75</td>
<td>0.0037</td>
</tr>
<tr>
<td>19</td>
<td>Carroll-IL</td>
<td>GmuD93</td>
<td>22</td>
<td>3</td>
<td>0.09</td>
<td>0.25</td>
<td>0.0053</td>
</tr>
<tr>
<td>20</td>
<td>Muscatine-IA</td>
<td>GmuD121</td>
<td>14</td>
<td>8</td>
<td>0.43</td>
<td>0.66</td>
<td>0.0069</td>
</tr>
<tr>
<td>21</td>
<td>McHenry-IL</td>
<td>GmuD121</td>
<td>61</td>
<td>6</td>
<td>0.59</td>
<td>0.65</td>
<td>0.0079</td>
</tr>
<tr>
<td>22</td>
<td>Bremer-IA</td>
<td>GmuD90</td>
<td>13</td>
<td>8</td>
<td>0.54</td>
<td>0.61</td>
<td>0.0145</td>
</tr>
<tr>
<td>23</td>
<td>McHenry-IL</td>
<td>GmuD87</td>
<td>61</td>
<td>10</td>
<td>0.74</td>
<td>0.74</td>
<td>0.0238</td>
</tr>
<tr>
<td>24</td>
<td>Wright-IA</td>
<td>GmuD90</td>
<td>9</td>
<td>4</td>
<td>0.78</td>
<td>0.61</td>
<td>0.0314</td>
</tr>
<tr>
<td>25</td>
<td>Grant-NE</td>
<td>GmuD87</td>
<td>21</td>
<td>3</td>
<td>0.14</td>
<td>0.22</td>
<td>0.0317</td>
</tr>
<tr>
<td>26</td>
<td>Will-IL</td>
<td>GmuD21</td>
<td>21</td>
<td>3</td>
<td>0.33</td>
<td>0.42</td>
<td>0.0317</td>
</tr>
<tr>
<td>27</td>
<td>Scott-MN</td>
<td>GmuD90</td>
<td>6</td>
<td>4</td>
<td>0.33</td>
<td>0.76</td>
<td>0.0484</td>
</tr>
<tr>
<td>28</td>
<td>Bremer-IA</td>
<td>GmuD93</td>
<td>13</td>
<td>5</td>
<td>0.46</td>
<td>0.61</td>
<td>0.0566</td>
</tr>
<tr>
<td>29</td>
<td>Wright-IA</td>
<td>GmuD93</td>
<td>9</td>
<td>3</td>
<td>0.11</td>
<td>0.22</td>
<td>0.0589</td>
</tr>
<tr>
<td>30</td>
<td>Bremer-IA</td>
<td>GmuD21</td>
<td>13</td>
<td>3</td>
<td>0.54</td>
<td>0.59</td>
<td>0.0634</td>
</tr>
<tr>
<td>31</td>
<td>Worth-IA</td>
<td>GmuD95</td>
<td>9</td>
<td>4</td>
<td>0.67</td>
<td>0.67</td>
<td>0.0641</td>
</tr>
<tr>
<td>32</td>
<td>Wright-IA</td>
<td>GmuD93</td>
<td>8</td>
<td>3</td>
<td>0.00</td>
<td>0.23</td>
<td>0.0669</td>
</tr>
<tr>
<td>33</td>
<td>Clinton-IA</td>
<td>GmuD90</td>
<td>7</td>
<td>6</td>
<td>0.43</td>
<td>0.68</td>
<td>0.0696</td>
</tr>
<tr>
<td>34</td>
<td>Worth-IA</td>
<td>GmuD121</td>
<td>8</td>
<td>5</td>
<td>0.50</td>
<td>0.72</td>
<td>0.0733</td>
</tr>
<tr>
<td></td>
<td>County-Name</td>
<td>GmuD</td>
<td>N</td>
<td>L</td>
<td>H</td>
<td>M</td>
<td></td>
</tr>
<tr>
<td>---</td>
<td>--------------</td>
<td>------</td>
<td>---</td>
<td>---</td>
<td>---</td>
<td>---</td>
<td>---</td>
</tr>
<tr>
<td>35</td>
<td>Wright-IA</td>
<td>GmuD55</td>
<td>9</td>
<td>4</td>
<td>0.56</td>
<td>0.53</td>
<td>0.0736</td>
</tr>
<tr>
<td>36</td>
<td>Bremer-IA</td>
<td>GmuD55</td>
<td>13</td>
<td>7</td>
<td>0.69</td>
<td>0.76</td>
<td>0.0736</td>
</tr>
<tr>
<td>37</td>
<td>Muscatine-IA</td>
<td>GmuD88</td>
<td>14</td>
<td>12</td>
<td>0.86</td>
<td>0.93</td>
<td>0.0751</td>
</tr>
<tr>
<td>38</td>
<td>Muscatine-IA</td>
<td>GmuD55</td>
<td>14</td>
<td>6</td>
<td>0.64</td>
<td>0.63</td>
<td>0.0768</td>
</tr>
<tr>
<td>39</td>
<td>McHenry-IL</td>
<td>GmuD90</td>
<td>61</td>
<td>4</td>
<td>0.61</td>
<td>0.47</td>
<td>0.0868</td>
</tr>
<tr>
<td>40</td>
<td>Winnebago-IA</td>
<td>GmuD90</td>
<td>6</td>
<td>3</td>
<td>0.17</td>
<td>0.32</td>
<td>0.0896</td>
</tr>
<tr>
<td>41</td>
<td>Scott-MN</td>
<td>GmuD93</td>
<td>9</td>
<td>4</td>
<td>0.44</td>
<td>0.65</td>
<td>0.0898</td>
</tr>
<tr>
<td>42</td>
<td>Carroll-IL</td>
<td>GmuD88</td>
<td>22</td>
<td>8</td>
<td>0.77</td>
<td>0.79</td>
<td>0.0926</td>
</tr>
<tr>
<td>43</td>
<td>Wright-IA</td>
<td>GmuD88</td>
<td>9</td>
<td>6</td>
<td>1.00</td>
<td>0.81</td>
<td>0.1077</td>
</tr>
<tr>
<td>44</td>
<td>Winnebago-IA</td>
<td>GmuD93</td>
<td>7</td>
<td>4</td>
<td>0.43</td>
<td>0.63</td>
<td>0.1083</td>
</tr>
<tr>
<td>45</td>
<td>Carroll-IL</td>
<td>GmuD90</td>
<td>19</td>
<td>2</td>
<td>0.63</td>
<td>0.44</td>
<td>0.109</td>
</tr>
<tr>
<td>46</td>
<td>McHenry-IL</td>
<td>GmuD93</td>
<td>61</td>
<td>2</td>
<td>0.15</td>
<td>0.19</td>
<td>0.1203</td>
</tr>
<tr>
<td>47</td>
<td>Carroll-IL</td>
<td>GmuD87</td>
<td>21</td>
<td>9</td>
<td>0.67</td>
<td>0.83</td>
<td>0.1401</td>
</tr>
<tr>
<td>48</td>
<td>Wright-IA</td>
<td>GmuD87</td>
<td>9</td>
<td>4</td>
<td>0.44</td>
<td>0.71</td>
<td>0.1421</td>
</tr>
<tr>
<td>49</td>
<td>Winnebago-IA</td>
<td>GmuD88</td>
<td>7</td>
<td>6</td>
<td>0.86</td>
<td>0.88</td>
<td>0.1453</td>
</tr>
<tr>
<td>50</td>
<td>Grant-NE</td>
<td>GmuD88</td>
<td>24</td>
<td>5</td>
<td>0.58</td>
<td>0.73</td>
<td>0.153</td>
</tr>
<tr>
<td>51</td>
<td>Winnebago-IA</td>
<td>GmuD55</td>
<td>7</td>
<td>4</td>
<td>0.43</td>
<td>0.49</td>
<td>0.1647</td>
</tr>
<tr>
<td>52</td>
<td>Jones-IA</td>
<td>GmuD95</td>
<td>6</td>
<td>4</td>
<td>0.33</td>
<td>0.56</td>
<td>0.1969</td>
</tr>
<tr>
<td>53</td>
<td>Clinton-IA</td>
<td>GmuD21</td>
<td>7</td>
<td>2</td>
<td>0.14</td>
<td>0.36</td>
<td>0.2302</td>
</tr>
<tr>
<td>54</td>
<td>Bremer-IA</td>
<td>GmuD88</td>
<td>13</td>
<td>13</td>
<td>0.92</td>
<td>0.92</td>
<td>0.2307</td>
</tr>
<tr>
<td>55</td>
<td>Clinton-IA</td>
<td>GmuD93</td>
<td>7</td>
<td>2</td>
<td>0.14</td>
<td>0.36</td>
<td>0.231</td>
</tr>
<tr>
<td>56</td>
<td>Will-IL</td>
<td>GmuD121</td>
<td>21</td>
<td>7</td>
<td>0.62</td>
<td>0.68</td>
<td>0.2388</td>
</tr>
<tr>
<td>57</td>
<td>Jones-IA</td>
<td>GmuD55</td>
<td>5</td>
<td>4</td>
<td>0.60</td>
<td>0.64</td>
<td>0.2411</td>
</tr>
<tr>
<td>58</td>
<td>Winnebago-IA</td>
<td>GmuD87</td>
<td>7</td>
<td>6</td>
<td>0.86</td>
<td>0.84</td>
<td>0.2675</td>
</tr>
<tr>
<td>59</td>
<td>Scott-MN</td>
<td>GmuD95</td>
<td>9</td>
<td>3</td>
<td>0.56</td>
<td>0.69</td>
<td>0.278</td>
</tr>
<tr>
<td>60</td>
<td>Scott-MN</td>
<td>GmuD87</td>
<td>9</td>
<td>4</td>
<td>0.56</td>
<td>0.71</td>
<td>0.355</td>
</tr>
<tr>
<td>61</td>
<td>Jones-IA</td>
<td>GmuD90</td>
<td>6</td>
<td>5</td>
<td>0.67</td>
<td>0.76</td>
<td>0.3778</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>---</td>
<td>-----</td>
<td>-----</td>
<td>-----</td>
<td>-----</td>
<td>-----</td>
<td>-----</td>
<td>-----</td>
</tr>
<tr>
<td>62</td>
<td>Worth-IA</td>
<td>GmuD87</td>
<td></td>
<td>9</td>
<td>7</td>
<td>0.78</td>
<td>0.75</td>
</tr>
<tr>
<td>63</td>
<td>Will-IL</td>
<td>GmuD90</td>
<td></td>
<td>20</td>
<td>2</td>
<td>0.40</td>
<td>0.51</td>
</tr>
<tr>
<td>64</td>
<td>Clinton-IA</td>
<td>GmuD87</td>
<td></td>
<td>7</td>
<td>5</td>
<td>0.57</td>
<td>0.67</td>
</tr>
<tr>
<td>65</td>
<td>Clinton-IA</td>
<td>GmuD88</td>
<td></td>
<td>7</td>
<td>6</td>
<td>0.71</td>
<td>0.80</td>
</tr>
<tr>
<td>66</td>
<td>Muscatine-IA</td>
<td>GmuD87</td>
<td></td>
<td>14</td>
<td>9</td>
<td>0.79</td>
<td>0.87</td>
</tr>
<tr>
<td>67</td>
<td>Clinton-IA</td>
<td>GmuD121</td>
<td></td>
<td>7</td>
<td>4</td>
<td>0.71</td>
<td>0.69</td>
</tr>
<tr>
<td>68</td>
<td>Grant-NE</td>
<td>GmuD93</td>
<td></td>
<td>24</td>
<td>6</td>
<td>0.46</td>
<td>0.58</td>
</tr>
<tr>
<td>69</td>
<td>Worth-IA</td>
<td>GmuD55</td>
<td></td>
<td>9</td>
<td>3</td>
<td>0.78</td>
<td>0.54</td>
</tr>
<tr>
<td>70</td>
<td>Worth-IA</td>
<td>GmuD90</td>
<td></td>
<td>9</td>
<td>3</td>
<td>0.67</td>
<td>0.54</td>
</tr>
<tr>
<td>71</td>
<td>Worth-IA</td>
<td>GmuD88</td>
<td></td>
<td>9</td>
<td>9</td>
<td>0.89</td>
<td>0.90</td>
</tr>
<tr>
<td>72</td>
<td>Scott-MN</td>
<td>GmuD21</td>
<td></td>
<td>9</td>
<td>4</td>
<td>0.44</td>
<td>0.54</td>
</tr>
<tr>
<td>73</td>
<td>Jones-IA</td>
<td>GmuD87</td>
<td></td>
<td>5</td>
<td>3</td>
<td>0.60</td>
<td>0.73</td>
</tr>
<tr>
<td>74</td>
<td>Winnebago-IA</td>
<td>GmuD21</td>
<td></td>
<td>7</td>
<td>3</td>
<td>0.71</td>
<td>0.58</td>
</tr>
<tr>
<td>75</td>
<td>Bremer-IA</td>
<td>GmuD87</td>
<td></td>
<td>13</td>
<td>10</td>
<td>0.85</td>
<td>0.84</td>
</tr>
<tr>
<td>76</td>
<td>Scott-MN</td>
<td>GmuD88</td>
<td></td>
<td>9</td>
<td>8</td>
<td>0.89</td>
<td>0.87</td>
</tr>
<tr>
<td>77</td>
<td>Jones-IA</td>
<td>GmuD88</td>
<td></td>
<td>6</td>
<td>7</td>
<td>0.83</td>
<td>0.83</td>
</tr>
<tr>
<td>78</td>
<td>Grant-NE</td>
<td>GmuD55</td>
<td></td>
<td>24</td>
<td>5</td>
<td>0.88</td>
<td>0.73</td>
</tr>
<tr>
<td>79</td>
<td>Carroll-IL</td>
<td>GmuD55</td>
<td></td>
<td>21</td>
<td>5</td>
<td>0.76</td>
<td>0.72</td>
</tr>
<tr>
<td>80</td>
<td>Bremer-IA</td>
<td>GmuD121</td>
<td></td>
<td>13</td>
<td>4</td>
<td>0.62</td>
<td>0.54</td>
</tr>
<tr>
<td>81</td>
<td>Clinton-IA</td>
<td>GmuD95</td>
<td></td>
<td>7</td>
<td>4</td>
<td>0.71</td>
<td>0.57</td>
</tr>
<tr>
<td>82</td>
<td>Clinton-IA</td>
<td>GmuD55</td>
<td></td>
<td>7</td>
<td>3</td>
<td>0.57</td>
<td>0.56</td>
</tr>
<tr>
<td>83</td>
<td>Grant-NE</td>
<td>GmuD90</td>
<td></td>
<td>20</td>
<td>2</td>
<td>0.45</td>
<td>0.45</td>
</tr>
<tr>
<td>84</td>
<td>Grant-NE</td>
<td>GmuD21</td>
<td></td>
<td>24</td>
<td>3</td>
<td>0.12</td>
<td>0.12</td>
</tr>
<tr>
<td>85</td>
<td>Jones-IA</td>
<td>GmuD121</td>
<td></td>
<td>5</td>
<td>2</td>
<td>0.40</td>
<td>0.36</td>
</tr>
<tr>
<td>86</td>
<td>Jones-IA</td>
<td>GmuD93</td>
<td></td>
<td>5</td>
<td>2</td>
<td>0.20</td>
<td>0.20</td>
</tr>
<tr>
<td>87</td>
<td>Jones-IA</td>
<td>GmuD21</td>
<td></td>
<td>5</td>
<td>2</td>
<td>0.60</td>
<td>0.56</td>
</tr>
<tr>
<td>88</td>
<td>McHenry-IL</td>
<td>GmuD21</td>
<td></td>
<td>61</td>
<td>3</td>
<td>0.25</td>
<td>0.22</td>
</tr>
</tbody>
</table>
Table A.6: Estimates of genotype and allele counts, as well as expected and observed heterozygosities, across loci in all populations of *Emys blandingii* sampled. Significant deviations from HWE after sequential Bonferroni correction are shown in gray, although the P-values given in the table are uncorrected. These tests were performed using 100,000 iterations in Arlequin v.3.0

