

# Genetic variation and modern human origins

by

Michael DeGiorgio

A dissertation submitted in partial fulfillment  
of the requirements for the degree of  
Doctor of Philosophy  
(Bioinformatics)  
in the University of Michigan  
2011

Doctoral Committee:

Associate Professor Noah A. Rosenberg, Chair  
Professor Michael L. Boehnke  
Assistant Professor Jun Li  
Assistant Professor Patricia Wittkopp  
Assistant Professor Sebastian K. Zöllner

© Michael DeGiorgio 2011  

---

All Rights Reserved

To my parents, Meryl and Sal

## ACKNOWLEDGEMENTS

The road to obtaining a doctoral degree is full of obstacles and I would like to extend my gratitude to all the individuals who have helped me accomplish my goals. In the following I will acknowledge, and detail the contributions of, the individuals who have had the greatest influence on my path to becoming who I am now.

I would like to begin by extending my intense gratitude to my advisor Dr. Noah Rosenberg. My relationship with Noah began in April 2007 when I started as a rotation student in his lab. From that date, Noah has helped mold me into a scientist through his high expectations, attention to detail, mathematical rigor, and careful and unbelievably patient reviewing and editing of manuscripts. Noah has always kept an open door, allowing for discussion of personal issues as well as issues relevant to academic success. He has provided someone for me to look up to.

I would also like to thank the other members of my dissertation committee Drs. Mike Boehnke, Jun Li, Trisha Wittkopp, and Sebastian Zöllner for their valuable help and guidance. I am also grateful for their commitment to helping me obtain a post-doctoral position.

I would like to acknowledge my friend and colleague Dr. James Degnan. I have interacted with James since I first joined Noah's lab and, ever since, he has been like a second mentor to me.

I would like to thank the members of the Rosenberg lab for always being available to give me scientific advice and for providing friendly faces to discuss personal issues. In particular, I would like to acknowledge my cubicle mates Lucy Huang and Zach

Szpiech for being there for both emotional and academic support each weekday. In addition, I would like to thank Dr. Trevor Pemberton for always cooking and baking delicious items so that I never become thin. I would also like to give a very special thanks to my amazing student Ivana Jankovic who helped me appreciate both the difficulties and pleasures of mentoring and teaching. Ivana contributed equally to the work in Chapter III of this dissertation and, therefore, has had a significant impact on my path to finishing a doctoral degree.

I would next like to acknowledge the individuals who had a large influence on the path that I have taken to a doctoral degree in bioinformatics. I would like to thank Drs. Raquel Assis and Mark Choi for encouraging me to move to biology and providing me with the necessary resources to make the transition. I thank Arup Guha for sparking my interest in research and encouraging me to apply to graduate school. I would like to acknowledge Dr. Roger Goldwyn for piquing my interest in mathematics. I am deeply grateful to Alan Lucas who introduced me to, and patiently and enthusiastically mentored me in, computer science while I was in high school.

I would like to thank my parents Sal and Meryl DeGiorgio for giving me everything they could. They always taught me to strive to reach my goals and dreams and they provided as much emotional and financial support as they could along the way. My parents have nurtured every interest and passion that I have and had. They have also supported every transition that I have made in my path to becoming an adult. Without their unconditional love and support, I would never have become the person I am today.

I would also like to extend my intense gratitude to Alan and Janus Lucas, Mario and Georgina Pigna, and Dr. Randy and Rekha Stein. They are like second parents to me and have always welcomed me into their homes and provided me with a comfortable place to hang out.

I would like to thank my lifelong friends Mark Choi, Ryan “Raindrop” Deaunovich,

Tim Emanuel, Adam Lucas, Kris Pigna, and Maha Stein. They were an integral part in molding who I am today.

I would like to thank my extended family for their influence on who I am. In particular, I would like to acknowledge my grandmothers Nellie DeGiorgio and Carmen Rodriguez, my uncles Frank Charon and Dominick DeGiorgio, my aunt Doreen Charon, and my cousin Tony Valle.

A very special thanks goes to my longtime girlfriend Raquel Assis for supporting me during every step of my career. Raquel nurtured my interest in applying mathematics to biology, helped me grow as a better writer and scientist, has accepted and helped me improve upon my weaknesses and faults, and has been supportive in every decision I have made. I am indebted to her beyond what words can express.

Finally, I would like to acknowledge the contributions of all coauthors of Chapters II-VIII of this dissertation. Chapters that an individual is a coauthor on are indicated in parentheses to the right of the individual's name. The coauthors and their respective contributions are Richard Cronn (VIII), James Degnan (V and VI), Andrew Eckert (VIII), Mattias Jakobsson (IV), Ivana Jankovic (III, equal contribution), Aaron Liston (VIII), David Neale (VIII), Noah Rosenberg (II-V, VII, and VIII), and John Syring (VIII).

# TABLE OF CONTENTS

DEDICATION . . . . .	ii
ACKNOWLEDGEMENTS . . . . .	iii
LIST OF FIGURES . . . . .	x
LIST OF TABLES . . . . .	xvii
CHAPTER	
<b>I. Introduction . . . . .</b>	<b>1</b>
<b>II. An Unbiased Estimator of Gene Diversity in Samples         Containing Related Individuals . . . . .</b>	<b>10</b>
2.1 Introduction . . . . .	10
2.2 Theory . . . . .	12
2.3 Data from Human Populations . . . . .	17
2.4 Simulations . . . . .	18
2.4.1 Simulation Procedure . . . . .	18
2.4.2 Simulation Results . . . . .	19
2.5 Application to Data . . . . .	22
2.5.1 Notation . . . . .	22
2.5.2 Mean of the Estimator . . . . .	22
2.5.3 Gene Diversity vs. Distance from Africa . . . . .	24
2.6 Discussion . . . . .	25
2.7 Acknowledgments . . . . .	27
<b>III. Unbiased estimation of gene diversity in samples containing         related individuals: exact variance and arbitrary ploidy . . . . .</b>	<b>41</b>
3.1 Introduction . . . . .	41
3.2 Theory . . . . .	43

3.2.1	An unbiased estimator . . . . .	44
3.2.2	Variance of the estimator . . . . .	49
3.2.3	The X chromosome case . . . . .	51
3.3	Data analysis . . . . .	53
3.3.1	Data . . . . .	53
3.3.2	Data analysis methods . . . . .	54
3.3.3	Effect of parameters on the estimators . . . . .	56
3.3.4	Application to data . . . . .	59
3.4	Discussion . . . . .	61
3.5	Acknowledgments . . . . .	62
3.6	Appendix A . . . . .	73
3.7	Appendix B . . . . .	76

**IV. Explaining worldwide patterns of human genetic variation using a coalescent-based serial founder model of migration outward from Africa . . . . . 88**

4.1	Introduction . . . . .	88
4.2	Results . . . . .	90
4.2.1	Overview of models . . . . .	90
4.2.2	Simulations . . . . .	91
4.2.3	Basic model . . . . .	92
4.2.4	Archaic admixture . . . . .	94
4.2.5	Archaic persistence model . . . . .	95
4.2.6	Instantaneous divergence model . . . . .	97
4.3	Discussion . . . . .	99
4.4	Materials and Methods . . . . .	101
4.4.1	Heterozygosity . . . . .	101
4.4.2	Linkage disequilibrium . . . . .	102
4.4.3	Ancestral allele frequency spectrum . . . . .	102
4.5	Acknowledgments . . . . .	103

**V. Coalescence-time distributions in a serial founder model of human evolutionary history . . . . . 115**

5.1	Introduction . . . . .	115
5.2	Serial founder model . . . . .	118
5.2.1	Model . . . . .	118
5.2.2	Coalescence times . . . . .	119
5.2.3	Pairwise homozygosity and heterozygosity . . . . .	123
5.2.4	Pairwise $F_{ST}$ . . . . .	125
5.3	Patterns observed in human population data . . . . .	125
5.4	Modern serial founder model . . . . .	127
5.4.1	Motivation and model . . . . .	127
5.4.2	Patterns generated by the model . . . . .	128



5.5	Nested regions model . . . . .	130
5.5.1	Motivation and model . . . . .	130
5.5.2	Patterns generated by the model . . . . .	130
5.6	Instantaneous divergence model . . . . .	131
5.6.1	Motivation and model . . . . .	131
5.6.2	Patterns generated by the model . . . . .	133
5.7	Archaic serial founder model . . . . .	134
5.7.1	Motivation and model . . . . .	134
5.7.2	Patterns generated by the model . . . . .	135
5.8	Discussion . . . . .	136
5.9	Acknowledgments . . . . .	141

**VI. Fast and consistent estimation of species trees using supermatrix rooted triples . . . . . 151**

6.1	Introduction . . . . .	151
6.2	Methods . . . . .	154
6.2.1	Supermatrix rooted triple (SMRT) . . . . .	154
6.2.2	Simulation . . . . .	155
6.2.3	Empirical example . . . . .	156
6.3	Results for simulations . . . . .	157
6.3.1	Four taxa . . . . .	157
6.3.2	Five taxa . . . . .	158
6.3.3	Model violations . . . . .	159
6.4	Results for yeast data . . . . .	161
6.5	Theory . . . . .	163
6.6	Discussion . . . . .	173
6.6.1	Overview of results and implications . . . . .	173
6.6.2	Taxon sampling for species tree inference . . . . .	175
6.6.3	Rooted triple consensus . . . . .	176
6.6.4	Bayesian approaches . . . . .	178
6.6.5	Other sources of discordance . . . . .	178
6.6.6	Summary . . . . .	179
6.7	Acknowledgments . . . . .	180

**VII. Consistency of phylogenetic consensus methods in the presence of ancestral population structure . . . . . 205**

7.1	Introduction . . . . .	205
7.1.1	Model . . . . .	207
7.1.2	Counterexample . . . . .	209
7.2	Consistency and inconsistency of methods . . . . .	215
7.2.1	Uniquely favored topology . . . . .	216
7.2.2	Average coalescence times . . . . .	218
7.2.3	Average ranks of coalescences . . . . .	219

7.2.4	Uniquely favored rooted triples . . . . .	221
7.2.5	Minimizing deep coalescences . . . . .	222
7.2.6	Majority-rule . . . . .	224
7.2.7	Minimum coalescence time . . . . .	224
7.3	Simulations . . . . .	227
7.3.1	Performance of methods on true gene trees . . . . .	227
7.3.2	GLASS/Maximum Tree from inferred gene trees . . . . .	230
7.4	Discussion . . . . .	232
7.5	Acknowledgments . . . . .	233
7.6	Appendix . . . . .	242
7.6.1	Average coalescence times . . . . .	242
7.6.2	Average ranks of coalescences . . . . .	244
7.6.3	Uniquely favored rooted triples . . . . .	245
7.6.4	Minimizing deep coalescences . . . . .	246
7.6.5	Majority-rule . . . . .	247
<b>VIII.</b>	<b>An empirical evaluation of species tree inference strategies</b>	
	<b>using a multilocus dataset from North American pines . . . . .</b>	<b>248</b>
8.1	Introduction . . . . .	248
8.2	Methods . . . . .	251
8.2.1	North American white pine dataset . . . . .	251
8.2.2	Overview of the analysis . . . . .	252
8.2.3	Creating datasets . . . . .	253
8.2.4	Inferring gene trees . . . . .	257
8.2.5	Inferring species trees . . . . .	257
8.2.6	Multivariate analysis . . . . .	260
8.3	Results . . . . .	262
8.3.1	Clade size . . . . .	263
8.3.2	Clustering of strategies . . . . .	264
8.3.3	Clade flow . . . . .	266
8.3.4	Representative topologies . . . . .	268
8.4	Discussion . . . . .	270
8.5	Acknowledgments . . . . .	272
<b>IX.</b>	<b>Conclusion . . . . .</b>	<b>289</b>
<b>BIBLIOGRAPHY . . . . .</b>		<b>293</b>

## LIST OF FIGURES

### Figure

1.1	Processes affecting both genetic and phenotypic variation. . . . .	8
1.2	Simple models representing hypotheses for modern human origins. .	9
2.1	Mean squared error (MSE) as a function of sample size $m$ for three different estimators. . . . .	34
2.2	Heat maps of simulated mean squared error (MSE), variance, and bias squared for each estimator applied to a full sample of 40 and a reduced sample of 20 individuals, as functions of the mixture of types of relative pairs included in the sample. . . . .	35
2.3	Heat maps of simulated mean squared error (MSE), variance, and bias squared for each estimator applied to a full sample of 40 and a reduced sample of 20 individuals, as functions of the mixture of types of relative pairs included in the sample. . . . .	36
2.4	Mean squared error (MSE), variance, and bias squared for each estimator applied to a full sample of 30 and a reduced sample of 15 individuals, as functions of parametric gene diversity, considering simulated values based on each of the 783 loci. . . . .	37
2.5	Comparison of the mean of $\widehat{H}_{1048} - \widehat{H}_{952}$ and the mean of $\widetilde{H}_{1048} - \widetilde{H}_{952}$ .	38
2.6	Comparison of the mean difference of an estimator ( $\widehat{H}_{603}$ or $\widetilde{H}_{603}$ ) from $\widehat{H}_{507}$ with the standard deviation of the estimator. . . . .	39
2.7	Gene diversity vs. geographic distance from Addis Ababa, Ethiopia.	40
3.1	Mean squared error, variance, and bias squared for each estimator, obtained analytically using the variance approximation (eq. 3.18), as a function of heterozygosity for 36 loci. . . . .	67

3.2	Mean squared error as a function of sample size (number of pairs = number of individuals / 2) calculated analytically using the variance approximation (eq. 3.18) based on allele frequencies at the DXS1068 locus ( $H = 0.7344$ ). . . . .	68
3.3	Mean squared error as a function of sample size (number of pairs = number of individuals / 2) calculated analytically using the variance approximation (eq. 3.18), based on allele frequencies at the DXS1068 locus ( $H = 0.7344$ ) for male-female relative pairs in which the females were removed to calculate $\hat{H}_{reduced}$ . The range of each plot is truncated at 0.020. . . . .	69
3.4	Mean squared error (MSE), variance, and bias squared of $\hat{H}_{full}$ , $\tilde{H}_{full}$ , and $\hat{H}_{reduced}$ , calculated analytically using the variance approximation (eq. 3.18), as functions of the configuration of $t_1$ male-male ( $\Phi = 1/2$ ), $u_1$ male-female ( $\Phi = 1/2$ ), and $v_2$ female-female ( $\Phi = 3/8$ ) pairs in 20 total relative pairs, based on allele frequencies at the ATCT003 locus ( $H = 0.7794$ ). . . . .	70
3.5	Comparison of the difference between the mean of $\hat{H}_{485}$ across loci and the mean of $\hat{H}_{446}$ with the difference between the mean of $\tilde{H}_{485}$ and the mean of $\hat{H}_{446}$ . . . . .	71
3.6	Comparison of the difference between the mean of the estimator and the mean of $\hat{H}_{446}$ and standard deviation of the estimator, for the estimators $\tilde{H}_{485}$ and $\hat{H}_{485}$ . . . . .	72
3.7	Identity states. . . . .	87
4.1	Patterns of heterozygosity, LD, and the ancestral allele frequency spectrum observed in human population-genetic data. . . . .	104
4.2	Ancestral allele frequency spectra calculated using a resampling technique applied to the data of <i>Jakobsson et al.</i> (2008). . . . .	105
4.3	Models. . . . .	106
4.4	Patterns of heterozygosity, LD, and the ancestral allele frequency spectrum in simulations of the basic serial founder model. . . . .	107
4.5	Patterns of heterozygosity, LD, and the ancestral allele frequency spectrum in simulations of the serial founder model with symmetric migration at rate $M = 40$ between neighboring populations. . . . .	108