<table>
<thead>
<tr>
<th></th>
<th>Location</th>
<th>Locus</th>
<th>N</th>
<th>A1</th>
<th>A2</th>
<th>HET</th>
</tr>
</thead>
<tbody>
<tr>
<td>89</td>
<td>Will-IL</td>
<td>GmuD55</td>
<td>21</td>
<td>4</td>
<td>0.33</td>
<td>0.30</td>
</tr>
<tr>
<td>90</td>
<td>Winnebago-IA</td>
<td>GmuD121</td>
<td>7</td>
<td>2</td>
<td>0.43</td>
<td>0.36</td>
</tr>
<tr>
<td>91</td>
<td>Winnebago-IA</td>
<td>GmuD95</td>
<td>7</td>
<td>10</td>
<td>1.00</td>
<td>0.96</td>
</tr>
<tr>
<td>92</td>
<td>Worth-IA</td>
<td>GmuD21</td>
<td>9</td>
<td>2</td>
<td>0.44</td>
<td>0.47</td>
</tr>
<tr>
<td>93</td>
<td>Wright-IA</td>
<td>GmuD121</td>
<td>9</td>
<td>3</td>
<td>0.56</td>
<td>0.46</td>
</tr>
<tr>
<td>Group</td>
<td>Populations</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>--------</td>
<td>-----------------------------------------------------------------------------</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Group 1</td>
<td>Bremer-IA, Butler-IA, Clinton-IA, Jones-IA, Linn-IA, Muscatine-IA, Tama-IA, Carroll-IL, Upper Mississippi outside/on border of the Des Moines Lobe of the Laurentide Ice Sheet</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Group 2</td>
<td>Scott-MN (Minnesota – Upper Mississippi Watersheds)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Group 3</td>
<td>Grant-NE (Missouri Watershed)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Group 4</td>
<td>McHenry-IL Will-IL (Southern Lake Michigan Crescent Watershed Illinois)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Group 5</td>
<td>Winnebago-IA Worth-IA Wright-IA Palo Alto-IA (Green-IA Guthrie-IA) (Upper Mississippi inside the Southern Des Moines Lobe of the Laurentide Ice Sheet)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Table A.3 Groups of populations of *Emys blandingii* based on watershed and Laurentide Ice Sheet distribution (Model 3)