4.6	Patterns of heterozygosity, LD, and the ancestral allele frequency spectrum in simulations of the serial founder model with symmetric migration at rate $M = 1$ between neighboring populations. . . . .	109
4.7	Patterns of heterozygosity, LD, and the ancestral allele frequency spectrum in simulations of the serial founder model with archaic admixture ( $\gamma = 0.05$ ). . . . .	110
4.8	Patterns of heterozygosity, LD, and the ancestral allele frequency spectrum in simulations of the serial founder model with archaic admixture ( $\gamma = 0.1$ ). . . . .	111
4.9	LD ( $r^2$ ) as a function of physical distance for population 25 in the serial founder model with archaic admixture at rate $\gamma$ . . . . .	112
4.10	Patterns of heterozygosity, LD, and the ancestral allele frequency spectrum in simulations of the archaic persistence model. . . . .	113
4.11	Patterns of heterozygosity, LD, and the ancestral allele frequency spectrum in simulations of the instantaneous divergence model. . . . .	114
5.1	Serial founder model. . . . .	142
5.2	Distributions of coalescence times in the serial founder model. . . . .	143
5.3	Expected heterozygosity for a pair of lineages sampled from population 4 of Figure 5.2A (eq. 5.8), as a function of population size for bottlenecks and bottleneck length measured in generations. . . . .	144
5.4	Patterns of within- and between-population summary statistics observed in human population-genetic data. . . . .	145
5.5	Models to which the general serial founder model reduces. . . . .	146
5.6	Patterns of genetic variation in a modern serial founder model. . . . .	147
5.7	Patterns of genetic variation in a nested regions model. . . . .	148
5.8	Patterns of genetic variation in the instantaneous divergence model. . . . .	149
5.9	Patterns of genetic variation in an archaic serial founder model, as a function of varying divergence time $\tau_D$ . . . . .	150
6.1	Four- and five-taxon clocklike species tree topologies. . . . .	183

6.2	Schematic of our simulation procedure. . . . .	184
6.3	Results of simulations for the four-taxon tree $((AB)C)D$ (Figure 6.1A) generated under a Jukes-Cantor model with $\theta = 0.01$ and a molecular clock, and analyzed under maximum likelihood assuming a molecular clock and a Jukes-Cantor model. . . . .	185
6.4	Results of simulations for the four-taxon tree $((AB)(CD))$ (Figure 6.1B) generated under a Jukes-Cantor model with $\theta = 0.01$ and a molecular clock, and analyzed under maximum likelihood assuming a molecular clock and a Jukes-Cantor model. . . . .	186
6.5	Results of simulations for the five-taxon tree $((((AB)C)D)E)$ (Figure 6.1C) generated under a Jukes-Cantor model with $\theta = 0.01$ and a molecular clock, and analyzed under maximum likelihood assuming a molecular clock and a Jukes-Cantor model. . . . .	187
6.6	Results of simulations for the five-taxon tree $((((AB)C)(DE))$ (Figure 6.1D) generated under a Jukes-Cantor model with $\theta = 0.01$ and a molecular clock, and analyzed under maximum likelihood assuming a molecular clock and a Jukes-Cantor model. . . . .	188
6.7	Results of simulations for the five-taxon tree $((((AB)(CD))E)$ (Figure 6.1E) generated under a Jukes-Cantor model with $\theta = 0.01$ and a molecular clock, and analyzed under maximum likelihood assuming a molecular clock and a Jukes-Cantor model. . . . .	189
6.8	Results of simulations for the four-taxon tree $((AB)C)D$ (Figure 6.1A) generated under a Jukes-Cantor model with $\theta = 0.01$ and a violation of the molecular clock, and analyzed under maximum likelihood assuming a molecular clock and a Jukes-Cantor model. . . . .	190
6.9	Results of simulations for the four-taxon tree $((AB)(CD))$ (Figure 6.1B) generated under a Jukes-Cantor model with $\theta = 0.01$ and a violation of the molecular clock, and analyzed under maximum likelihood assuming a molecular clock and a Jukes-Cantor model. . . . .	191
6.10	Results of simulations for the five-taxon tree $((((AB)C)D)E)$ (Figure 6.1C) generated under a Jukes-Cantor model with $\theta = 0.01$ and a violation of the molecular clock, and analyzed under maximum likelihood assuming a molecular clock and a Jukes-Cantor model. . . . .	192

- 6.11 Results of simulations for the five-taxon tree (((AB)C)(DE)) (Figure 6.1D) generated under a Jukes-Cantor model with  $\theta = 0.01$  and a violation of the molecular clock, and analyzed under maximum likelihood assuming a molecular clock and a Jukes-Cantor model. . 193
- 6.12 Results of simulations for the five-taxon (((AB)(CD))E) (Figure 6.1E) generated under a Jukes-Cantor model with  $\theta = 0.01$  and a violation of the molecular clock, and analyzed under maximum likelihood assuming a molecular clock and a Jukes-Cantor model. . 194
- 6.13 Results of simulations for the four-taxon tree (((AB)C)D) (Figure 6.1A) generated under a General Time-Reversible model with shape parameter  $\alpha = 1$ , relative frequencies for nucleotides (A, C, G, T) = (0.1, 0.2, 0.3, 0.4), relative rates of substitutions (A $\leftrightarrow$ C, A $\leftrightarrow$ G, A $\leftrightarrow$ T, C $\leftrightarrow$ G, C $\leftrightarrow$ T, G $\leftrightarrow$ T) = (1.5, 1.5, 0.5, 10.5, 1.0, 6.0),  $\theta = 0.01$ , and a molecular clock, and analyzed under maximum likelihood assuming a molecular clock and a Jukes-Cantor model. . 195
- 6.14 Results of simulations for the four-taxon tree ((AB)(CD)) (Figure 6.1B) generated under a General Time-Reversible model with shape parameter  $\alpha = 1$ , relative frequencies for nucleotides (A, C, G, T) = (0.1, 0.2, 0.3, 0.4), relative rates of substitutions (A $\leftrightarrow$ C, A $\leftrightarrow$ G, A $\leftrightarrow$ T, C $\leftrightarrow$ G, C $\leftrightarrow$ T, G $\leftrightarrow$ T) = (1.5, 1.5, 0.5, 10.5, 1.0, 6.0),  $\theta = 0.01$ , and a molecular clock, and analyzed under maximum likelihood assuming a molecular clock and a Jukes-Cantor model. . 196
- 6.15 Results of simulations for the five-taxon tree (((((AB)C)D)E) (Figure 6.1C) generated under a General Time-Reversible model with shape parameter  $\alpha = 1$ , relative frequencies for nucleotides (A, C, G, T) = (0.1, 0.2, 0.3, 0.4), relative rates of substitutions (A $\leftrightarrow$ C, A $\leftrightarrow$ G, A $\leftrightarrow$ T, C $\leftrightarrow$ G, C $\leftrightarrow$ T, G $\leftrightarrow$ T) = (1.5, 1.5, 0.5, 10.5, 1.0, 6.0),  $\theta = 0.01$ , and a molecular clock, and analyzed under maximum likelihood assuming a molecular clock and a Jukes-Cantor model. . 197
- 6.16 Results of simulations for the five-taxon tree (((AB)C)(DE)) (Figure 6.1D) generated under a General Time-Reversible model with shape parameter  $\alpha = 1$ , relative frequencies for nucleotides (A, C, G, T) = (0.1, 0.2, 0.3, 0.4), relative rates of substitutions (A $\leftrightarrow$ C, A $\leftrightarrow$ G, A $\leftrightarrow$ T, C $\leftrightarrow$ G, C $\leftrightarrow$ T, G $\leftrightarrow$ T) = (1.5, 1.5, 0.5, 10.5, 1.0, 6.0),  $\theta = 0.01$ , and a molecular clock, and analyzed under maximum likelihood assuming a molecular clock and a Jukes-Cantor model. . 198

6.17	Results of simulations for the five-taxon tree $((((AB)(CD))E)$ (Figure 6.1E) generated under a General Time-Reversible model with shape parameter $\alpha = 1$ , relative frequencies for nucleotides $(A, C, G, T) = (0.1, 0.2, 0.3, 0.4)$ , relative rates of substitutions $(A \leftrightarrow C, A \leftrightarrow G, A \leftrightarrow T, C \leftrightarrow G, C \leftrightarrow T, G \leftrightarrow T) = (1.5, 1.5, 0.5, 10.5, 1.0, 6.0)$ , $\theta = 0.01$ , and a molecular clock, and analyzed under maximum likelihood assuming a molecular clock and a Jukes-Cantor model. . . . .	199
6.18	Proportion of times SMRT-ML recovers the estimated species tree or at least one false clade for random subsets of genes from the original data set. . . . .	200
6.19	Bootstrap support percentages for nodes in the SMRT-ML yeast analysis using the 106-gene data set. . . . .	201
6.20	A three-taxon gene tree within a model species tree with notation used in the paper. . . . .	202
6.21	Results of simulations for a six-taxon tree with topology $(((((AB)(CD))E)F)$ . . . . .	203
6.22	Results of simulations for a four-taxon tree with topology $((((AB)C)D)$ . . . . .	204
7.1	Model for the relationship among species A, B and C in a fixed species tree $\sigma$ . . . . .	237
7.2	Counterexample used to prove that consensus methods are misleading. . . . .	238
7.3	Four possible coalescence ranks of four-taxon trees. . . . .	239
7.4	Simulation results for the three-taxon species tree $((AB)C)$ . . . . .	240
7.5	Inference of species trees using GLASS/Maximum Tree under a Jukes-Cantor substitution model (per-site mutation rate $\theta = 0.01$ ) when gene trees are generated under the three-taxon species tree $\sigma = ((A:1.0, B:1.0):0.1, C:1.1)$ . . . . .	241
8.1	Flow diagram representing the procedure in which we obtained results on the behavior of phylogenetic inference strategies. . . . .	277
8.2	Schematic for creating the four subsets $\mathcal{D}_s$ , $\mathcal{D}_{s,0}$ , $\mathcal{D}_p$ , and $\mathcal{D}_{p,0}$ from dataset $\mathcal{D}$ (see Table 8.2). . . . .	279
8.3	Distribution of clade size for all 72 phylogenetic inference strategies. . . . .	280



8.4	Principal components analysis of phylogenetic inference strategies. . .	281
8.5	Procrustes analysis of the principal component plots in Figure 8.4. . .	282
8.6	Cluster and correlation analysis of phylogenetic inference strategies. . .	284
8.7	Heat map representing the “flow” of clades between phylogenetic inference strategies. . . . .	285
8.8	Consensus trees of phylogenetic inference strategies averaged over outgroups. . . . .	286
8.9	Consensus trees of phylogenetic inference strategies averaged over outgroups and gene tree inference methods. . . . .	287
8.10	Consensus trees of phylogenetic inference strategies averaged over outgroups and species tree construction methods. . . . .	288

## LIST OF TABLES

### Table

2.1	Joint distribution of the numbers of $i$ alleles carried by individuals $j$ and $k$ given their descent configuration $S$ , assuming allele $i$ has frequency $p$ . . . . .	28
2.2	The 26 populations containing relatives in the H1048 data set . . .	30
2.3	Mean squared error (MSE), variance, and bias squared of estimates for data simulated based on allele frequencies at two loci (AAT263P and ACT3F12) . . . . .	31
2.4	Statistical tests applied to the mean gene diversity across loci . . . .	33
3.1	Relationship types with corresponding X-linked kinship coefficients .	63
3.2	Symbols used for relative pair types. . . . .	64
3.3	Comparison of exact, approximate, and simulation variances . . . .	65
3.4	Types of relative pairs in populations from the dataset of 485 individuals reported by <i>Jakobsson et al.</i> (2008) . . . . .	66
6.1	Probabilities of concordant and most probable discordant gene trees and performance of SM-ML and SMRT-ML with 6000 loci . . . . .	181
7.1	Notation . . . . .	234
7.2	Summary of the behavior of consensus methods . . . . .	236
8.1	Phylogenetic inference strategies . . . . .	273
8.2	Datasets . . . . .	276

# CHAPTER I

## Introduction

In recent years, novel high-throughput genotyping and sequencing methods and dramatic increases in computational power have enabled geneticists to study human variation at a fine scale. This variation can be separated into genetic and phenotypic variation, both of which are influenced by evolutionary processes such as migration, mutation, genetic drift, and natural selection (see Figure 1.1).

Genetic variation is directly influenced by neutral processes such as migration, mutation, and genetic drift. Migration can act to increase genetic variation within a population through immigration, causing diverse alleles to be assimilated into a population's gene pool. Similarly, migration can act to decrease genetic variation within a population through emigration, causing alleles to be removed from a population. Mutation causes an increase in genetic variation through the creation of new alleles not previously present within a population. In contrast, genetic drift causes a decrease in genetic variation through the random fixation or loss of alleles within a population.

Phenotypic variation is directly influenced by both genetic variation and selection. Genetic variation influences phenotypic variation by the generation of novel phenotypic traits. Selection, in turn, acts on these traits, promoting either variation or lack of variation. Selection thereby indirectly influences genetic variation through its

action on phenotypic variation. Because of the complex interplay among the many neutral and selective forces acting on variation, to elucidate the demographic and adaptive processes that have led to modern human evolution, we must understand the neutral processes that have shaped the background levels of genetic variation among populations. Therefore, in this thesis, I will focus only on neutral processes that affect genetic variation at multiple levels: variation within populations, variation among populations within a species, and variation among species.

In addition to its importance within population genetics, the study of genetic variation is important to other fields, including medical genetics (e.g., *Tishkoff and Kidd*, 2004), forensic science (e.g., *Evetts and Weir*, 1998), anthropology (e.g., *Cavalli-Sforza and Feldman*, 2003), and the many other disciplines for which knowledge about population relationships is important. For example, in disease-gene mapping, it is necessary to carefully correct for population stratification in association studies to circumvent the problem of falsely associating a genetic variant with a disease (*Marchini et al.*, 2004; *Clayton et al.*, 2005; *McCarthy et al.*, 2008). Similarly, the correlations of allelic states across loci within populations directly influence studies that seek to isolate disease-causing variants (*Kruglyak*, 1999; *Hirschhorn et al.*, 2002; *Hirschhorn and Daly*, 2005; *Rosenberg et al.*, 2010). As a consequence, knowledge of patterns of genetic variation among populations from around the world can improve the power of association studies, thereby ultimately influencing our understanding of the genetic basis of human diseases.