<table>
<thead>
<tr>
<th>Group 1</th>
<th>Bremer-IA, (Butler-IA), Clinton-IA, Jones-IA, (Linn-IA), Winnebago-IA, Worth-IA, Wright-IA, (Tama-IA)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Group 2</td>
<td>Muscatine-IA, Scott-MN</td>
</tr>
<tr>
<td>Group 3</td>
<td>Carroll-IL, Grant-NE, Will-IL</td>
</tr>
<tr>
<td>Group 4</td>
<td>McHenry-IL</td>
</tr>
<tr>
<td>Group 5</td>
<td>(Green-IA, Guthrie-IA, Palo Alto-IA)</td>
</tr>
</tbody>
</table>

Table A.4 Groups of populations of *Emys blandingii* identified by STRUCTURE upon exclusion of the *GmuD95* locus (Fig 3). Populations in parentheses were excluded from IM and BayesAss analyses, owing to small sample sizes. Group 3 was also split into Grant-NE and (Carroll-IL, Will-IL) to resolve ancestral splits between these populations for IM, but not BayesAss, analyses.
Table A.5  Expected ($H_e$) and observed heterozygosities ($H_o$) across 201 genotyped individuals sampled from 12 populations of *Emys blandingii* with a sample size of greater than 5, estimated using GDA v.1.1. $N$ denotes the total number individuals genotyped at that locus, $H_e$ is the expected heterozygosity, and $H_o$ is the observed heterozygosity.

<table>
<thead>
<tr>
<th>Locus</th>
<th>N</th>
<th>$H_e$</th>
<th>$H_o$</th>
<th>%Missing</th>
</tr>
</thead>
<tbody>
<tr>
<td>1 GmuD121</td>
<td>190.00</td>
<td>0.75</td>
<td>0.53</td>
<td>5.47</td>
</tr>
<tr>
<td>2 GmuD95</td>
<td>195.00</td>
<td>0.88</td>
<td>0.36</td>
<td>2.99</td>
</tr>
<tr>
<td>3 GmuD93</td>
<td>200.00</td>
<td>0.69</td>
<td>0.26</td>
<td>0.50</td>
</tr>
<tr>
<td>4 GmuD90</td>
<td>190.00</td>
<td>0.73</td>
<td>0.55</td>
<td>5.47</td>
</tr>
<tr>
<td>5 GmuD88</td>
<td>201.00</td>
<td>0.94</td>
<td>0.83</td>
<td>0.00</td>
</tr>
<tr>
<td>6 GmuD87</td>
<td>197.00</td>
<td>0.85</td>
<td>0.64</td>
<td>1.99</td>
</tr>
<tr>
<td>7 GmuD55</td>
<td>190.00</td>
<td>0.74</td>
<td>0.66</td>
<td>5.47</td>
</tr>
<tr>
<td>8 GmuD21</td>
<td>201.00</td>
<td>0.68</td>
<td>0.33</td>
<td>0.00</td>
</tr>
<tr>
<td>9 All</td>
<td>195.50</td>
<td>0.78</td>
<td>0.52</td>
<td>2.74</td>
</tr>
<tr>
<td>Locus1</td>
<td>Locus2</td>
<td>$\chi^2$</td>
<td>df</td>
<td>P</td>
</tr>
<tr>
<td>------------</td>
<td>------------</td>
<td>---------</td>
<td>----</td>
<td>-----</td>
</tr>
<tr>
<td>GmuD121</td>
<td>GmuD21</td>
<td>29.983</td>
<td>20</td>
<td>0.07</td>
</tr>
<tr>
<td>GmuD88</td>
<td>GmuD55</td>
<td>28.923</td>
<td>20</td>
<td>0.089</td>
</tr>
<tr>
<td>GmuD93</td>
<td>GmuD21</td>
<td>30.919</td>
<td>22</td>
<td>0.098</td>
</tr>
<tr>
<td>GmuD121</td>
<td>GmuD88</td>
<td>24.575</td>
<td>18</td>
<td>0.137</td>
</tr>
<tr>
<td>GmuD93</td>
<td>GmuD55</td>
<td>28.295</td>
<td>22</td>
<td>0.166</td>
</tr>
<tr>
<td>GmuD95</td>
<td>GmuD90</td>
<td>26.711</td>
<td>22</td>
<td>0.222</td>
</tr>
<tr>
<td>GmuD88</td>
<td>GmuD87</td>
<td>26.404</td>
<td>22</td>
<td>0.235</td>
</tr>
<tr>
<td>GmuD95</td>
<td>GmuD55</td>
<td>24.096</td>
<td>20</td>
<td>0.238</td>
</tr>
<tr>
<td>GmuD55</td>
<td>GmuD21</td>
<td>22.729</td>
<td>20</td>
<td>0.302</td>
</tr>
<tr>
<td>GmuD95</td>
<td>GmuD93</td>
<td>23.648</td>
<td>22</td>
<td>0.366</td>
</tr>
<tr>
<td>GmuD93</td>
<td>GmuD90</td>
<td>24.509</td>
<td>24</td>
<td>0.433</td>
</tr>
<tr>
<td>GmuD90</td>
<td>GmuD21</td>
<td>22.241</td>
<td>22</td>
<td>0.446</td>
</tr>
<tr>
<td>GmuD95</td>
<td>GmuD87</td>
<td>22.221</td>
<td>22</td>
<td>0.447</td>
</tr>
<tr>
<td>GmuD87</td>
<td>GmuD55</td>
<td>22.103</td>
<td>22</td>
<td>0.454</td>
</tr>
<tr>
<td>GmuD95</td>
<td>GmuD88</td>
<td>18.796</td>
<td>20</td>
<td>0.535</td>
</tr>
<tr>
<td>GmuD90</td>
<td>GmuD55</td>
<td>20.621</td>
<td>22</td>
<td>0.544</td>
</tr>
<tr>
<td>GmuD95</td>
<td>GmuD21</td>
<td>18.362</td>
<td>20</td>
<td>0.564</td>
</tr>
<tr>
<td>GmuD121</td>
<td>GmuD93</td>
<td>17.073</td>
<td>20</td>
<td>0.648</td>
</tr>
<tr>
<td>GmuD88</td>
<td>GmuD21</td>
<td>15.529</td>
<td>20</td>
<td>0.745</td>
</tr>
<tr>
<td>GmuD121</td>
<td>GmuD55</td>
<td>16.31</td>
<td>22</td>
<td>0.8</td>
</tr>
<tr>
<td>GmuD121</td>
<td>GmuD87</td>
<td>15.983</td>
<td>22</td>
<td>0.817</td>
</tr>
<tr>
<td>GmuD121</td>
<td>GmuD95</td>
<td>14.198</td>
<td>20</td>
<td>0.82</td>
</tr>
<tr>
<td>GmuD121</td>
<td>GmuD90</td>
<td>15.23</td>
<td>22</td>
<td>0.852</td>
</tr>
<tr>
<td>GmuD87</td>
<td>GmuD21</td>
<td>15.075</td>
<td>22</td>
<td>0.859</td>
</tr>
<tr>
<td>GmuD93</td>
<td>GmuD88</td>
<td>14.936</td>
<td>22</td>
<td>0.865</td>
</tr>
<tr>
<td>GmuD90</td>
<td>GmuD87</td>
<td>14.369</td>
<td>22</td>
<td>0.888</td>
</tr>
<tr>
<td>GmuD93</td>
<td>GmuD87</td>
<td>11.715</td>
<td>24</td>
<td>0.983</td>
</tr>
<tr>
<td>GmuD90</td>
<td>GmuD88</td>
<td>5.3</td>
<td>20</td>
<td>1</td>
</tr>
</tbody>
</table>

Table A.7  Pair-wise tests of linkage disequilibrium, performed with 100000 permutations in Genepop v.4.1, on populations of Emys blandingii. The first two columns show the pairs of microsatellite loci tested, the third column shows the $\chi^2$ values from Fisher’s Exact Test, the fourth column indicates the number of degrees of freedom, and the fifth column contains the P-values. No significant linkage disequilibrium was detected at the genotyped loci, hence no corrections were performed for multiple comparisons.
Table A.8  Estimates of Weir and Cockerham $F$-statistics: $F_{is}$ (population differentiation between individuals among sampled locations), $F_{it}$ (population differentiation among individuals), and $F_{st}$ (population differentiation between populations), using GDA v.1.1. All analyses were performed with and without the $GmuD95$ locus for 202 individuals (from 12 well-sampled populations of *Emys blandingii*) using only a priori information on their sampling locations (no ‘grouping’). 95% confidence intervals were established by bootstrapping with 10,000 replicates.

<table>
<thead>
<tr>
<th></th>
<th>With $GmuD95$ Locus</th>
<th></th>
<th>Without $GmuD95$ Locus</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>95% CI</td>
<td></td>
<td>95% CI</td>
<td></td>
</tr>
<tr>
<td></td>
<td>$F_{is}$</td>
<td>Lower 0.027 0.275</td>
<td>$F_{is}$</td>
<td>Lower 0.010 0.165</td>
</tr>
<tr>
<td></td>
<td>0.136</td>
<td></td>
<td>$F_{it}$</td>
<td></td>
</tr>
<tr>
<td></td>
<td>0.363</td>
<td>Lower 0.233 0.502</td>
<td>$F_{it}$</td>
<td></td>
</tr>
<tr>
<td></td>
<td>0.263</td>
<td></td>
<td>$F_{st}$</td>
<td>Lower 0.198 0.471</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>0.270 0.178 0.379</td>
<td></td>
</tr>
<tr>
<td>Without Locus</td>
<td>GmuD95</td>
<td>Sum of Squares</td>
<td>Estimated Variance</td>
<td>% Variation</td>
</tr>
<tr>
<td>---------------</td>
<td>--------</td>
<td>----------------</td>
<td>-------------------</td>
<td>------------</td>
</tr>
<tr>
<td>Grouped by Geographical Distance (Model 1)</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Among 5 groups</td>
<td></td>
<td>127.03</td>
<td>0.1051 (P &lt; 0.001)</td>
<td>5.10</td>
</tr>
<tr>
<td>Among 12 populations within 5 groups</td>
<td></td>
<td>91.39</td>
<td>0.4959 (P &lt; 0.001)</td>
<td>24.09</td>
</tr>
<tr>
<td>Within 12 populations</td>
<td></td>
<td>571.31</td>
<td>1.4574 (P = 0.438±0.004)</td>
<td>70.80</td>
</tr>
<tr>
<td>Grouped by Watershed (Model 2)</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Among 4 groups</td>
<td></td>
<td>81.05</td>
<td>0.0792 (P &lt; 0.001)</td>
<td>4.82</td>
</tr>
<tr>
<td>Among 12 populations within 4 groups</td>
<td></td>
<td>99.95</td>
<td>0.4325 (P &lt; 0.001)</td>
<td>26.34</td>
</tr>
<tr>
<td>Within 12 populations</td>
<td></td>
<td>443.07</td>
<td>1.1303 (P = 0.337±0.005)</td>
<td>68.84</td>
</tr>
<tr>
<td>Grouped by Watershed and Pleistocene Distribution (Model 3)</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Among 5 groups</td>
<td></td>
<td>93.39</td>
<td>0.0793 (P &lt; 0.001)</td>
<td>4.85</td>
</tr>
<tr>
<td>Among 12 populations within 5 groups</td>
<td></td>
<td>88.51</td>
<td>0.4254 (P &lt; 0.001)</td>
<td>26.02</td>
</tr>
<tr>
<td>Within 12 populations</td>
<td></td>
<td>443.07</td>
<td>1.1303 (P = 0.498±0.005)</td>
<td>69.18</td>
</tr>
</tbody>
</table>

Table A.9  AMOVA results showing genetic variance for *Emys blandingii* populations under the hypothesized models specified in Tables S1, S2, and S3.
<table>
<thead>
<tr>
<th>Cluster1</th>
<th>Cluster2</th>
<th>$\theta_1$</th>
<th>$\theta_2$</th>
<th>$\theta_A$</th>
</tr>
</thead>
<tbody>
<tr>
<td>IA</td>
<td>McHenry-IL</td>
<td>7604.11(6955.86-8098.82)</td>
<td>1408.01(895.51-2460.58)</td>
<td>3168.73(1949.01-3646.39)</td>
</tr>
<tr>
<td>IA</td>
<td>Grant-NE</td>
<td>15242.33(14867.03-15992.93)</td>
<td>1013.03(920.61-1873.21)</td>
<td>5365.1(2243.28-7258.66)</td>
</tr>
<tr>
<td>Carroll-IL, Will-IL</td>
<td>McHenry-IL</td>
<td>6556.97(5761.63-8599.52)</td>
<td>2328.32(1777.24-2681.01)</td>
<td>3400.17(2672.75-3995.34)</td>
</tr>
<tr>
<td>IA</td>
<td>Scott-MN, Muscatine-IA</td>
<td>7424.99(6785.27-8184.12)</td>
<td>4424.14(2580.99-5283.48)</td>
<td>2452.25(1829.59-3782.86)</td>
</tr>
<tr>
<td>Scott-MN, Muscatine-IA</td>
<td>Carroll-IL, Will-IL</td>
<td>10377.31(8587.51-10863.29)</td>
<td>1244.96(747.88-2433.44)</td>
<td>5209.4(3324.77-5612.4)</td>
</tr>
<tr>
<td>Scott-MN, Muscatine-IA</td>
<td>McHenry-IL</td>
<td>5319.04(4050.77-5882.06)</td>
<td>1198.6(559.35-2311.79)</td>
<td>1395.69(121.49-2462.46)</td>
</tr>
<tr>
<td>Carroll-IL, Will-IL</td>
<td>Grant-NE</td>
<td>6059.88(4631.9-7587.28)</td>
<td>749.99(465.64-1325.82)</td>
<td>5694.27(4471.53-6433.61)</td>
</tr>
<tr>
<td>Grant-NE</td>
<td>McHenry-IL</td>
<td>10016.52(8935.96-14040.19)</td>
<td>2510.18(1650.49-3568.25)</td>
<td>5623.18(3917.03-6092.38)</td>
</tr>
<tr>
<td>Scott-MN, Muscatine-IA</td>
<td>Grant-NE</td>
<td>1169299.7(1169299.7-1171670.31)*</td>
<td>1901.65(1199.78-2080.97)</td>
<td>472745.56(474167.35-475589.15)</td>
</tr>
<tr>
<td>IA</td>
<td>Carroll-IL, Will-IL</td>
<td>1681177.01(1682882.9237.53)(1199.78-1684588.84)*</td>
<td>2080.97(567215.99-568921.9)</td>
<td></td>
</tr>
</tbody>
</table>