Because of the importance of genetic variation, we need to be able to properly assess it by developing accurate estimators of its properties. There now exist many publicly available large-scale multilocus datasets from human individuals around the world. These datasets include the Human Genome Diversity Project-Centre d'Etude du Polymorphisme Humain (HGDP-CEPH) Cell Line Panel (*Cann et al.*, 2002; *Cavalli-Sforza*, 2005), the HapMap Project (*International HapMap Consortium*,

2005, 2007; *International HapMap 3 Consortium*, 2010), and the 1000 Genomes Project (*The 1000 Genomes Project Consortium*, 2010). Recent work facilitated by these datasets has led to important advancements in our understanding of worldwide human genetic variation. For example, the HGDP has been useful for understanding the population-genetic processes that have led to modern human evolution (*Rosenberg et al.*, 2002; *Prugnolle et al.*, 2005; *Wang et al.*, 2007; *Friedlaender et al.*, 2008; *Jakobsson et al.*, 2008; *Li et al.*, 2008; *Tishkoff et al.*, 2009). Models based on HGDP data have been shown to mimic trends observed from the data (*Ramachandran et al.*, 2005; *Liu et al.*, 2006; *DeGiorgio et al.*, 2009; *Hunley et al.*, 2009). Additionally, datasets from the HapMap and 1000 Genomes Projects have been, and will be in the future, useful for mapping genetic variants that are associated with disease (*International HapMap 3 Consortium*, 2010; *Nielsen*, 2010; *The 1000 Genomes Project Consortium*, 2010; *Jostins et al.*, 2011).

One important consideration when using these datasets is that they often contain related individuals (*Rosenberg*, 2006; *Pemberton et al.*, 2010). The inclusion of relatives within a dataset is not necessarily problematic, especially because random sampling of a population with a small size is likely to yield a dataset with related individuals. However, if the presence of related individuals is not properly taken into account, then estimates of genetic variation (e.g., expected heterozygosity) from these datasets could be biased. Chapters II and III tackle the problem of correcting the bias in estimates of a specific measure of genetic variation (termed gene diversity or expected heterozygosity) that is generated by the inclusion of related individuals within samples.

Chapter II (*DeGiorgio and Rosenberg*, 2009) develops an unbiased estimator of gene diversity for samples containing related individuals at diploid autosomal loci. This tool will be useful in assessing autosomal genetic variation within humans. However, as Chapter II does not show that my estimator is unbiased on samples

at X-linked loci or at loci from non-diploid individuals, the derivation of an estimator that can account for arbitrary ploidy would have wide applicability to studies of humans as well as other diploid or possibly non-diploid species.

Chapter III (*DeGiorgio et al.*, 2010) extends the work of Chapter II to the most general case of an unbiased estimator of gene diversity for samples containing related individuals with arbitrary ploidy. This chapter also presents the first derivation of the exact variance of the estimator of gene diversity in samples containing related individuals. Previous variance calculations were only obtained in approximation for an unbiased estimator of gene diversity in samples containing unrelated individuals at diploid autosomal loci (*Weir*, 1989). Because many population-genetic datasets contain relatives, my estimators will be valuable tools for assessing genetic variation within those populations (e.g., *Jankovic et al.*, 2010).

Using estimators of genetic variation applied to large-scale genomic datasets, it is possible to further investigate evolutionary hypotheses. One set of evolutionary hypotheses pertains to the origin of anatomically modern humans. A popular hypothesis for modern human origins is the out-of-Africa hypothesis (Figure 1.2A), which states that all non-African populations of modern human descend from a common ancestral population that lived in Africa 150,000-200,000 years ago and migrated out 50,000-100,000 years ago (*Relethford*, 2008). A competing hypothesis for modern human origins is the multiregional hypothesis (Figure 1.2B), which states that modern humans descend from the continual mating of distinct groups of archaic hominids such as *Homo erectus* (*Relethford*, 2008). The out-of-Africa and multiregional hypotheses are not completely disjoint, however. A relaxed version of the out-of-Africa hypothesis states that modern humans mated with archaic Neanderthal populations during their expansion out of Africa, leading to the introgression of archaic human genes into the modern gene pool. This hypothesis has spurred a heated debate that remains unresolved (*Currat and Excoffier*, 2004; *Serre*

*et al.*, 2004; *Garrigan and Hammer*, 2006; *Green et al.*, 2006; *Noonan et al.*, 2006; *Plagnol and Wall*, 2006; *Wall et al.*, 2009; *Green et al.*, 2010). In Chapters IV and V, I investigate these hypotheses of modern human origins using models of evolutionary history.

Chapter IV (*DeGiorgio et al.*, 2009) is a simulation study in which I qualitatively compare and contrast patterns of within-population genetic variation predicted by models of evolutionary history with the analogous patterns observed from worldwide human genetic data. I investigate several models of evolutionary history, including one that represents the out-of-Africa hypothesis (serial founder model), one that represents a version of the multiregional hypothesis (archaic persistence model), one that represents a relaxed version of the out-of-Africa hypothesis involving admixture with archaic humans (serial founder model with archaic admixture), and a model that illustrates the evolutionary process that drives within-population human genetic variation (instantaneous divergence model).

An alternative to investigating the patterns of genetic variation predicted by models of human evolutionary history through simulation studies is to make analytical predictions of genetic variation under the models. Analytical predictions are useful for multiple reasons, as they enable the generation of hypotheses as well as model parameter estimation and hypothesis testing. Although simulation studies have the advantage that they can commonly investigate more complex models than can be explored through analytical theory, analytical results can offer a faster and more comprehensive exploration of the space of parameter values for a model that would otherwise be investigated through simulations. Chapter V complements the work of Chapter IV through the rigorous development of analytical, in contrast to simulation-based, results for models of human evolutionary history. I recapitulate several patterns of within-population genetic variation obtained through simulations in Chapter IV as well as extend the set of patterns by including those involving

between-population genetic variation. I find in Chapter V that investigating patterns of genetic variation between populations (e.g., using inter-population expected homozygosity and  $F_{ST}$ ) can provide additional insight into human evolutionary history and enables us to distinguish patterns predicted by models of evolutionary history that were not possible to distinguish using genetic variation within populations (e.g., using intra-population expected heterozygosity). However, although within- and between-population genetic relationships are useful for understanding human evolution, they neglect information that can be extracted from our evolutionary relationships with other species.

We can determine our relationships with other species by constructing phylogenetic trees from sequence data across a set of species. However, the estimation of phylogenetic trees is a complex task that requires knowledge beyond simple sequence differences. For example, due to diverse evolutionary phenomena, trees estimated at different genomic loci (or gene trees) can disagree (*Rannala and Yang, 2008*). The phenomenon that generates gene tree discordance on which I focus in this dissertation is incomplete lineage sorting, in which sets of sampled lineages fail to coalesce in the population in which they are first capable of coalescing (*Degnan and Rosenberg, 2009*). Due to the discordance of gene trees, methods for inferring species trees from multilocus data can be misled by the conflicting signals observed over the set of loci investigated (*Degnan and Rosenberg, 2006; Kubatko and Degnan, 2007; Degnan et al., 2009*). As genetic datasets rapidly increase in size, it is becoming increasingly important to develop and evaluate phylogenetic methods that can overcome these obstacles, and to produce accurate estimates of species trees from multilocus data. Chapters VI, VII, and VIII develop and analyze the performance of methods for estimating species trees from multilocus datasets.

In Chapter VI (*DeGiorgio and Degnan, 2010*), I develop a method (SuperMatrix Rooted Triple, or SMRT) for inferring species trees from multilocus sequence data



that is both computationally efficient and accurate when applied to genome-scale data. Additionally, using a DNA substitution model applied to genealogies generated under the coalescent, I prove that SMRT is a statistically consistent estimator of species tree topologies. Statistical consistency is a desirable property because it is reasonable to expect that as more data are gathered, evidence should accumulate in support of the true value of the parameter being estimated.

Chapter VII presents a subsequent study that investigates the statistical consistency of various species tree inference methods when the relationship among species is obscured due to non-random mating (or population structure) in ancient species. This feature of non-random mating in ancient species is not typically included in phylogenetic models. However, because extant species are commonly structured, it is reasonable to believe that ancient species were also structured.

In addition to analyzing the statistical consistency of species tree inference methods, it is important to understand how these methods perform in practice. Empirical phylogenetic datasets tend to have fewer than 100 loci available and, therefore, the performance of a species tree inference method as the number of loci gets large (*i.e.*, statistical consistency) may not be relevant in practice. Typically, the performance of species tree inference methods is studied through simulations, which can explore only a small evolutionary parameter space. An alternative approach is to evaluate the performance of methods on a space of parameters defined by the actual evolutionary history of a group of species. Chapter VIII presents such a study, in which I evaluate the performance of species tree inference methods using an empirical multilocus dataset from North American pines. I utilize techniques from multivariate statistical analysis, such as principal components and cluster and correlation analyses, to thoroughly investigate the strengths and weaknesses of each method. These results will be useful for investigators within the community of researchers interested in phylogeny of closely related species.

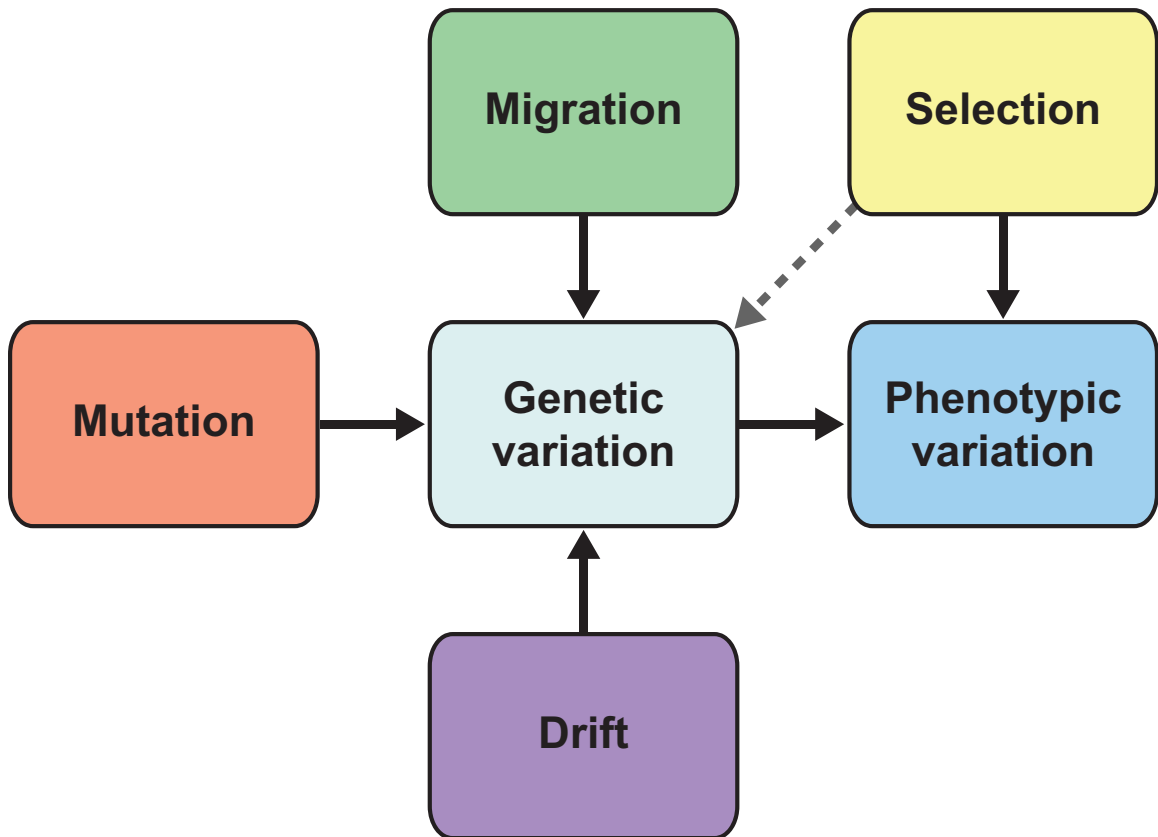


Figure 1.1: Processes affecting both genetic and phenotypic variation. A solid black arrow indicates a direct influence and a dotted gray arrow indicates an indirect influence.

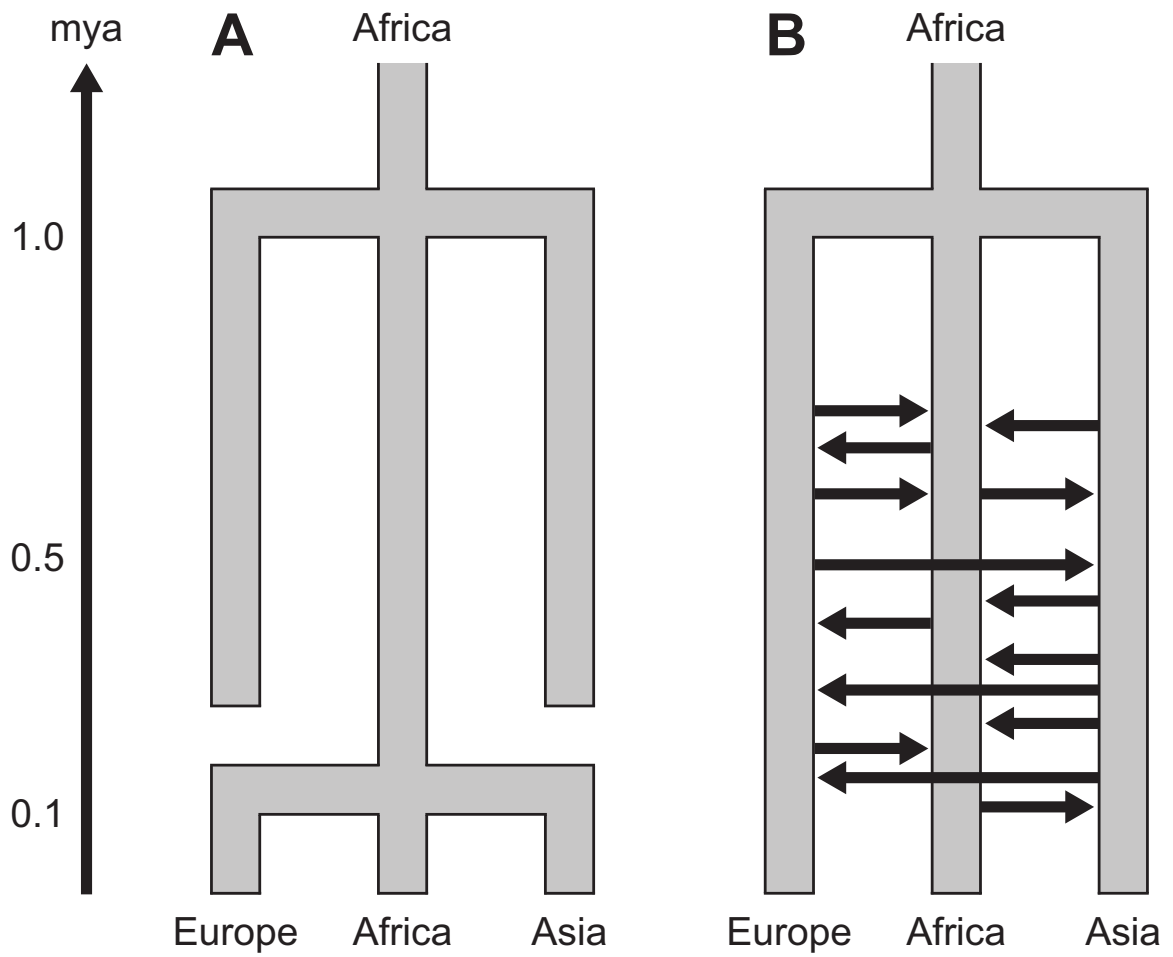


Figure 1.2: Simple models representing hypotheses for modern human origins. (A) Out-of-Africa hypothesis. (B) Multiregional hypothesis.