Table A.10 Maximum likelihood estimates of effective population sizes ($\theta$) for *Emys blandingii*. The values shown are HiPt values or the values with the most number of counts (median). The values in parentheses represent 95% confidence intervals around the mean. Values with a * failed to converge in our IM analyses.
<table>
<thead>
<tr>
<th>Cluster1</th>
<th>Cluster2</th>
<th>t</th>
<th>m1</th>
<th>m2</th>
</tr>
</thead>
<tbody>
<tr>
<td>IA</td>
<td>McHenry-IL</td>
<td>353250(185250-410750)</td>
<td>1.05(0.85-2.65)</td>
<td>10.65(6.75-23.35)</td>
</tr>
<tr>
<td>IA</td>
<td>Grant-NE</td>
<td>231750(187250-485250)</td>
<td>0.15(0.15-1.35)</td>
<td>2.95(1.95-6.85)</td>
</tr>
<tr>
<td>Carroll-IL,</td>
<td>McHenry-IL</td>
<td>197750(166250-400750)</td>
<td>1.55(1.25-2.75)</td>
<td>3.05(1.95-8.85)</td>
</tr>
<tr>
<td>Will-IL</td>
<td>IA</td>
<td>185250(165250-347750)</td>
<td>1.85(0.95-3.35)</td>
<td>1.65(1.05-3.95)</td>
</tr>
<tr>
<td>Scott-MN,</td>
<td>Carroll-IL, Will-IL</td>
<td>170250(117250-392250)</td>
<td>1.65(0.55-4.95)</td>
<td>6.25(4.45-28.55)</td>
</tr>
<tr>
<td>Muscatine-IA</td>
<td>Scott-MN, Muscatine-IA</td>
<td>79750(67750-461750)</td>
<td>1.65(0.85-6.45)</td>
<td>5.25(3.45-31.85)</td>
</tr>
<tr>
<td>Carroll-IL,</td>
<td>Grant-NE</td>
<td>22550(16550-90650)</td>
<td>2.25(0.75-3.85)</td>
<td>8.25(3.65-21.35)</td>
</tr>
<tr>
<td>Will-IL</td>
<td>Grant-NE</td>
<td>22250(20550-95450)*</td>
<td>0.85(0.65-4.75)</td>
<td>4.15(1.15-7.95)</td>
</tr>
<tr>
<td>Grant-NE</td>
<td>McHenry-IL</td>
<td>1250(1150-2750)</td>
<td>7.15(5.45-9.65)</td>
<td>7.25(5.65-12.45)</td>
</tr>
<tr>
<td>Scott-MN,</td>
<td>Grant-NE</td>
<td>850(950-1150)</td>
<td>13.45(11.75-16.95)</td>
<td>22.65(17.05-26.25)</td>
</tr>
<tr>
<td>Muscatine-IA</td>
<td>Carroll-IL, Will-IL</td>
<td>850(950-1150)</td>
<td>13.45(11.75-16.95)</td>
<td>22.65(17.05-26.25)</td>
</tr>
</tbody>
</table>

Table A.11 Maximum likelihood estimates of migration rates per generation (m) and estimated time since splitting in years (t) for *Emys blandingii*. The values shown are HiPt values or the values with the most number of counts (median). The values within parentheses represent 95% confidence intervals around the mean. Values with a * failed to converge in our IM analyses.
<table>
<thead>
<tr>
<th></th>
<th>Group1</th>
<th>Group2</th>
<th>Group3</th>
<th>Group4</th>
</tr>
</thead>
<tbody>
<tr>
<td>Group1</td>
<td>0.961 (0.932)</td>
<td>- 0.020 (-0.001)</td>
<td>- 0.012 (-0.004)</td>
<td>- 0.006 (-0.006)</td>
</tr>
<tr>
<td></td>
<td>0.990</td>
<td>0.042</td>
<td>0.029</td>
<td>0.018</td>
</tr>
<tr>
<td>Group2</td>
<td>0.012 (-0.011)</td>
<td>0.945 (0.896)</td>
<td>- 0.030 (-0.009)</td>
<td>- 0.013 (-0.011)</td>
</tr>
<tr>
<td></td>
<td>-0.036</td>
<td>0.993</td>
<td>0.069</td>
<td>0.036</td>
</tr>
<tr>
<td>Group3</td>
<td>0.019 (0.001)</td>
<td>- 0.005 (-0.004)</td>
<td>0.972 (0.951)</td>
<td>- 0.005 (-0.004)</td>
</tr>
<tr>
<td></td>
<td>0.036</td>
<td>0.013</td>
<td>0.993</td>
<td>0.014</td>
</tr>
<tr>
<td>Group4</td>
<td>0.006 (-0.005)</td>
<td>- 0.006 (-0.005)</td>
<td>- 0.005 (-0.005)</td>
<td>0.984 (0.966)</td>
</tr>
<tr>
<td></td>
<td>0.016</td>
<td>0.016</td>
<td>0.016</td>
<td>1.001</td>
</tr>
</tbody>
</table>

Table A.12  Estimates of recent migration rates, with 95% confidence intervals around the means in parentheses, for groups of *Emys blandingii* populations. These estimates were obtained using 10,000,000 iterations, with a burn in of 100,000 iterations, and sampling every 1000th value from the distribution, as suggested by Rannala (2007). These groups are those obtained from STRUCTURE, as listed in Table S4. Group 5 was excluded from these analyses owing to small sample sizes. Group 3 was not split into Grant-NE and (Carroll-IL, Will-IL) as with IM analyses, since BayesAss estimates recent migrations.
Table A.13  Pair-wise differentiation estimates between microsatellite flanking region sequences of *Emys blandingii* from Carroll-IL, Will-IL, Grant-NE, and McHenry-IL. These estimates were determined using DnaSP 5.10 (Librado et al. 2009) by using all the sequences from these populations.
APPENDIX B. APPENDIX TO MULTICLUST - FAST MULTINOMIAL CLUSTERING OF MULTILOCUS GENOTYPES TO INFERN GENETIC SUBPOPULATION STRUCTURE

EM Algorithm

In general, the Expectation-Maximization (EM) algorithm works by iteratively computing the conditional expectation of the complete-data log likelihood in the E-step, followed by maximization of expected complete-data log likelihood in the M-step until the observed log likelihood no longer increases (Dempster et al., 1977). The complete-data likelihood is the likelihood of the observed data and the missing data. The missing data are the population assignments of each individual, in the mixture model, or allele, in the admixture model. The missing data may also include literal missing allele observations when the data are incompletely observed. Our derivations are formulated in terms of the allele counts \( N = \{ n_{ila} \} \), but could be naturally adjusted to use the data matrix \( X = \{ x_{ilm} \} \).

Mixture Model

If we pretend to know \( V_{ik} \), which indicates if the individual \( i \) comes from the \( k \)th subpopulation, then the complete-data likelihood is

\[
L_c(\Theta, V \mid X) = \prod_{i=1}^{I} \prod_{k=1}^{K} \left[ \eta_k L \prod_{l=1}^{L} \prod_{a=1}^{A_l} p_{kla}^{n_{ila}} \right] V_{ik}.
\]  (B.1)

For iteration \( t + 1 \), the E-step needs the conditional expectation of matrix \( V = [V_{ik}]_{I \times K} \), given the current parameter estimate \( \Theta^{(t)} \) and the observed data \( X \). Let \( v_{ik}^{(t)} = \mathbb{E}[V_{ik} \mid \Theta^{(t)}, X] \). Then,

\[
v_{ik}^{(t)} = \frac{\eta_k^{(t)} \prod_{l=1}^{L} \prod_{a=1}^{A_l} (p_{kla}^{(t)})^{n_{ila}}}{\sum_{j=1}^{K} \eta_j^{(t)} \prod_{l=1}^{L} \prod_{a=1}^{A_l} (p_{jla}^{(t)})^{n_{ila}}}. \]  (B.2)
The M-step maximizes the conditional expectation of the complete-data log likelihood, producing
\[ \eta_{ik}^{(t+1)} = \frac{\sum_{i=1}^{I} v_{ik}^{(t)}}{I} \quad \text{and} \quad p_{kla}^{(t+1)} = \frac{\sum_{i=1}^{I} v_{ik}^{(t)} n_{ila}}{M \sum_{j=1}^{I} v_{ij}^{(t)}}. \]

The derivations follow from the standard EM algorithm for mixture models (Fraley and Raftery, 2002).