## CHAPTER II

# An Unbiased Estimator of Gene Diversity in Samples Containing Related Individuals

### 2.1 Introduction

Gene diversity, or expected heterozygosity, is a frequently used measure of genetic variation applied in diverse areas of population genetics. Together with its counterpart, gene identity or expected homozygosity, it has been used to quantify genetic variation in populations (*Driscoll et al.*, 2002; *Hoelzel et al.*, 2002), evaluate genetic divergence and population relationships (*Nei*, 1973; *Ramachandran et al.*, 2005), detect inbreeding (*Li and Horvitz*, 1953), measure linkage disequilibrium (*Ohta*, 1980; *Sabatti and Risch*, 2002), and test for the influence of natural selection (*Watterson*, 1978; *Depaulis and Veuille*, 1998; *Sabeti et al.*, 2002).

Consider a polymorphic locus with  $I$  distinct alleles and a population with parametric allele frequencies  $p_1, p_2, \dots, p_I$ , where  $p_i \in [0, 1]$ , and  $\sum_{i=1}^I p_i = 1$ . The term “gene diversity”, which is defined as

$$H = 1 - \sum_{i=1}^I p_i^2, \quad (2.1)$$

was proposed by *Nei* (1973), though the use of eq. 2.1 as a measure of diversity dates

to considerably earlier (*Gini*, 1912; *Simpson*, 1949; *Gibbs and Martin*, 1962).

Now consider a sample of  $n$  observations of alleles, in which the number of observations of allelic type  $i$  is  $n_i$ . The count estimate of  $p_i$  is  $\hat{p}_i = n_i/n$ . If no inbred or related individuals are included in the sample, then an unbiased estimator of gene diversity is (*Nei and Roychoudhury*, 1974)

$$\hat{H} = \frac{n}{n-1} \left( 1 - \sum_{i=1}^I \hat{p}_i^2 \right). \quad (2.2)$$

If relatives or inbred individuals are included in the sample, then  $\hat{H}$  is no longer an unbiased estimator of  $H$ . To understand why this statement is true, suppose that a sample contains a pair of close relatives. Because these individuals are related, they may share one or two alleles identically by descent (IBD) at a locus (compared to zero alleles shared IBD in unrelated individuals). As a result, estimation of  $p_i$  is based on fewer independent observations than for a sample not containing any relatives. Although  $\mathbb{E}[\hat{p}_i] = p_i$  when relatives are included,  $\text{Var}[\hat{p}_i]$  is greater than it would be had no relatives been included. Observe that the computation of  $\mathbb{E}[\hat{H}]$  involves a negative coefficient for  $\mathbb{E}[\hat{p}_i^2]$ . Because  $\mathbb{E}[\hat{p}_i^2] = \text{Var}[\hat{p}_i] + \mathbb{E}[\hat{p}_i]^2$ ,  $\mathbb{E}[\hat{H}]$  decreases as  $\text{Var}[\hat{p}_i]$  increases. Thus, the inclusion of relatives results in a downward bias, so that  $\mathbb{E}[\hat{H}] < H$ . For the case in which inbred unrelated individuals with known inbreeding coefficients are included in a sample, *Weir* (1989, 1996) provided the expectation of  $1 - \sum_{i=1}^I \hat{p}_i^2$ , producing an unbiased estimator of gene diversity

$$\hat{H}_{Weir} = \frac{n}{n-1-\bar{f}} \left( 1 - \sum_{i=1}^I \hat{p}_i^2 \right), \quad (2.3)$$

where  $\bar{f}$  is the average inbreeding coefficient across individuals (see also *Shete* (2003)). When inbred individuals are included,  $\bar{f} \neq 0$  and it follows that  $\mathbb{E}[\hat{H}] < \mathbb{E}[\hat{H}_{Weir}] = H$ .

In this article, we conduct a detailed investigation of the case in which a sample includes related individuals. We derive an unbiased estimator of  $H$  for samples containing related individuals with known levels of relationship. Our derivation makes use of a formula of *Bourgain et al.* (2003) and *McPeck et al.* (2004) for the variance of count estimates of allele frequencies in samples containing inbred and related individuals. The resulting estimator incorporates kinship coefficients, the same quantitative descriptors of pairwise relationships that have been used in diverse problems involving relatives—such as evaluation of phenotypic covariances in families (*Lange*, 2002), estimation of relatedness parameters (*Weir et al.*, 2006), and quantitative-trait linkage analysis (*Almasy and Blangero*, 1998). When a sample consists only of unrelated non-inbred individuals, our new estimator  $\tilde{H}$  reduces to the standard estimator  $\hat{H}$ , and it reduces to  $\hat{H}_{Weir}$  if inbred but not related individuals are included. Using data simulated based on allele frequencies from human populations, we find that the new estimator  $\tilde{H}$  corrects for bias generated by inclusion of related individuals and that it attains a mean squared error (MSE) comparable to that of  $\hat{H}$ . We apply this new estimator to microsatellite data from human population samples containing relatives and show that, compared to the standard estimator, it produces estimates closer to those obtained when excluding relatives from the analysis.

## 2.2 Theory

We assume that gene diversity is estimated from  $n/2$  diploid individuals. Our aim is to obtain a bias-correction factor that can be incorporated into a new estimator of gene diversity,  $\tilde{H}$ . We begin by computing  $Var[\hat{p}_i]$  in a sample that may include relatives or inbred individuals.  $Var[\hat{p}_i]$  was reported by *Bourgain et al.* (2003) and *McPeck et al.* (2004); we provide an alternative derivation that uses a generalization of the simpler method of *Broman* (2001). This approach was originally applied in a setting that did not consider inbreeding, and we generalize the computation to

include inbreeding. Note that the variances of other estimators of allele frequencies have previously been derived in fairly general settings (*McPeck et al.*, 2004), and that the estimator  $\hat{p}_i$  is not a maximum likelihood estimator when related individuals are included in a sample (*Boehnke*, 1991). However, our interest here is specifically on the count-based estimator of allele frequencies, as it is this estimator that is used in the standard estimator of gene diversity in eq. 2.2.

Define  $X_k$  to be the number of alleles of type  $i$  that are carried by individual  $k$  at a particular locus.  $X_k$  can equal 0, 1, or 2, and  $\mathbb{E}[X_k] = 2p_i$ . Regardless of the relationships among individuals  $1, 2, \dots, n/2$ , an unbiased estimator for  $p_i$ , the frequency of allele  $i$ , is

$$\hat{p}_i = \frac{1}{n} \sum_{k=1}^{n/2} X_k. \quad (2.4)$$

The variance of  $\hat{p}_i$  is given by

$$Var[\hat{p}_i] = \frac{1}{n^2} \sum_{j=1}^{n/2} \sum_{k=1}^{n/2} Cov(X_j, X_k). \quad (2.5)$$

Suppose that individuals  $j$  and  $k$  are related. The coefficient of kinship between individuals  $j$  and  $k$ ,  $\Phi_{j,k}$ , is the probability that two alleles chosen at the locus — one from individual  $j$  and the other from individual  $k$  — are identical by descent. In the special case of  $j = k$ , the kinship coefficient is  $\Phi_{k,k} = (1/2)(1 + f_k)$ , where  $f_k$  is the inbreeding coefficient for individual  $k$  (*Lange*, 2002, p. 81).

Conditional on the nature of the relationship between individuals  $j$  and  $k$  and on their inbreeding coefficients, the four alleles in the two individuals can take on one of nine condensed identity states (*Jacquard*, 1974, p. 107). Let  $\Delta_s = \mathbb{P}[S = s]$ , where the condensed identity state  $S$  ranges from 1 to 9 and the probability is conditional on the type of the relationship. Using Table 2.1 and the fact that the kinship coefficient for the pair of individuals equals  $\Delta_1 + (1/2)(\Delta_3 + \Delta_5 + \Delta_7) + (1/4)\Delta_8$  (*Jacquard*,

1974, p. 108), we obtain

$$\begin{aligned}\mathbb{E}[X_j X_k] &= \sum_{a=0}^2 \sum_{b=0}^2 \sum_{s=1}^9 ab \Delta_s \mathbb{P}[X_j = a, X_k = b | S = s] \\ &= 4\Phi_{j,k} p_i (1 - p_i) + 4p_i^2.\end{aligned}$$

Since  $\mathbb{E}[X_j] = \mathbb{E}[X_k] = 2p_i$ , it follows that

$$\begin{aligned}\text{Cov}(X_j, X_k) &= \mathbb{E}[X_j X_k] - \mathbb{E}[X_j] \mathbb{E}[X_k] \\ &= 4\Phi_{j,k} p_i (1 - p_i).\end{aligned}\tag{2.6}$$

Inserting the covariance into eq. 2.5 yields

$$\begin{aligned}\text{Var}[\hat{p}_i] &= \frac{4p_i(1 - p_i)}{n^2} \sum_{j=1}^{n/2} \sum_{k=1}^{n/2} \Phi_{j,k} \\ &= \bar{\Phi} p_i (1 - p_i),\end{aligned}\tag{2.7}$$

where  $\bar{\Phi} = \frac{1}{(n/2)^2} \sum_{j=1}^{n/2} \sum_{k=1}^{n/2} \Phi_{j,k}$  is the average kinship coefficient across pairs of individuals (including comparisons of individuals with themselves). This result can be seen to be equivalent to the variance reported by *McPeck et al.* (2004), p. 361.

**Proposition II.1.** *Consider a locus with  $I$  distinct alleles, allele frequencies  $p_i \in [0, 1]$  and  $\sum_{i=1}^I p_i = 1$ . Suppose a sample from a population has  $n/2$  possibly related and inbred individuals. Then an unbiased estimator for gene diversity is*

$$\tilde{H} = \frac{1}{1 - \bar{\Phi}} \left( 1 - \sum_{i=1}^I \hat{p}_i^2 \right),\tag{2.8}$$

where  $\Phi_{j,k}$  is the kinship coefficient of individuals  $j$  and  $k$  and  $\bar{\Phi} = \frac{1}{(n/2)^2} \sum_{j=1}^{n/2} \sum_{k=1}^{n/2} \Phi_{j,k}$  is the average kinship coefficient across pairs of individuals.

*Proof.* We need to show that  $\mathbb{E}[\tilde{H}] = H$ . Observing that  $\mathbb{E}[\hat{p}_i^2] = \text{Var}[\hat{p}_i] + \mathbb{E}[\hat{p}_i]^2$  and



$\mathbb{E}[\widehat{p}_i] = p_i$ , we apply eq. 2.4 and then the variance of  $\widehat{p}_i$  in eq. 2.7 to get

$$\begin{aligned}\mathbb{E}[\widetilde{H}] &= \frac{1}{1 - \overline{\Phi}} \left[ 1 - \sum_{i=1}^I (\text{Var}[\widehat{p}_i] + p_i^2) \right] \\ &= \frac{1}{1 - \overline{\Phi}} \left[ 1 - \sum_{i=1}^I (\overline{\Phi} p_i (1 - p_i) + p_i^2) \right] \\ &= H. \quad \square\end{aligned}$$

**Corollary II.2.** *Consider a locus with  $I$  distinct alleles, allele frequencies  $p_i \in [0, 1]$  and  $\sum_{i=1}^I p_i = 1$ . Suppose a sample from a population has  $n/2$  possibly related and inbred individuals. Let  $\mathcal{R}$  be the set of distinct types of relative pairs in the sample. Further, let  $n_R$  be the number of pairs of individuals with relationship type  $R \in \mathcal{R}$  and let  $\Phi_R$  be the kinship coefficient for each of these pairs. Then an unbiased estimator for gene diversity is*

$$\widetilde{H} = \frac{n(n-1)}{n(n-1-\bar{f}) - 8 \sum_{R \in \mathcal{R}} n_R \Phi_R} \widehat{H}, \quad (2.9)$$

where  $\bar{f} = \frac{1}{n/2} \sum_{k=1}^{n/2} f_k$  is the average inbreeding coefficient across individuals and  $f_k$  is the inbreeding coefficient for individual  $k$ .

*Proof.* Applying the definitions of  $\overline{\Phi}$  and  $\Phi_{k,k}$  and the fact that  $\Phi_{j,k} = 0$  for a pair of “unrelated” individuals,

$$\begin{aligned}\overline{\Phi} &= \frac{1}{(n/2)^2} \sum_{j=1}^{n/2} \sum_{k=1}^{n/2} \Phi_{j,k} \\ &= \frac{4}{n^2} \left( \sum_{k=1}^{n/2} \Phi_{k,k} + 2 \sum_{j=1}^{n/2} \sum_{k=j+1}^{n/2} \Phi_{j,k} \right) \\ &= \frac{1}{n^2} \left( n + n\bar{f} + 8 \sum_{R \in \mathcal{R}} n_R \Phi_R \right).\end{aligned}$$

Inserting this value for  $\overline{\Phi}$  into eq. 3.10 we obtain the desired result. □

Note that if no related individuals are included in the sample, then  $\mathcal{R}$  is the empty set, thus reducing  $\tilde{H}$  to  $\hat{H}_{Weir}$ ; if additionally no related individuals are included, then  $\bar{f} = 0$  and  $\tilde{H}$  reduces to  $\hat{H}$ .

**Corollary II.3.** *Consider a locus with  $I$  distinct alleles, allele frequencies  $p_i \in [0, 1]$  and  $\sum_{i=1}^I p_i = 1$ . Suppose a sample from a population has  $n/2$  non-inbred individuals, among which  $q$  parent-offspring pairs,  $r$  full-sib pairs, and  $s$  second-degree (avuncular, grandparent-grandchild, and half-sib) relative pairs are included. Assuming the sample has no other relative pairs, an unbiased estimator for gene diversity is*

$$\tilde{H} = \frac{n(n-1)}{n(n-1) - 2q - 2r - s} \hat{H}. \quad (2.10)$$

*Proof.* The kinship coefficients are  $\Phi_P = 1/4$  for parent-offspring pairs,  $\Phi_F = 1/4$  for full-sib pairs, and  $\Phi_S = 1/8$  for second-degree pairs. If an individual  $k$  is not inbred, then  $f_k = 0$ . For a sample without inbred individuals,  $\bar{f} = 0$ . Inserting the quantity and kinship coefficient for each of the three types of relative pairs into eq. 3.13, we obtain eq. 2.10.  $\square$

**Corollary II.4.** *Consider a locus with  $I$  distinct alleles, allele frequencies  $p_i \in [0, 1]$  and  $\sum_{i=1}^I p_i = 1$ . Suppose a sample from a population has  $n/2$  possibly related and inbred individuals. Let  $\mathcal{R}$  be the set of distinct types of relative pairs in the sample. Further, let  $n_R$  be the number of pairs of individuals with relationship type  $R \in \mathcal{R}$  and let  $\Phi_R$  be the kinship coefficient for each of these pairs. Then the bias of  $\hat{H}$  is always negative, increases in magnitude as  $H$  increases, and is given by*

$$\text{bias}(\hat{H}) = -\frac{n\bar{f} + 8 \sum_{R \in \mathcal{R}} n_R \Phi_R}{n(n-1)} H, \quad (2.11)$$

where  $\bar{f} = \frac{1}{n/2} \sum_{k=1}^{n/2} f_k$  is the average inbreeding coefficient across individuals and  $f_k$  is the inbreeding coefficient for individual  $k$ .