**Admixture Model**

If we pretend to know \( D_{ilmk} \), which indicates if the \( m \)th allele at locus \( l \) in individual \( i \) comes from the \( k \)th subpopulation, then the complete-data likelihood can be written as
\[
L_c(\Theta, D | X) = \prod_{i=1}^{I} \prod_{l=1}^{L} \prod_{m=1}^{M} \prod_{k=1}^{K} \prod_{a=1}^{A_l} (\eta_{ik} p_{kla})^{I(x_{ilm}=a)D_{ilmk}}.
\]

Here, we introduce indicator function \( I(\cdot) \), which takes value 1 when its argument is true and is otherwise 0. When \( x_{ilm} = a \), we compute
\[
d_{ilk}^{(t)}(a) = \mathbb{E}[D_{ilmk} | \Theta^{(t)}, X] = \frac{\eta_{ik}^{(t)} p_{kla}^{(t)}}{\sum_{j=1}^{K} \eta_{ij}^{(t)} p_{jla}^{(t)}}
\]
by Bayes’ rule. Since this expression is independent of \( m \), we can rearrange sums in the expected complete-data log likelihood to get
\[
\mathbb{E}[\log L_c | \Theta^{(t)}, X] = \sum_{i=1}^{I} \sum_{l=1}^{L} \sum_{k=1}^{K} \sum_{a=1}^{A_l} n_{ila} d_{ilk}^{(t)}(a) (\log \eta_{ik} + \log p_{kla}). \tag{B.4}
\]

If we let \( n_{ikla}^{(t)} = n_{ila} d_{ilk}^{(t)}(a) \) be the expected number of \( a \) alleles at locus \( l \) in individual \( i \) descendent from group \( k \) at the \( t \)th iteration, then maximization of \( \text{(B.4)} \) yields
\[
\eta_{ik}^{(t+1)} = \frac{\sum_{l=1}^{L} \sum_{a=1}^{A_l} n_{ikla}^{(t)}}{\sum_{k=1}^{K} \sum_{l=1}^{L} \sum_{a=1}^{A_l} n_{ikla}^{(t)}} \quad \text{and} \quad p_{kla}^{(t+1)} = \frac{\sum_{i=1}^{I} n_{ikla}^{(t)}}{\sum_{i=1}^{I} \sum_{a=1}^{A_l} n_{ikla}^{(t)}}. \tag{B.5}
\]

**Missing Values**

Genetic data are replete with missing information. As Smouse et al. (1990) does for the mixture model and Tang et al. (2005) does for the admixture model, we assume the missing data is missing completely at random and is ignorable in Rubin’s sense (Rubin, 1976).
genetic data can be split into observed and missing parts, $X = \{ X^{\text{obs}}, X^{\text{mis}} \}$. The likelihood function becomes

$$L(\Theta \mid X^{\text{obs}}) = \sum_{X^{\text{mis}}} L(\Theta, X^{\text{mis}} \mid X^{\text{obs}}).$$

In the E-step of EM algorithm, we need to calculate the expectation over the missing alleles as well as the unknown population assignments.

**Mixture Model with Missing Data**

We will typically signal missing data by encoding missing $x_{ilm}$ with a special value, such as $−9$, $0$ or ‘‘−‘’. Then, the observed data likelihood is unchanged from (3.1), if we understand $p_{kl}^{(t)}$, for example, to be $1$. The complete-data likelihood includes random variables $V_{ik}$ and $I(x_{ilm} = a)$ when $x_{ilm}$ is missing, so the expected complete-data log likelihood needed for the E-step is

$$E \left[ l_{c}(\Theta \mid V, X) \mid X^{\text{obs}}, \Theta^{(t)} \right] = \sum_{i=1}^{I} \sum_{l=1}^{L} \sum_{k=1}^{K} \sum_{a=1}^{A_l} \left( n_{ila} + n_{i0}^{(t)} p_{kla} \right) \log p_{kla},$$

where $v_{ik}^{(t)}$ is as in the case with no missing data, and $n_{i0}$ is the number of missing alleles in individual $i$ at locus $l$. We have used the independence of alleles within an individual, so conditional probability $P(x_{ilm} = a \mid V_{ik} = 1, X^{\text{obs}}, \Theta^{(t)})$ is $p_{kla}^{(t)}$. The M-step yields the very same update for $\eta_{ik}^{(t+1)}$, but now

$$p_{kla}^{(t+1)} = \frac{\sum_{i=1}^{I} v_{ik}^{(t)} (n_{ila} + n_{i0}^{(t)} p_{kla})}{M \sum_{i=1}^{I} v_{ik}^{(t)}}.$$

**Admixture Model with Missing Data**

The expected complete-data log likelihood for the admixture model is

$$\sum_{i=1}^{I} \sum_{l=1}^{L} \sum_{k=1}^{K} \sum_{a=1}^{A_l} \left( n_{ila} d_{ilk}^{(t)}(a) + n_{i0} \eta_{ik}^{(t)} p_{kla}^{(t)} \right) \left( \log \eta_{ik} + \log p_{kla} \right).$$

This equation is derived by recognizing that $E \left[ D_{ilmk} I(x_{ilm} = a) \mid X^{\text{obs}}, \Theta^{(t)} \right]$ is the previously computed $d_{ilk}^{(t)}(a)$ of Eq.(B.3) when $x_{ilm} = a$ is observed and $\eta_{ik}^{(t)} p_{kla}^{(t)}$ when $x_{ilm}$ is missing. The M-step equations(B.5) are valid if we define $n_{i0}^{(t)} = n_{ila} d_{ilk}(a) + n_{i0} \eta_{ik}^{(t)} p_{kla}^{(t)}$. 


Quadratic Programming Updates

Under the admixture model specified by MULTICLUST, we define a set of observed alleles \( \{a_1, a_2, \ldots, a_m\} \) at a locus, \( l \in L \), where \( L \) is the set of genotyped loci at \( I \) individuals. The frequency of observing an allele \( a \) at locus \( l \) in a subpopulation \( k \) is denoted by \( p_{kla} \). Let us define the number of alleles of type \( a \) at a locus \( l \) in an individual \( i \) as \( n_{ila} = \sum_{m=1}^{M} I(x_{ilm} = a) \), where \( I(x_{ilm} = a) \) is an indicator function, and it is equal to 1 if the \( m^{th} \) allele at locus \( l \) in individual \( i \) is \( a \). Also, note that \( n_{ila} \in \{0, 1, 2, \ldots, P\} \) where \( P \) is the ploidy level, or the maximum number of alleles at a locus. Let us denote the admixture proportions by \( \eta_{ik} \), which is the proportion of individual \( i \) that is derived from an ancestral subpopulation \( k \). Hence the logarithmic observed data likelihood, given parameter set \( \theta = \{\eta_{ik}, p_{kla}\} \) can be defined as:

\[
L(\theta|X) = \sum_{i=1}^{I} \sum_{l=1}^{L} \sum_{p=1}^{P} \left[ n_{ila} \ln \left( \sum_{k=1}^{K} \eta_{ik} p_{kla} \right) \right]
\]

Under the EM algorithm, updates are defined as before. But in order to use the quadratic programming method of Alexander et al. (2009) we require defining the first order partial derivatives of the logarithmic data likelihood, with respect to the parameters, \( \theta = \{\eta_{ik}, p_{kla}\} \):

\[
\frac{\partial L}{\partial \eta_{ik}} = \sum_{l=1}^{L} \sum_{p=1}^{P} \left[ \frac{n_{ila} \cdot p_{kla}}{\sum_{k=1}^{K} \eta_{ik} p_{kla}} \right]
\]

\[
\frac{\partial L}{\partial p_{kla}} = \sum_{i=1}^{I} \sum_{l=1}^{L} \sum_{p=1}^{P} \left[ \frac{n_{ila} \cdot \eta_{ik}}{\sum_{k=1}^{K} \eta_{ik} p_{kla}} \right]
\]

The second order partial derivatives of the logarithmic data likelihood with respect to the parameters \( \theta = \{\eta_{ik}, p_{kla}\} \) can be defined as:

\[
\frac{\partial L^2}{\partial \eta_{ik} \partial \eta_{jk}} = -\sum_{l=1}^{L} \sum_{p=1}^{P} \left[ \frac{n_{ila} \cdot p_{kla} \cdot p_{jla}}{\left( \sum_{k=1}^{K} \eta_{ik} p_{kla} \right)^2} \right]
\]

\[
\frac{\partial L^2}{\partial p_{kla} \partial p_{jla}} = -\sum_{i=1}^{I} \sum_{l=1}^{L} \sum_{p=1}^{P} \left[ \frac{n_{ila} \cdot \eta_{ik} \cdot \eta_{j}}{\left( \sum_{k=1}^{K} \eta_{ik} p_{kla} \right)^2} \right]
\]

The problem of solving for the most likely parameter space can also be specified now as a sequential quadratic programming problem, as utilized in Alexander et al. (2009). For function \( f(x) \), the quadratic programming problem can be written as a minimization:
\[
\text{minimize } \frac{1}{2} x^T H x + x^T f \\
\text{subject to } x^T a = b, l_i \leq x_i \leq u_i, i = 1, \ldots, n
\]

where \( x = (x_1, x_2, \ldots, x_k) \) is the vector of \( K \) (number of ancestral subpopulations) parameters to be optimized, \( H \) is a \( K \times K \) Hessian matrix (symmetric, positive, semi-definite) comprised of the second order partial derivatives above, while updating the \( \eta_{ik} \) blocks, and is a \( P \times P \) Hessian matrix while updating \( p_{kla} \)'s. \( f \) is the Jacobian, real vector of size \( K \) (while updating \( \eta_{ik} \)), and size \( P \) (while updating \( p_{kla} \), comprised of the first order partial derivatives from the above step, \( b \) is a scalar, and is 0, \((l_1, l_2, \ldots, l_k)\) and \((u_1, u_2, \ldots, u_k)\) are the real lower and upper bounds on the parameter set, \( x \). Recall that the parameter set being currently optimized could be either the ancestral allele frequencies \( p_{kla} \) or the admixture proportions, \( \eta_{ik} \), depending on the current run of the block relaxation (in conjunction with EM) algorithm.

This QP problem can be solved by LibQP (http://cmp.felk.cvut.cz/xfrancv/libqp/html/).

We will now address details of the optimization problem for a modified admixture model than for SNP's (Single Nucleotide Polymorphisms), as proposed by Falush et al. (2003), Pritchard et al. (2000b), and improved upon by Alexander et al. (2009). This model allows for infinite allelic states, equivalent to microsatellite or SSR markers, AFLP’s, etc. that are commonly utilized in population genetics projects. The new accelerated optimization algorithm operates via a block update relaxation method - after performing an initial estimate of both parameters (mixing proportions and allele frequencies), we switch over to the block relaxation method, that alternates between updates of either parameter. Roughly, the algorithm for optimization proceeds in the following steps: (1) Perform initial runs of the EM algorithm (as described previously), and obtain initial estimates of parameters, \( \theta = \{\eta_{ik}, p_{kla}\} \), (2) Perform the block relaxation update on one of the parameters, until the logarithmic likelihood does not improve, by using Newton’s method, and (3) If the likelihood is no longer improving, then switch to update the other parameter, and repeat.

Let us define the estimates obtained in each iteration \( q \) (of either EM algorithm, or the block update relaxation algorithm) as \( X_q \). So during the current iteration, \( q (q = 0, 1, \ldots, and \)
\( q = 0 \) for the initialization step, \( X_q = (\eta_{i1}, \eta_{i2}, \ldots, \eta_{ik}) \), where \( i \) indexes individuals, and \( k \) indexes ancestral subpopulations. Alternately, if we are optimizing the allele frequencies in the current iteration, \( q \), then \( X_q = (p_{kl1}, p_{kl2}, \ldots, p_{klm}) \), where \( m \) indexes all observed alleles at locus \( l \) in subpopulation \( k \). We also define the \( U \) and \( V \) column vectors as \( U_1 = X_1 - X_0 \) and \( V_1 = X_2 - X_1 \).

We could thereon use a Newton’s method update proposed by Alexander et al. (2009) or a quasi-Newton method proposed by Zhou et al. (2011) both of which use the above defined vectors, \( U \) and \( V \) from previous iterates to perform updates. The previously defined EM algorithm is encoded as a fallback, in cases where updating using the quasi-Newton method does not improve the logarithmic likelihood of the data. The ascent property of the EM algorithm is key to this step, since it ensures that the likelihood is always increasing. Thereon, after performing a couple of EM runs, we proceed to update using the Newton’s or quasi-Newton’s method, and so on.

The Newton’s update obtained by solving the second order Taylor series expansion, and substituting an approximation for \( M \), as proposed by Alexander et al. (2009) can be defined as:

\[
X_{n+1} = X_n - (I - M)^{-1} \frac{\partial L}{\partial X_n}
\]

where

\[
(I - M)^{-1} = I + V[U^tU - U^tV]^{-1}U^t
\]

Here, \( X_n \) is the set of estimated parameters, \( \theta \) (either \( \eta_{ik} \) or \( p_{kl} \)) in the current iteration, \( n \).

Similarly, the quasi-Newton update as proposed by Zhou et al. (2011) can be defined as:

\[
X_{n+1} = X_n - 2SU + S^2(V - U)
\]

where

\[
S = -\sqrt{\frac{U^tU}{(V - U)^t(V - U)}}
\]
Algorithm 1: BlockRelaxation-Zhou

\begin{algorithm}
begin
repeat
\begin{algorithmic}
\State Randomly pick to update either $\eta_{ik}$ or $p_{kla}$.
\State Run EM twice to get $\eta_{ik}$ and $p_{kla}$, and save estimates for all alleles and individuals.
\State Set $q = 0$
\If {$\eta_{ik}$ was picked}
\State Set $y = 0$ and $t = I$
\ElseIf {$p_{kla}$ was picked}
\State Set $y = 1$ and $t = L$
\EndIf
\For {$x \in t$}
\State Set $z = 0$ and $q = 0$
\repeat
\State Calculate log likelihood $L_1$
\If {$q == 0$ \&\& $y == 0$}
\State Set $X_0 = \{\eta_{i1}, \eta_{i2}, \ldots, \eta_{ik}\}$ from initialization
\State Set $X_1 = \{\eta_{i1}, \eta_{i2}, \ldots, \eta_{ik}\}$ from iteration 1 of EM
\State Set $X_2 = \{\eta_{i1}, \eta_{i2}, \ldots, \eta_{ik}\}$ from iteration 2 of EM
\ElseIf {$q == 0$ \&\& $y == 1$}
\State Set $X_0 = \{p_{kl1}, p_{kl2}, \ldots, p_{klm}\}$ from initialization
\State Set $X_1 = \{p_{kl1}, p_{kl2}, \ldots, p_{klm}\}$ from iteration 1 of EM
\State Set $X_2 = \{p_{kl1}, p_{kl2}, \ldots, p_{klm}\}$ from iteration 2 of EM
\EndIf
\State Calculate $U = X_1 - X_0$ and $V = X_2 - X_1$
\State Quasi Newton Update to get $X_{n+1}$
\State Calculate log likelihood $L_2$
\If {$L_2 > L_1$}
\State Set $X_z = X_{n+1}$
\State $z = z + 1$
\EndIf
\If {$z == 2$}
\State Set $z = 0$
\EndIf
\Until {$L_1 > L_2$}
\Until {$L_2 - L_1 < 10^{-4}$}
\end{algorithmic}
end
\end{algorithm}
APPENDIX C. APPENDIX TO ESTIMATING RELATEDNESS USING ADMIXTURE PROPORTIONS IN STRUCTURED POPULATIONS

EM Algorithm

An alternate to obtaining maximum likelihood estimates of relatedness coefficients is proposed here, through the Expectation-Maximization (EM) procedure of Dempster et al. (1977). The likelihood of the relatedness between a pair of individuals (here conditioned on observed IBS states), can be written, similar to that of Thomson (1977) as

\[ L(\Delta) = \sum_{j=1,2,\ldots,9} P(S_i | D_j) \Delta_j. \]

This is an example of a mixture model, with the likelihood specified in terms of the mixing proportions (here the relatedness parameter vector, \( \Delta \)), and the parametric distribution, \( P(S_i | D_j) \) (Table 4.1). Estimation of the parameters can thereon proceed using an Expectation-Maximization algorithm (Dempster et al. (1977)), iteratively, in two steps. In the expectation step,

\[ V_{ij}^{(t+1)} = \frac{\Delta_j^{(t)} \prod_{l=1}^{L} P(x_l = s_i) \{n_x = s_i\} \{n_y = d_i\}}{\sum_{j'=1}^{9} \Delta_{j'}^{(t)} \prod_{l=1}^{L} P(x_l = s_i) \{n_x = s_i\} \{n_y = d_i\}} \quad (C.1) \]

Where \( V_{ij} \) is the unobserved distribution that a genotype in IBS state \( i \) is in IBD state \( j \), \( P(x = s_i) \) is the probability of observing a genotype at a locus \( l \) in IBS state \( i \), and \( \{n_x = s_i\} \), and \( \{n_y = d_i\} \) are indicator functions which are \( 1 \) if the genotype at locus \( l \) is in IBS state \( i \), and in IBD state \( j \).