*Proof.* As shown in Corollary II.2,  $\tilde{H} = c\hat{H}$ , where  $c = n(n-1)/[n(n-1-\bar{f}) - 8\sum_{R\in\mathcal{R}} n_R\Phi_R]$ . Rearranging and taking the expected value gives  $\mathbb{E}[\hat{H}] = \mathbb{E}[\tilde{H}]/c = H/c$ . The desired result follows from simplifying the expression for  $\text{bias}(\hat{H})$ , or  $(1-c)H/c$ .  $\square$

## 2.3 Data from Human Populations

To examine the behavior of  $\tilde{H}$  in a realistic setting, we performed simulations and data analysis using microsatellite loci from the H1048 and H952 subsets (*Rosenberg, 2006*) of the Human Genome Diversity Project-Centre d'Etude du Polymorphisme Humain (HGDP-CEPH) Cell Line Panel (*Cann et al., 2002; Cavalli-Sforza, 2005*). The H1048 subset consists of 1048 individuals in 53 populations. Among the 53 populations, the samples from 26 of them contain at least one pair of closely related individuals with either a first-degree (parent-offspring, full-sib) or second-degree (avuncular, grandparent-grandchild, half-sib) relationship (Table 2.2). The H952 subset is a collection of 952 individuals included in the larger H1048 subset. No two of the 952 individuals are believed to have a first- or second-degree relationship. Levels of relationship in H1048, as estimated previously from microsatellite genotypes (*Rosenberg, 2006*), were treated here as known with certainty. Since no cycles were observed in pedigrees from the HGDP-CEPH panel (*Rosenberg, 2006*), we assumed that none of the panel members were inbred. Genotypes at 783 autosomal microsatellite loci (*Ramachandran et al., 2005; Rosenberg et al., 2005*) were investigated in the H1048 and H952 data sets.

## 2.4 Simulations

### 2.4.1 Simulation Procedure

Simulations based on the microsatellite loci were used to examine the properties of  $\tilde{H}$  and  $\hat{H}$ . For each of the 783 loci, we treated allele frequencies estimated from the H952 subset of individuals as true allele frequencies. The parametric gene diversity  $H$  was obtained for a locus as one minus the sum of the squares of these allele frequencies. All of our simulations assumed no inbreeding.

For a given locus, individual genotypes were simulated by sampling two alleles independently from the allele frequency distribution. To simulate a related individual with a given level of relationship to another individual, the number of alleles shared IBD with its relative was drawn under the appropriate probability distribution for the specified type of relative pair (parent-offspring, full-sib, or second-degree). This number of shared alleles (0, 1, or 2) was copied from a random individual that had already been generated and that had not yet been paired with a relative; if the number of alleles copied was 1, then an allele was chosen at random from the previously generated individual. The rest of the alleles, if any, were sampled independently from the allele frequency distribution. Gene diversity was estimated using  $\tilde{H}$  and  $\hat{H}$  for samples with and without related individuals. We applied  $\hat{H}$  both to entire samples as well to samples in which the “second” member of each relative pair was discarded. For each locus, simulated sets of individuals were obtained 100,000 times, and  $\hat{H}$ ,  $\tilde{H}$ ,  $\hat{H}^2$ , and  $\tilde{H}^2$  were averaged across all replicates. The true value for gene diversity,  $H$ , was then subtracted from the mean of  $\hat{H}$  and  $\tilde{H}$  to calculate bias for each estimator (and the result was squared to give bias squared). Variance of  $\hat{H}$  was calculated by subtracting the square of the mean of  $\hat{H}$  from the mean of  $\hat{H}^2$  (variance of  $\tilde{H}$  was calculated analogously). MSE was then calculated by adding variance and bias squared. Note that in our simulations, relative pairs were all disjoint, so that no

individual was contained in multiple relative pairs; however, in our derivations, it is not required for relative pairs to be disjoint for  $\tilde{H}$  to be unbiased.

### 2.4.2 Simulation Results

To illustrate the performance of the estimators across the span of gene diversities present in the human microsatellite dataset, loci were placed in increasing order by assumed parametric gene diversity, and six equally spaced loci—with the 112th, 224th, 336th, 448th, 560th, and 672nd highest values of gene diversity—were chosen for analysis. Similar results were obtained with all six loci (not shown), and therefore, among the six loci only the locus with the lowest gene diversity (AAT263P,  $H = 0.6778$ ) and the locus with the highest gene diversity (ACT3F12,  $H = 0.8263$ ) were chosen for display. For both loci, Table 2.3 shows the simulated MSE, variance, and bias squared for the different estimators, considering three different sample sizes and three combinations of the number of related individuals for each sample size. Since the simulation results are based on 100,000 replicate datasets, each of the quantities presented is small. However, it is possible to observe differences in the properties of the three estimators. Among the three estimators,  $\hat{H}$  applied to full samples gives the lowest variance,  $\tilde{H}$  produces slightly higher variance, and  $\hat{H}$  applied to samples with related individuals removed produces the highest variance. Bias squared is very close to zero for  $\hat{H}$  applied to samples with related individuals removed, as well as for  $\tilde{H}$ , but it is noticeably higher for  $\hat{H}$  applied to full samples containing relatives. For the locus with the lower value of  $H$  (0.6778),  $\hat{H}$  applied to full samples has the smallest MSE in all cases tested, although  $\tilde{H}$  has MSE very close to that of  $\hat{H}$ . However, for the locus with the higher value of  $H$  (0.8263), MSE is always smallest for  $\tilde{H}$ . Therefore,  $\tilde{H}$  is not only unbiased, but it also has MSE comparable to—and sometimes smaller than—that of  $\hat{H}$ .

It is instructive to investigate the influence of specific variables on the MSE,

variance, and bias squared of  $\tilde{H}$  and  $\hat{H}$ , by varying the simulation parameters over the space of gene diversities, sample sizes, and possible sets of relative pairs, and calculating MSE, variance, and bias squared for each scenario. We use  $\hat{H}_{full}$  and  $\tilde{H}_{full}$  to denote  $\hat{H}$  and  $\tilde{H}$  applied to a sample of individuals. For  $\hat{H}$  applied to a sample in which related individuals are removed, we use the notation  $\hat{H}_{reduced}$ .

Figure 2.1 displays the effect of sample size on MSE for each of the estimators, for scenarios in which all simulated individuals belong to relative pairs of a particular type. Here, the full and reduced samples consist of  $m$  and  $m/2$  individuals, respectively. When  $q = m/2$ ,  $r = m/2$ , or  $s = m/2$ , MSE is consistently lower for  $\hat{H}_{full}$  and  $\tilde{H}_{full}$  (which have virtually identical MSE and therefore have overlapping lines in the graph) than for  $\hat{H}_{reduced}$ . As the sample size increases, the MSEs of all estimators approach zero.

We next examined how the three estimators performed in simulated samples containing the same sample size and total number of relative pairs, but with different combinations involving different numbers of parent-offspring, full-sib, and second-degree pairs. The same two loci that were analyzed in Table 2.3 and Figure 2.1 were investigated to show the effect of the combination of relative pairs at differing degrees of gene diversity. Figures 2.2 and 2.3 illustrate MSE, variance, and bias squared for each estimator as functions of the combination of types of relative pairs in a full sample of size 40 and a reduced sample of size 20 individuals. Each point in a triangle represents the number of parent-offspring, full-sib, and second-degree relative pairs in a sample; the sum of these quantities is equal to half the sample size. MSE and variance are always lower for  $\hat{H}_{full}$  and  $\tilde{H}_{full}$  than for  $\hat{H}_{reduced}$ , which relies on a smaller sample size, and  $\hat{H}_{full}$  and  $\tilde{H}_{full}$  show similar trends. Bias squared for the unbiased  $\tilde{H}_{full}$  is similar to that for  $\hat{H}_{reduced}$ , which eliminates relatives from the sample, whereas it is much larger for  $\hat{H}_{full}$ . As the number of first-degree pairs is increased (decreasing the number of second-degree pairs), both variance and MSE

increase. For  $\widehat{H}_{full}$ , as can be predicted from eq. 2.11, bias squared also increases with an increase in the number of first-degree pairs. Since they are both unbiased estimators,  $\widetilde{H}_{full}$  and  $\widehat{H}_{reduced}$  display no particular pattern for bias squared.

Finally, we studied the trends in MSE, variance, and bias squared for the estimators over the space of gene diversities, holding the full sample size fixed at 30 individuals and the reduced sample size fixed at 15. Unlike the analyses in Table 2.3 and Figures 2.1-2.3, which show results based on two representative loci, this analysis used simulations based on all 783 microsatellites. We considered a scenario in which the sample of 30 individuals consisted of 15 parent-offspring pairs. Figure 2.4 illustrates that for all three estimators, MSE and variance tend to decrease as gene diversity increases. Since  $\widetilde{H}_{full}$  and  $\widehat{H}_{reduced}$  are both unbiased, bias squared shows no trend for these estimators. However, since bias for  $\widehat{H}_{full}$  is linear with respect to gene diversity (eq. 2.11), bias squared is quadratic. On the basis of eq. 2.11, we predict  $[bias(\widehat{H}_{full})]^2 = (-\frac{8 \times 15 \times (1/4)}{60 \times 59} H)^2 \approx (7.182 \times 10^{-5}) H^2$ , and a close match to this prediction was observed. The regression displayed in Figure 2.4 has regression model  $[bias(\widehat{H}_{full})]^2 = (7.187 \times 10^{-5}) H^2$ .

Three main results can be observed in our simulations. First,  $\widetilde{H}$  is unbiased and has comparable bias in samples containing relatives to that obtained by applying  $\widehat{H}$  to samples with relatives removed. Using  $\widetilde{H}$ , or excluding relatives and using  $\widehat{H}$ , reduces the bias compared to using  $\widehat{H}$  without excluding relatives. Second,  $\widetilde{H}$  has comparable (but consistently slightly higher) variance to the values obtained with  $\widehat{H}$  in samples containing relatives. Both  $\widetilde{H}$  and  $\widehat{H}$  have lower variance in full samples of individuals than that of  $\widehat{H}$  in reduced samples that exclude relatives. Third, because  $\widetilde{H}$  has less bias than  $\widehat{H}$  in samples containing relatives,  $\widetilde{H}$  has comparable, and sometimes smaller, MSE to  $\widehat{H}$  (although its variance is larger). Both estimators have lower MSE than  $\widehat{H}$  applied to subsets that exclude relatives.

The properties of the estimators depend on a number of parameters. All estimators

have lower MSE as sample size increases. In addition, the MSEs of  $\hat{H}$  and  $\tilde{H}$  are smaller when second-degree relative pairs are investigated, in comparison to scenarios that include an equivalent number of first-degree pairs. Furthermore, the MSEs of  $\hat{H}$  and  $\tilde{H}$  are generally smaller for loci with larger gene diversities, with the magnitude of the bias of  $\hat{H}$  increasing linearly with increasing gene diversity.

We can conclude that for samples containing relatives,  $\tilde{H}$  has comparable variance to  $\hat{H}$ , with a considerable reduction of bias.  $\tilde{H}$  has comparable bias in a full sample to that of  $\hat{H}$  applied to a reduced sample excluding relatives, with a considerable reduction of variance. Thus,  $\tilde{H}$  combines into a single estimator the desirable properties possessed by  $\hat{H}$  applied to samples with relatives and by  $\hat{H}$  applied to samples without relatives.

## 2.5 Application to Data

### 2.5.1 Notation

For convenience, we use the following notation:  $\hat{H}_{952}$  and  $\hat{H}_{1048}$  for application of  $\hat{H}$  to the samples of 952 and 1048 individuals, respectively, and  $\tilde{H}_{952}$  and  $\tilde{H}_{1048}$  for application of  $\tilde{H}$  to these samples. Note that because the H952 data set contains no relative pairs,  $\tilde{H}_{952} = \hat{H}_{952}$ , and there is no need to consider  $\tilde{H}_{952}$  separately. We also use the notation  $\hat{H}_{507}$ ,  $\hat{H}_{603}$ , and  $\tilde{H}_{603}$  when restricting our analysis to the 26 populations containing at least one relative pair; for each of the 27 remaining populations, the estimators  $\hat{H}$  and  $\tilde{H}$  produce identical values.

### 2.5.2 Mean of the Estimator

For investigating the properties of  $\hat{H}$  and  $\tilde{H}$  applied to the H1048 data set, since the true value of  $H$  is unknown for the actual data, we treated the value of  $\hat{H}_{952}$  for each locus as a substitute “true” value. Because  $\hat{H}$  is unbiased when applied to



data not containing relatives,  $\widehat{H}_{952}$  provides a sensible proxy for the unknown true gene diversity. This approach enabled us to consider how estimates of  $H$  from data including relatives might differ from estimates based on the same data excluding all relatives. For each of the 53 populations, we computed the means of  $\widehat{H}_{952}$ ,  $\widehat{H}_{1048}$ , and  $\widetilde{H}_{1048}$  across the 783 microsatellite loci. Since the true  $H$  is unknown and bias cannot be calculated, we instead examine the mean of  $\widehat{H}_{1048}$  across loci minus the mean of  $\widehat{H}_{952}$  across loci, and the mean of  $\widetilde{H}_{1048}$  across loci minus the mean of  $\widehat{H}_{952}$  across loci.

Figure 2.5 shows comparisons of the mean of  $\widehat{H}_{1048} - \widehat{H}_{952}$  across loci and the mean of  $\widetilde{H}_{1048} - \widehat{H}_{952}$  across loci. In general, the three estimators produce similar estimates in a given population. However, notice that in Figure 2.5A,  $\widehat{H}_{1048}$  is reduced compared with  $\widehat{H}_{952}$ , a likely consequence of the bias of  $\widehat{H}$  when applied to samples containing relatives. When  $\widetilde{H}_{1048}$  is used in place of  $\widehat{H}_{1048}$ , since  $\widetilde{H}_{1048}$  corrects for the inclusion of known related individuals, there is a considerable reduction in the magnitude of the difference between the mean of the estimator ( $\widehat{H}_{1048}$  or  $\widetilde{H}_{1048}$ ) across loci and the mean of  $\widehat{H}_{952}$  across loci (Figure 2.5B). These observations are reflected in Wilcoxon signed rank tests that compare paired lists of mean heterozygosities across loci for the 53 populations (Table 2.4). The  $p$ -value for a comparison of  $\widehat{H}_{1048}$  with  $\widehat{H}_{952}$  was  $8.804 \times 10^{-6}$ , suggesting that inclusion of relatives in a sample has a statistically significant impact on  $\widehat{H}$ . In contrast,  $\widetilde{H}_{1048}$  and  $\widehat{H}_{952}$  showed no significant difference, with a  $p$ -value of 0.703 for the Wilcoxon signed rank test. Similar results were obtained for other comparisons of the three estimators. The mean across populations of  $\widehat{H}_{952} - \widetilde{H}_{1048}$  ( $3.262 \times 10^{-4}$ ) was smaller than for  $\widehat{H}_{952} - \widehat{H}_{1048}$  ( $2.387 \times 10^{-3}$ ); the same was true for the mean of  $|\widehat{H}_{952} - \widetilde{H}_{1048}|$  ( $6.660 \times 10^{-4}$ ) compared with the mean of  $|\widehat{H}_{952} - \widehat{H}_{1048}|$  ( $2.387 \times 10^{-3}$ ).