In the maximization step, we estimate:

\[ \Delta_j^{(t+1)} = \frac{\sum_{i=1}^{9} V_{ij}^{(t)}}{9} \text{ and } P(x = s_i)^{(t+1)} = \frac{\sum_{i=1}^{9} V_{ij}^{(t)} n_x = s_i}{\sum_{i=1}^{9} V_{ij}^{(t)}} \]
Other Relatedness Estimators

MULTICLUST and similar methods produce a vector of posterior probabilities that the $m$th allele at locus $l$ in individual $x$ can be defined as $V_{xlm} = \{K_{xlm1}, K_{xlm2}, \ldots, K_{xlmK}\}$, and denotes the probability that the allele at locus $l$ in individual $x$ was derived from a certain subpopulation, $k$. We know that $\sum_{k}^{K} K_{xlmk} = 1$. We define the centroid vector of size $K$ at a locus $l$ in a diploid individual $x$ as

$$
\bar{V}_{xl} = (\bar{K}_{xl1}, \bar{K}_{xl2}, \ldots, \bar{K}_{xlk})
= \left( \frac{(K_{xl11} + K_{xl21})}{2} + \frac{(K_{xl12} + K_{xl22})}{2} + \ldots + \frac{(K_{xl1k} + K_{xl2k})}{2} \right)
$$

For each individual $x$ and $y$, at a locus $l$, we define the Euclidean distance between $x$ and $y$ as the distance between the two vectors, $\bar{V}_{xl}$ and $\bar{V}_{yl}$,

$$
\hat{r}_{lxy} = \sqrt{\sum_{k=1}^{K} (\bar{K}_{ylk} - \bar{K}_{xlk})^2}
$$

where $\bar{K}_{ylk}$ and $\bar{K}_{xlk}$ are elements of the centroid vectors $\bar{V}_{yl}$ and $\bar{V}_{xl}$ respectively, as defined above.

I now redefine three estimators of pairwise relatedness (Queller and Goodnight (QG) estimator((Queller and Goodnight, 1989)), Lynch and Ritland estimator ((Lynch and Ritland, 1999)) (LR), and Wang (W) estimator((Wang, 2002))) in terms of admixture proportions, and probabilities that we have obtained above.

The QG estimator ((Queller and Goodnight, 1989)) borrows from the works of Harpending (1979), and defines the relatedness coefficient between individuals in a subdivided population as (with notation changed for consistency, and modified to estimate pairwise relatedness):

$$
\hat{r}_{lxy}^{(1)} = \frac{\sum_{k=1}^{K} \sum_{l=1}^{L} (p_{ylm} - \bar{p}_{kl})}{\sum_{k=1}^{K} \sum_{l=1}^{L} (p_{xlm} - \bar{p}_{kl})}
$$

where $l$ indexes locus, $k$ indexes the subpopulation. The $p_{ylm}$ and $p_{xlm}$ are frequencies of an allele at allelic position (haplotype) $m$, in individual $y$ and $x$ respectively, and are equal to 1 if the individual is homozygous, 0.5 if heterozygous, and 0 if the allele is absent at position $m$. $\bar{p}_{kl}$ is the mean frequency of the allele in position $m$ at a locus $l$, in ancestral subpopulation $k$, and
and is estimated as a parameter in several methods described above (Pritchard et al., 2000b), (Liu et al., 2013)). This \( \hat{r}_{xy}^{(1)} \) is estimated with individual \( x \) as reference to individual \( y \). To make this estimator symmetric, \( \hat{r}_{xy}^{(2)} \) is calculated with \( y \) as reference to individual \( x \), and the average of the two is denoted as the Queller and Goodnight relatedness.

\[
\hat{r}_{xy} = \frac{\hat{r}_{xy}^{(1)} + \hat{r}_{xy}^{(2)}}{2} \tag{C.3}
\]

Both the Lynch and Ritland ((Lynch and Ritland, 1999)) and Wang ((Wang, 2002)) estimators are defined in terms of ‘higher order coefficients’ (see Lynch and Ritland (1999)), which are essentially coined \( \phi_{XY} \), and \( \Delta_{XY} \) for two individuals \( X \) and \( Y \) respectively. These are defined as the probability that both alleles at a diploid, co-dominant locus in an individual \( X \) are identical by descent (IBD) with both alleles at the same locus in individual \( Y \) (\( \phi_{XY} \)), and the probability that one allele at a locus in an individual \( X \) is IBD with one allele at the same locus in individual \( Y \) (\( \Delta_{XY} \)). These can be defined easily in joining terms of the conditional IBD probabilities previously derived as in Table 1 (and see Fig. 4.1). Correspondingly,

\[
\phi_{XY} = P(S_3 \mid \Delta_3) + P(S_5 \mid \Delta_5) + P(S_8 \mid \Delta_8) \tag{C.4}
\]

\[
\Delta_{XY} = P(S_1 \mid \Delta_1) + P(S_7 \mid \Delta_7) \tag{C.5}
\]

Pairwise genetic relatedness between two individuals \( X \) and \( Y \) can be defined *sensu* Lynch and Ritland (1999) as

\[
r_{XY} = \frac{\phi_{XY}}{2} + \Delta_{XY} \tag{C.6}
\]

The previously solved likelihood equation (and \( \Delta \)’s) can be used to estimate \( \phi_{XY} \), \( \Delta_{XY} \), and \( r_{XY} \) above. This would offer an alternate parametrization (using subpopulation allele frequencies and admixture proportions) for the estimators of Lynch and Ritland (1999) and Wang (2002).

Now I have four estimators of ancestral relatedness, that I can test to determine how they perform under different evolutionairy scenarios to infer relatedness.


Frankham, R., J. D. Ballou and D. A. Briscoe. 2002. Introduction to conservation genetics. Introduction to Conservation Genetics i. 0-521-63014-2 (cloth); 0-521-63985-9 (paper).


Lawson, Daniel John Falush, Daniel Chakravarti, A Green, E 978-0-8243-3713-1.


Li, Qizhai Yu, Kai.


ACKNOWLEDGEMENTS

Working on this dissertation has been a very rewarding experience for me, and none of this work would have been possible without the ample encouragement and support of both my advisors, Fred Janzen, and Karin Dorman. Both Fred and Karin have been extremely patient through my trials and quirks, and have provided me with timely and constructive feedback and appraisal throughout the length of my doctoral pursuit. They have both been very supportive of my persistence forward, and I could not thank them enough for this. I have learnt a lot from both their individual styles of mentoring students, and I will continue to strive to become as good of a researcher.

I thank my committee members, Anne Bronikowski, John Nason, and David Fernandez-Baca for their time, and support through this process. I have had several very insightful discussions with all of them, which have contributed in their own ways to the formulation of this thesis, and my development as a population geneticist, from a computer scientist.

A very special thank you goes to Trish Stauble, and the Bioinformatics and Computational Biology (BCB) program at Iowa State University. Trish has been one of the most important figures in my graduate life here at ISU, helping me throughout the process of application, to graduation. The constant support of the BCB program and its factions has been overwhelming, and I am very grateful for it.

A lot of my good friends and mentors have constantly encouraged me to push forward, without whose help and support, this work would not have been possible. I thank Aswini Sivaraman, Zachary Strohman, Karri Haen-Whitmer, Tonia Schwartz, Wei-Chen Chen, Vijay Kanagala, Tim Mitchell, John Obrycki, Amy Toth, Morgan Becker, and all members of the ‘Janzone’, the HHMI-ISU GTALC, the Bronikowski Lab, BCB Lab, and the Indian Students’ Association of ISU.

And finally, none of this work would have been possible without the financial support of the
BCB program, the Departments of Ecology, Evolution, and Organismal Biology (EEOB), and
Genetics, Developmental, and Cellular Biology (GDCB), the James Cornette Fellowship, Iowa
State University’s Office of Biotechnology, Sigma Xi Grants in Aid of Research Program, Howard
Hughes Medical Institute-Iowa State University Teaching Fellowships, Anne Bronikowski, and
Fred Janzen.

I dedicate this dissertation to my dearest parents, Lalitha Sethuraman and V. Sethuraman,
who have never said no.