Comparable results were obtained when using only the 26 populations that contained relative pairs. The Wilcoxon signed rank test produced a statistically

significant  $p$ -value of  $2.980 \times 10^{-8}$  for  $\widehat{H}_{603}$  compared with  $\widehat{H}_{507}$  and a non-significant  $p$ -value of 0.708 when comparing  $\widetilde{H}_{603}$  with  $\widehat{H}_{507}$ . The mean across populations of  $\widehat{H}_{507} - \widetilde{H}_{603}$  ( $6.649 \times 10^{-4}$ ) was smaller than for  $\widehat{H}_{507} - \widehat{H}_{603}$  ( $4.866 \times 10^{-3}$ ), as was the mean of  $|\widehat{H}_{507} - \widetilde{H}_{603}|$  ( $1.358 \times 10^{-3}$ ) relative to that of  $|\widehat{H}_{507} - \widehat{H}_{603}|$  ( $4.866 \times 10^{-3}$ ). In addition, similar numbers of populations had  $\widetilde{H}_{603} > \widehat{H}_{507}$  (12) and  $\widetilde{H}_{603} < \widehat{H}_{507}$  (14); by contrast there were no populations with  $\widehat{H}_{603} > \widehat{H}_{507}$ .

Because estimators often have a tradeoff between bias and variance, we investigated the relationship between the mean values across loci of  $\widehat{H}_{603} - \widehat{H}_{507}$  and  $\widetilde{H}_{603} - \widehat{H}_{507}$  and the standard deviations of  $\widehat{H}_{603}$  and  $\widetilde{H}_{603}$  across loci. We observed that compared to  $\widehat{H}_{603}$ ,  $\widetilde{H}_{603}$  produces a noticeable decrease in the mean difference from  $\widehat{H}_{507}$  with only a slight increase in the standard deviation (Figure 2.6). This result is somewhat analogous to the simulation-based result that  $\widetilde{H}$  has less bias than  $\widehat{H}$ , and comparable variance.

### 2.5.3 Gene Diversity vs. Distance from Africa

Based on an observed decline of gene diversity estimates with geographic distance from East Africa, *Ramachandran et al.* (2005) argued that the geographic expansion of modern humans can be described by a series of founder events originating in Africa. This analysis utilized the  $\widehat{H}$  estimator applied to the 783 microsatellites typed in the H1048 subset of individuals, excluding the Surui population. To evaluate how the results of *Ramachandran et al.* (2005) were affected by the bias of  $\widehat{H}$  in samples with close relatives, we analyzed the relationships of the three estimators of gene diversity— $\widehat{H}_{952}$ ,  $\widehat{H}_{1048}$ , and  $\widetilde{H}_{1048}$ —with geographic distance from East Africa (Figure 2.7). Distance from Addis Ababa was measured in kilometers via waypoint routes, and was based on the values from *Rosenberg et al.* (2005).

The three estimators produced relatively similar regressions (Figure 2.7), demonstrating that the close linear relationship of gene diversity and distance from

Africa is not greatly affected by inclusion of relatives in the analysis. We observed very similar values for the coefficients of determination ( $R^2$ ) of linear regressions when using  $\widehat{H}_{952}$ ,  $\widehat{H}_{1048}$ , and  $\widetilde{H}_{1048}$  (note that all three  $R^2$  values are higher than that reported by *Ramachandran et al.* (2005), whose lower value resulted from an error in the calculation of their Figure 4A). The Surui population, which has the smallest gene diversity and is the farthest population from Addis Ababa, deviates considerably from the regression line when using  $\widehat{H}_{1048}$  to measure gene diversity (Figure 2.7B). When excluding the large number of relatives present in the Surui sample ( $\widehat{H}_{952}$ ) or correcting for their inclusion ( $\widetilde{H}_{1048}$ ), the Surui population is not as extreme an outlier (Figures 2.7A and 2.7C).

## 2.6 Discussion

In this article, we have developed an unbiased estimator  $\widetilde{H}$  for gene diversity in samples containing related and inbred individuals. The bias-correction factor in this estimator, which we derived from the variance of allele frequency estimates, depends only on the average kinship coefficient between pairs of sampled individuals. Using data simulated based on allele frequency distributions from human populations, we found that  $\widetilde{H}$  performs well with regard to both bias and mean squared error. The bias generated by  $\widetilde{H}$  applied to data including relatives is approximately the same as the bias generated by the standard estimator  $\widehat{H}$  applied to data containing only unrelated individuals. The MSE for  $\widetilde{H}$  is comparable to—and often smaller than—the MSE of  $\widehat{H}$  when related individuals are included. Calculation of  $\widetilde{H}$  relies only on sample allele frequencies and on the average kinship coefficient and is therefore easy to perform when relationships among individuals are known. Thus, the new estimator  $\widetilde{H}$  offers a combination of unbiasedness, low MSE, and ease of computation, providing an improved approach to the estimation of gene diversity in samples containing relatives.

Using data from human populations, we found that  $\widetilde{H}$  largely corrected a reduction

in the standard estimator  $\hat{H}$ , producing estimates that were not significantly different from those obtained if we instead removed relatives from the data set and applied  $\hat{H}$ . This shift towards the values obtained in data without relatives occurred together with only a slight increase in standard deviation across loci relative to  $\hat{H}$ . However, by treating dependent observations as independent,  $\hat{H}$  perhaps produces a smaller variance than is appropriate in samples with relatives. Thus, we conclude that as an alternative to removing relatives from samples containing relative pairs,  $\tilde{H}$  can be applied to obtain suitable gene diversity estimates.

When we applied  $\tilde{H}$  to the human data, a few populations still produced a “bias”, in that  $\tilde{H}_{1048}$  remained considerably lower than  $\hat{H}_{952}$ . The most noticeable of these populations are the Surui, Karitiana, and Pima populations from the Americas (Figure 2.5B); the “bias” was larger for these low-diversity populations, whereas theory predicts less bias when diversity is lower (eq. 2.11). It should first be noted that unlike for the other populations, inferences about second-degree relationships obtained by *Rosenberg* (2006) were somewhat uncertain for the Surui and Karitiana populations. Thus, Table 2.2 and our analysis did not include inferred second-degree relationships in those populations, when in fact many are likely to be present. This is a likely reason why the “bias” in the Surui and Karitiana populations was only partially eliminated. For the Pima population, a likely explanation is that the sample contains many related individuals in extended families (*Rosenberg*, 2006), and our computation only adjusted for first- and second-degree relative pairs. If these higher-order relationships had been fully known, however, it would have been possible to use our estimator to adjust for them.

Our estimator adjusts for inbreeding by averaging over inbreeding coefficients for sampled individuals. It is important to note that the inbreeding coefficients that we have included are exact values obtained from pedigrees. If an estimated inbreeding coefficient was used in place of the exact value, then  $\tilde{H}$  would not necessarily produce







unbiased estimates in samples containing inbred individuals.  $\tilde{H}$  would also lead to a bias if relationships were misspecified. In our data example, relationships were assumed to be known, and for a dataset of the size used for inferring the relationships (*Rosenberg, 2006*) this assumption is generally sensible. However, for small datasets in which relationship inferences are uncertain, caution must be used when interpreting the bias of  $\tilde{H}$  applied to the same data from which relationships are estimated.

The estimators we have considered relate to within-population gene diversity. What if we consider the gene diversity between populations? Suppose we have samples from two populations,  $A$  and  $B$ , each containing related inbred individuals. The between-population analogue of gene diversity is  $\hat{H}_{A,B} = 1 - \sum_{i=1}^I \hat{p}_i \hat{q}_i$ , where  $\hat{p}_i$  and  $\hat{q}_i$  are estimates of the frequency of allele  $i$  at a given locus in populations  $A$  and  $B$ , respectively (*Nei, 1987*). Because the bias in within-population gene diversity estimates only arises from the quadratic  $\hat{p}_i^2$  term in eq. 2.1,  $E[\sum_{i=1}^I \hat{p}_i \hat{q}_i] = \sum_{i=1}^I p_i q_i$  (*Nei, 1987, p. 222*), and  $\hat{H}_{A,B}$  continues to be an unbiased estimator for between-population gene diversity in samples containing relatives.

## 2.7 Acknowledgments

We thank Ivana Jankovic, Yi-Ju Li, and two anonymous reviewers for helpful comments. This work was supported by National Institute of Health (NIH) training grant T32 GM070449, NIH grant R01 GM081441, and grants from the Burroughs Wellcome Fund and the Alfred P. Sloan Foundation.

Table 2.1: Joint distribution of the numbers of  $i$  alleles carried by individuals  $j$  and  $k$  given their descent configuration  $S$ , assuming allele  $i$  has frequency  $p$

$S$	Condensed identity state*	$X_j, X_k$	$\mathbb{P}[X_j, X_k   S]$
1		0, 0 2, 2	$1 - p$ $p$
2		0, 0 0, 2 2, 0 2, 2	$(1 - p)^2$ $p(1 - p)$ $p(1 - p)$ $p^2$
3		0, 0 0, 1 2, 1 2, 2	$(1 - p)^2$ $p(1 - p)$ $p(1 - p)$ $p^2$
4		0, 0 0, 1 0, 2 2, 0 2, 1 2, 2	$(1 - p)^3$ $2p(1 - p)^2$ $p^2(1 - p)$ $p(1 - p)^2$ $2p^2(1 - p)$ $p^3$
5		0, 0 1, 0 1, 2 2, 2	$(1 - p)^2$ $p(1 - p)$ $p(1 - p)$ $p^2$
6		0, 0 0, 2	$(1 - p)^3$ $p(1 - p)^2$

$S$	Condensed identity state*	$X_j, X_k$	$\mathbb{P}[X_j, X_k   S]$
7	$\begin{array}{c}   \\   \\   \\   \end{array}$	1,0	$2p(1-p)^2$
		1,2	$2p^2(1-p)$
		2,0	$p^2(1-p)$
		2,2	$p^3$
8	$\begin{array}{c}   \\   \\   \\   \end{array}$	0,0	$(1-p)^2$
		1,1	$2p(1-p)$
		2,2	$p^2$
		0,0	$(1-p)^3$
9	$\begin{array}{c} \bullet \\ \bullet \\ \bullet \\ \bullet \\ \bullet \\ \bullet \\ \bullet \\ \bullet \\ \bullet \\ \bullet \\ \bullet \\ \bullet \end{array}$	0,1	$p(1-p)^2$
		1,0	$p(1-p)^2$
		1,1	$p(1-p)$
		1,2	$p^2(1-p)$
		2,1	$p^2(1-p)$
		2,2	$p^3$
		0,0	$(1-p)^4$
		0,1	$2p(1-p)^3$
		0,2	$p^2(1-p)^2$
		1,0	$2p(1-p)^3$
		1,1	$4p^2(1-p)^2$
		1,2	$2p^3(1-p)$
2,0	$p^2(1-p)^2$		
2,1	$2p^3(1-p)$		
2,2	$p^4$		

\*The first row of dots represents the two alleles for individual  $j$ , and the second row represents the two alleles for individual  $k$ . Two alleles are identical by descent if there is a line connecting them.

Table 2.2: The 26 populations containing relatives in the H1048 data set

Population	Geographic region	Number of sampled individuals	Number of parent-offspring pairs	Number of full-sib pairs	Number of second-degree pairs	of
Bantu (Kenya)	Africa	12	0	1	0	
Biaka Pygmy	Africa	32	4	2	7	
Mandenka	Africa	24	0	0	2	
Mbuti Pygmy	Africa	15	2	0	1	
San	Africa	7	1	0	0	
Yoruba	Africa	25	2	2	0	
French	Europe	29	0	1	0	
Orcadian	Europe	16	1	0	0	
Bedouin	Middle East	48	1	0	1	
Druze	Middle East	47	1	2	2	
Mozabite	Middle East	30	0	1	0	
Palestinian	Middle East	51	0	1	5	
Balochi	Central/South Asia	25	0	1	0	
Hazara	Central/South Asia	24	0	1	1	
Kalash	Central/South Asia	25	1	0	1	
Sindhi	Central/South Asia	25	1	0	0	
Cambodian	East Asia	11	1	0	0	
Lahu	East Asia	10	1	1	0	
Naxi	East Asia	10	0	1	0	
Oroqen	East Asia	10	0	1	0	
Melanesian	Oceania	19	9	3	2	
Colombian	America	13	6	1	0	
Karitiana	America	24	6	6	0	
Maya	America	25	2	1	2	
Pima	America	25	15	6	10	
Surui	America	21	15	14	0	

Table modified from Supplementary Tables 16 and 19 of *Rosenberg (2006)*.



Table 2.3: Mean squared error (MSE), variance, and bias squared of estimates for data simulated based on allele frequencies at two loci (AAT263P and ACT3F12)

		AAT263P ( $H = 0.6778$ )			ACT3F12 ( $H = 0.8263$ )			
$m$	$(q, r, s)$	Estimator	MSE	Variance	Bias <sup>2</sup>	MSE	Variance	Bias <sup>2</sup>
10	(2, 0, 0)	$\hat{H}_{full}$	<b>9.196</b> $\times 10^{-3}$	<b>9.141</b> $\times 10^{-3}$	$5.454 \times 10^{-5}$	$3.774 \times 10^{-3}$	<b>3.694</b> $\times 10^{-3}$	$7.923 \times 10^{-5}$
		$\tilde{H}_{full}$	$9.337 \times 10^{-3}$	$9.337 \times 10^{-3}$	<b>6.387</b> $\times 10^{-8}$	<b>3.773</b> $\times 10^{-3}$	$3.773 \times 10^{-3}$	<b>4.249</b> $\times 10^{-8}$
		$\hat{H}_{reduced}$	$9.911 \times 10^{-3}$	$9.911 \times 10^{-3}$	$9.160 \times 10^{-8}$	$4.034 \times 10^{-3}$	$4.034 \times 10^{-3}$	$1.110 \times 10^{-7}$
	(2, 2, 0)	$\hat{H}_{full}$	<b>1.084</b> $\times 10^{-2}$	<b>1.064</b> $\times 10^{-2}$	$2.047 \times 10^{-4}$	$4.692 \times 10^{-3}$	<b>4.390</b> $\times 10^{-3}$	$3.020 \times 10^{-4}$
		$\tilde{H}_{full}$	$1.110 \times 10^{-2}$	$1.110 \times 10^{-2}$	<b>1.542</b> $\times 10^{-9}$	<b>4.581</b> $\times 10^{-3}$	$4.581 \times 10^{-3}$	$2.562 \times 10^{-10}$
		$\hat{H}_{reduced}$	$1.385 \times 10^{-2}$	$1.385 \times 10^{-2}$	$6.957 \times 10^{-9}$	$5.899 \times 10^{-3}$	$5.899 \times 10^{-3}$	<b>1.595</b> $\times 10^{-10}$
	(2, 0, 2)	$\hat{H}_{full}$	<b>9.885</b> $\times 10^{-3}$	<b>9.777</b> $\times 10^{-3}$	$1.078 \times 10^{-4}$	$4.236 \times 10^{-3}$	<b>4.066</b> $\times 10^{-3}$	$1.706 \times 10^{-4}$
		$\tilde{H}_{full}$	$1.009 \times 10^{-2}$	$1.009 \times 10^{-2}$	$1.048 \times 10^{-7}$	<b>4.197</b> $\times 10^{-3}$	$4.197 \times 10^{-3}$	<b>1.855</b> $\times 10^{-10}$
		$\hat{H}_{reduced}$	$1.363 \times 10^{-2}$	$1.363 \times 10^{-2}$	<b>5.363</b> $\times 10^{-8}$	$5.839 \times 10^{-3}$	$5.839 \times 10^{-3}$	$3.014 \times 10^{-9}$
(5, 2, 2)	$\hat{H}_{full}$	<b>5.107</b> $\times 10^{-3}$	<b>5.054</b> $\times 10^{-3}$	$5.273 \times 10^{-5}$	$2.030 \times 10^{-3}$	<b>1.959</b> $\times 10^{-3}$	$7.060 \times 10^{-5}$	
	$\tilde{H}_{full}$	$5.160 \times 10^{-3}$	$5.160 \times 10^{-3}$	<b>9.794</b> $\times 10^{-8}$	<b>2.000</b> $\times 10^{-3}$	$2.000 \times 10^{-3}$	<b>5.322</b> $\times 10^{-9}$	
	$\hat{H}_{reduced}$	$6.929 \times 10^{-3}$	$6.929 \times 10^{-3}$	$6.236 \times 10^{-7}$	$2.736 \times 10^{-3}$	$2.736 \times 10^{-3}$	$6.622 \times 10^{-9}$	
20	(5, 0, 0)	$\hat{H}_{full}$	<b>4.553</b> $\times 10^{-3}$	<b>4.535</b> $\times 10^{-3}$	$1.788 \times 10^{-5}$	$1.768 \times 10^{-3}$	<b>1.739</b> $\times 10^{-3}$	$2.916 \times 10^{-5}$
		$\tilde{H}_{full}$	$4.593 \times 10^{-3}$	$4.593 \times 10^{-3}$	<b>1.365</b> $\times 10^{-8}$	<b>1.762</b> $\times 10^{-3}$	$1.762 \times 10^{-3}$	$1.086 \times 10^{-8}$
		$\hat{H}_{reduced}$	$4.941 \times 10^{-3}$	$4.941 \times 10^{-3}$	$4.670 \times 10^{-8}$	$1.913 \times 10^{-3}$	$1.913 \times 10^{-3}$	<b>3.941</b> $\times 10^{-9}$
(2, 5, 2)	$\hat{H}_{full}$	<b>5.092</b> $\times 10^{-3}$	<b>5.043</b> $\times 10^{-3}$	$4.935 \times 10^{-5}$	$2.048 \times 10^{-3}$	<b>1.975</b> $\times 10^{-3}$	$7.219 \times 10^{-5}$	
	$\tilde{H}_{full}$	$5.148 \times 10^{-3}$	$5.148 \times 10^{-3}$	<b>5.843</b> $\times 10^{-9}$	<b>2.016</b> $\times 10^{-3}$	$2.016 \times 10^{-3}$	$5.047 \times 10^{-10}$	
	$\hat{H}_{reduced}$	$6.948 \times 10^{-3}$	$6.948 \times 10^{-3}$	$5.923 \times 10^{-9}$	$2.755 \times 10^{-3}$	$2.755 \times 10^{-3}$	<b>1.884</b> $\times 10^{-11}$	

		AAT263P ( $H = 0.6778$ )			ACT3F12 ( $H = 0.8263$ )			
$m$	$(q, r, s)$	Estimator	MSE	Variance	Bias <sup>2</sup>	MSE	Variance	Bias <sup>2</sup>
	(15, 0, 0)	$\hat{H}_{full}$	<b>3.580</b> × 10 <sup>-3</sup>	<b>3.548</b> × 10 <sup>-3</sup>	3.233 × 10 <sup>-5</sup>	1.396 × 10 <sup>-3</sup>	<b>1.346</b> × 10 <sup>-3</sup>	4.973 × 10 <sup>-5</sup>
		$\tilde{H}_{full}$	3.609 × 10 <sup>-3</sup>	3.609 × 10 <sup>-3</sup>	3.411 × 10 <sup>-9</sup>	<b>1.370</b> × 10 <sup>-3</sup>	1.370 × 10 <sup>-3</sup>	2.490 × 10 <sup>-9</sup>
		$\hat{H}_{reduced}$	4.924 × 10 <sup>-3</sup>	4.924 × 10 <sup>-3</sup>	<b>2.990</b> × 10 <sup>-10</sup>	1.903 × 10 <sup>-3</sup>	1.903 × 10 <sup>-3</sup>	<b>2.346</b> × 10 <sup>-9</sup>
30	(5, 5, 5)	$\hat{H}_{full}$	<b>3.370</b> × 10 <sup>-3</sup>	<b>3.345</b> × 10 <sup>-3</sup>	2.464 × 10 <sup>-5</sup>	1.294 × 10 <sup>-3</sup>	<b>1.260</b> × 10 <sup>-3</sup>	3.525 × 10 <sup>-5</sup>
		$\tilde{H}_{full}$	3.393 × 10 <sup>-3</sup>	3.393 × 10 <sup>-3</sup>	3.169 × 10 <sup>-8</sup>	<b>1.278</b> × 10 <sup>-3</sup>	1.278 × 10 <sup>-3</sup>	<b>2.062</b> × 10 <sup>-9</sup>
		$\hat{H}_{reduced}$	4.930 × 10 <sup>-3</sup>	4.930 × 10 <sup>-3</sup>	<b>1.154</b> × 10 <sup>-8</sup>	1.890 × 10 <sup>-3</sup>	1.890 × 10 <sup>-3</sup>	2.397 × 10 <sup>-8</sup>
	(0, 5, 5)	$\hat{H}_{full}$	<b>2.970</b> × 10 <sup>-3</sup>	<b>2.962</b> × 10 <sup>-3</sup>	7.105 × 10 <sup>-6</sup>	1.122 × 10 <sup>-3</sup>	<b>1.110</b> × 10 <sup>-3</sup>	1.181 × 10 <sup>-5</sup>
		$\tilde{H}_{full}$	2.988 × 10 <sup>-3</sup>	2.988 × 10 <sup>-3</sup>	<b>4.302</b> × 10 <sup>-8</sup>	<b>1.119</b> × 10 <sup>-3</sup>	1.119 × 10 <sup>-3</sup>	4.230 × 10 <sup>-9</sup>
		$\hat{H}_{reduced}$	3.623 × 10 <sup>-3</sup>	3.623 × 10 <sup>-3</sup>	4.632 × 10 <sup>-8</sup>	1.369 × 10 <sup>-3</sup>	1.369 × 10 <sup>-3</sup>	<b>2.294</b> × 10 <sup>-9</sup>

Sample size is indicated by  $m$ , and  $q$ ,  $r$ , and  $s$  represent the numbers of parent-offspring, full-sib, and second-degree pairs, respectively. Each value is based on 100,000 simulated datasets, and the same simulated datasets were used for all estimators and for all three quantities (MSE, variance, bias squared). We use  $\hat{H}_{full}$  and  $\tilde{H}_{full}$  to denote  $\hat{H}$  and  $\tilde{H}$  applied to a sample of  $m$  individuals. For  $\hat{H}$  applied to a sample of  $m$  individuals in which  $q + r + s$  related individuals are removed to create a sample of  $m - q - r - s$  individuals, we use the notation  $\hat{H}_{reduced}$ . Boldface type indicates the estimator with the smallest MSE, variance, or bias squared.

Table 2.4: Statistical tests applied to the mean gene diversity across loci

	$p$ -value for Wilcoxon signed rank test	Mean of $H_{reduced} - H_{full}$ across populations	Mean of $ H_{reduced} - H_{full} $ across populations	Fraction of populations $H_{full} > H_{reduced}$	of with
$\hat{H}_{952}$ vs. $\hat{H}_{1048}$	$8.804 \times 10^{-6}$	$2.387 \times 10^{-3}$	$2.387 \times 10^{-3}$	0	
$\hat{H}_{952}$ vs. $\tilde{H}_{1048}$	0.703	$3.262 \times 10^{-4}$	$6.660 \times 10^{-4}$	0.226	
$\hat{H}_{507}$ vs. $\hat{H}_{603}$	$2.980 \times 10^{-8}$	$4.866 \times 10^{-3}$	$4.866 \times 10^{-3}$	0	
$\hat{H}_{507}$ vs. $\tilde{H}_{603}$	0.708	$6.649 \times 10^{-4}$	$1.358 \times 10^{-3}$	0.462	

In the header line,  $H_{reduced}$  refers to  $\hat{H}_{952}$  or  $\hat{H}_{507}$  depending on which estimator is being considered; similarly,  $H_{full}$  refers to  $\hat{H}_{1048}$ ,  $\tilde{H}_{1048}$ ,  $\hat{H}_{603}$ , or  $\tilde{H}_{603}$ .

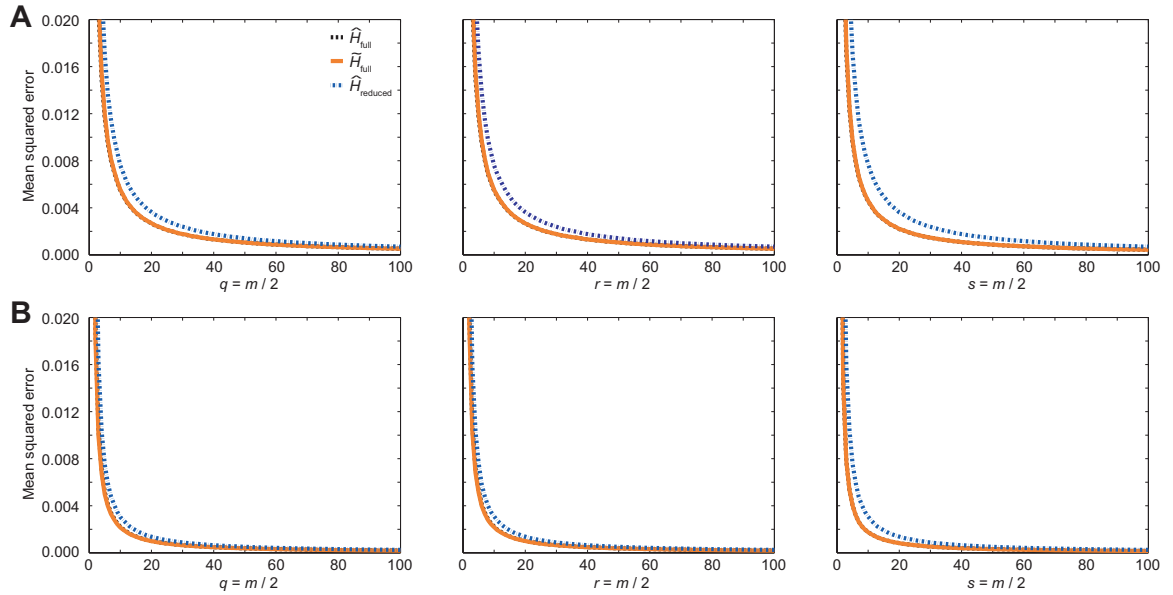


Figure 2.1: Mean squared error (MSE) as a function of sample size  $m$  for three different estimators. Each plot in a given row represents samples with a different type of relative pair. The numbers of parent-offspring, full-sib, and second-degree pairs are denoted by  $q$ ,  $r$ , and  $s$ , respectively. The full and reduced samples contain  $m$  and  $m/2$  individuals, respectively. The  $\tilde{H}_{full}$  curve is almost directly on top of the  $\hat{H}_{full}$  curve. **A.** Allele frequencies simulated based on observed frequencies at locus AAT263P ( $H = 0.6778$ ). **B.** Allele frequencies simulated based on observed frequencies at locus ACT3F12 ( $H = 0.8263$ ). The range of the plots is truncated at 0.02, so that the MSE for small sample sizes is not shown. Each point in the graphs is based on 100,000 simulated data sets, and the same simulated data were used for all three estimators.

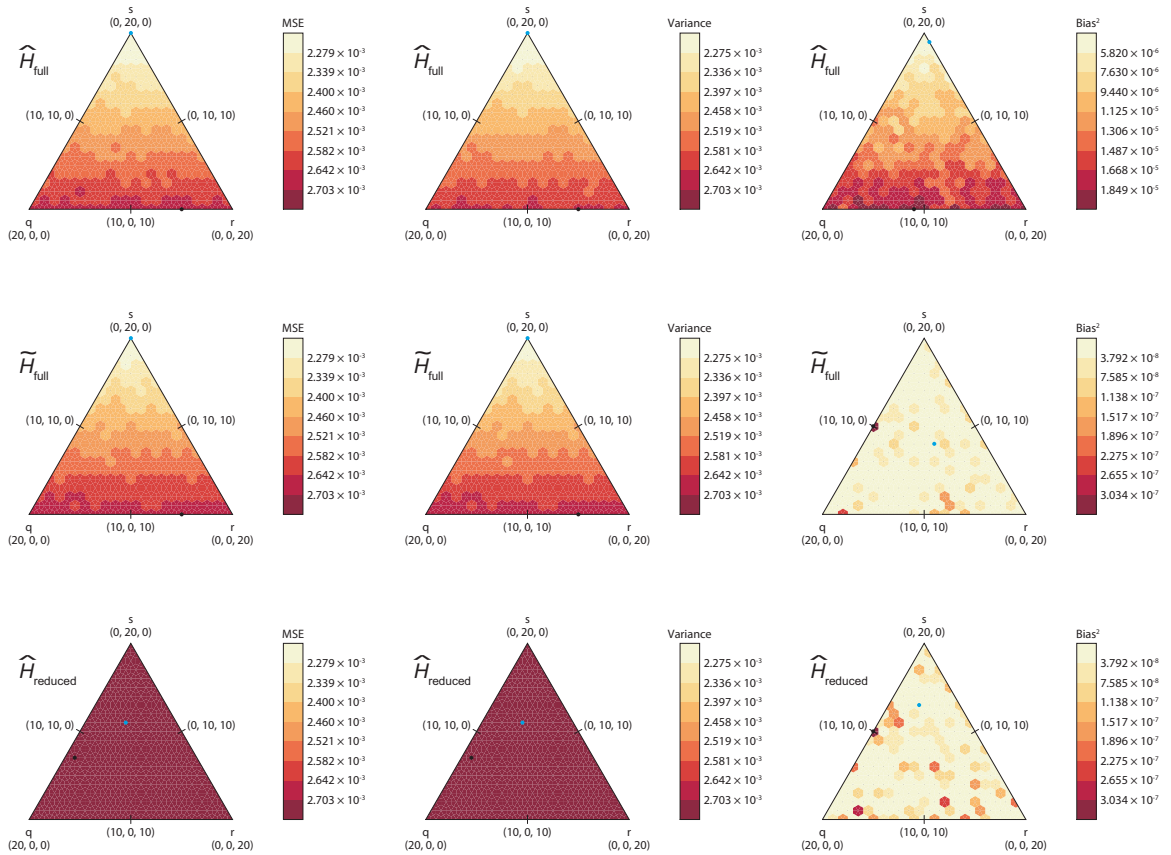


Figure 2.2: Heat maps of simulated mean squared error (MSE), variance, and bias squared for each estimator applied to a full sample of 40 and a reduced sample of 20 individuals, as functions of the mixture of types of relative pairs included in the sample. The simulation was based on allele frequencies at the AAT263P locus ( $H = 0.6778$ ). The sample of 40 individuals includes  $q$  parent-offspring,  $r$  full-sib, and  $s$  second-degree pairs. The three vertices correspond to samples that contain either all parent-offspring, all full-sib, or all second-degree pairs. Moving horizontally along the triangle changes the numbers of parent-offspring and full-sib pairs in the sample, and moving vertically changes the number of second-degree pairs. The numbers indicated on the scale are the cutoff values for each color. Each row of triangles represents a different estimator, and each column represents a different statistic. Blue and black dots represent the points at which the smallest and largest values occur in each triangle, respectively. Each point in the graphs is based on 100,000 simulated data sets, and the same simulations were used for all three estimators.

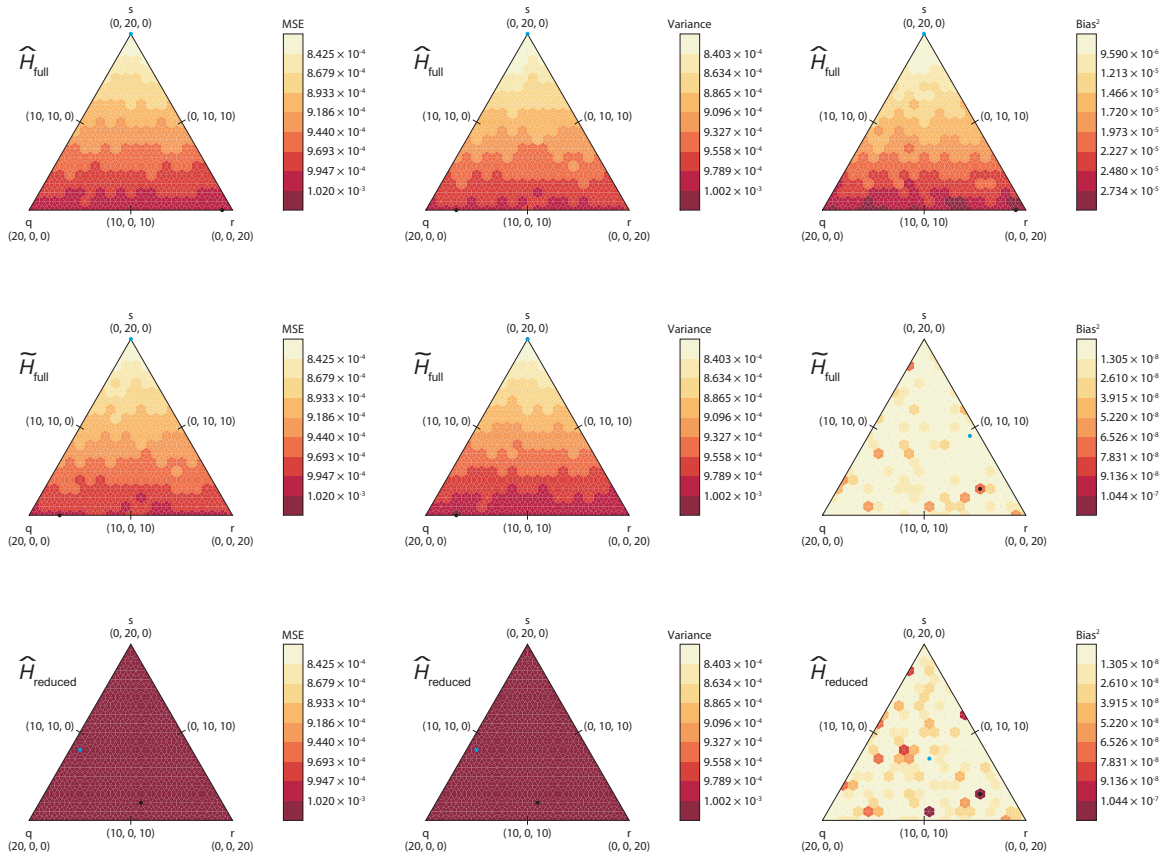


Figure 2.3: Heat maps of simulated mean squared error (MSE), variance, and bias squared for each estimator applied to a full sample of 40 and a reduced sample of 20 individuals, as functions of the mixture of types of relative pairs included in the sample. The simulation was based on allele frequencies at the ACT3F12 locus ( $H = 0.8263$ ). See Figure 2.2 caption for additional details.

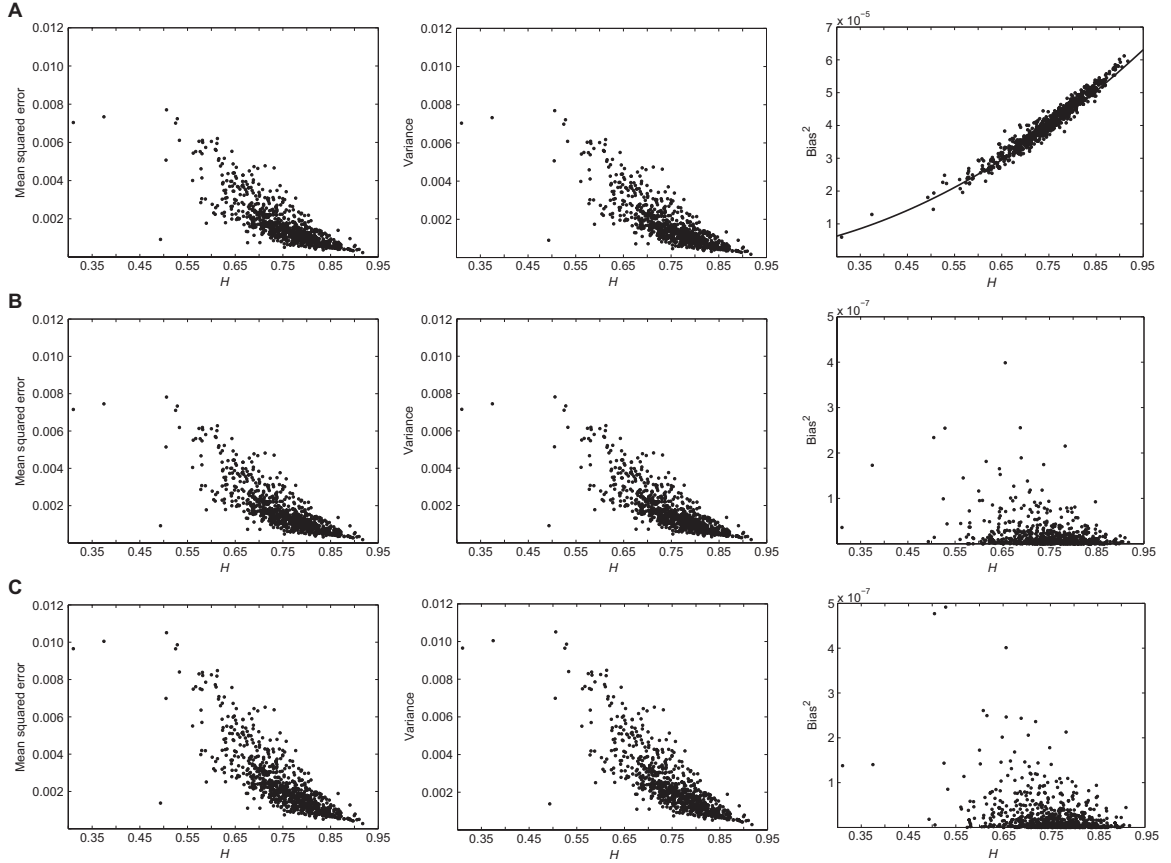


Figure 2.4: Mean squared error (MSE), variance, and bias squared for each estimator applied to a full sample of 30 and a reduced sample of 15 individuals, as functions of parametric gene diversity, considering simulated values based on each of the 783 loci. The simulations incorporated 30 individuals in 15 parent-offspring pairs. **A.**  $\hat{H}_{full}$ . A quadratic regression of bias squared on  $H$  (with the constant and linear terms forced to be 0) is given by  $(7.187 \times 10^{-5})H^2$ , with  $R^2 = 0.959$ . The Spearman correlation coefficient is  $-0.8364$  for  $H$  and MSE and  $-0.8394$  for  $H$  and variance. **B.**  $\tilde{H}_{full}$ . The Spearman correlation coefficient is  $-0.8394$  for  $H$  and MSE and  $-0.8394$  for  $H$  and variance. **C.**  $\hat{H}_{reduced}$ . The Spearman correlation coefficient is  $-0.8447$  for  $H$  and MSE and  $-0.8447$  for  $H$  and variance. Each point in the graphs is based on 100,000 simulated data sets, and the same simulations were used for all three estimators.

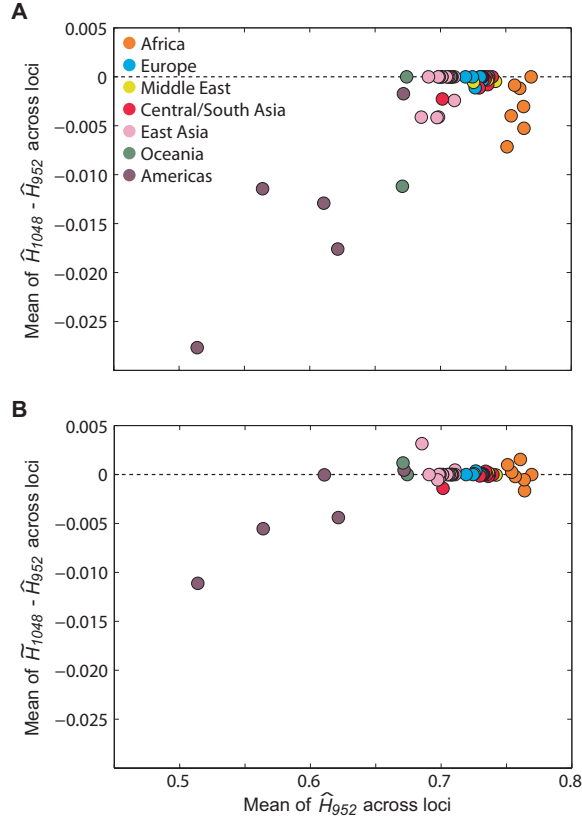


Figure 2.5: Comparison of the mean of  $\hat{H}_{1048} - \hat{H}_{952}$  and the mean of  $\tilde{H}_{1048} - \tilde{H}_{952}$ . Each population is represented by a point colored based on the geographic location of the population, and the dotted line represents zero difference between the full-data estimator and  $\hat{H}_{952}$ . Since 27 of the 53 populations do not contain related individuals, the gene diversities given by  $\hat{H}_{1048}$  and  $\tilde{H}_{1048}$  are the same for these populations. **A.** The mean of  $\hat{H}_{1048} - \hat{H}_{952}$ , displaying a reduction of  $\hat{H}$  when applied to samples containing related individuals. **B.** The mean of  $\tilde{H}_{1048} - \hat{H}_{952}$ , displaying a decrease in the magnitude of the difference between the full-data estimator and  $\hat{H}_{952}$ .



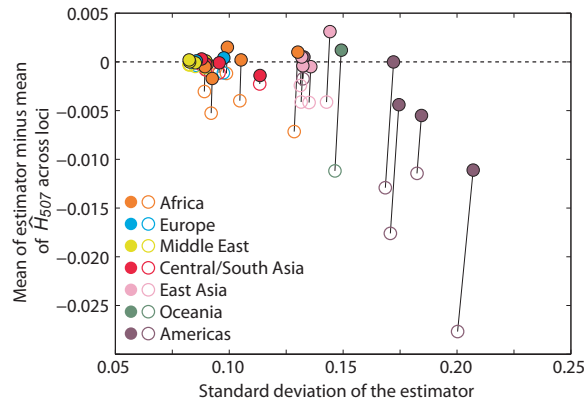


Figure 2.6: Comparison of the mean difference of an estimator ( $\hat{H}_{603}$  or  $\tilde{H}_{603}$ ) from  $\hat{H}_{507}$  with the standard deviation of the estimator. Each population is represented by a point colored based on the geographic location of the population. Open and filled circles represent the estimates for  $\hat{H}_{603}$  and  $\tilde{H}_{603}$ , respectively.

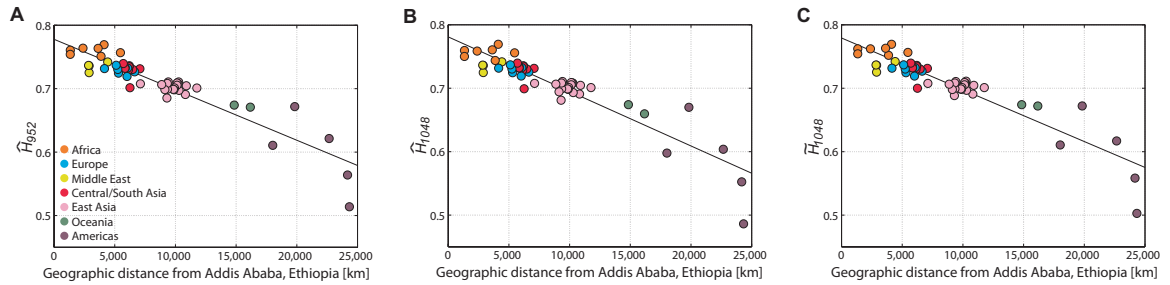


Figure 2.7: Gene diversity vs. geographic distance from Addis Ababa, Ethiopia. **A.**  $\hat{H}_{952}$  vs. distance from Addis Ababa. The linear regression is given by  $H = 0.7778 - (7.955 \times 10^{-6}) \times \text{distance}$ , with  $R^2 = 0.856$ . **B.**  $\hat{H}_{1048}$  vs. distance from Addis Ababa. The linear regression is given by  $H = 0.7809 - (8.595 \times 10^{-6}) \times \text{distance}$ , with  $R^2 = 0.844$ . **C.**  $\tilde{H}_{1048}$  vs. distance from Addis Ababa. The linear regression is given by  $H = 0.7792 - (8.161 \times 10^{-6}) \times \text{distance}$ , with  $R^2 = 0.849$ .

## CHAPTER III

# Unbiased estimation of gene diversity in samples containing related individuals: exact variance and arbitrary ploidy

### 3.1 Introduction

For a given locus, gene diversity, also known as expected heterozygosity, characterizes the proportion of heterozygous genotypes expected in a population under Hardy-Weinberg equilibrium (*Nei, 1973*). *Nei and Roychoudhury (1974)* devised an estimator of gene diversity that is unbiased for random samples of unrelated, non-inbred individuals. When inbred individuals or close relatives are included in a sample, however, this estimator has a downward bias (*Weir, 1989; DeGiorgio and Rosenberg, 2009*). To account for the effects of inbreeding in a sample of diploid individuals, *Weir (1989, 1996)* derived the expected value of gene diversity, producing an unbiased estimator of gene diversity that makes use of the mean inbreeding coefficient across sampled individuals, where the inbreeding coefficient of an individual is defined as the probability for a randomly chosen locus that the two alleles of the individual are inherited identically by descent from a common ancestor. Using the mean kinship coefficient across pairs of sampled individuals, *DeGiorgio and Rosenberg (2009)* extended this estimator to account for the bias produced in samples

containing close relatives, where the kinship coefficient between two individuals,  $j$  and  $k$ , is defined as the probability that an allele randomly selected from individual  $j$  at a random locus and an allele randomly selected from individual  $k$  at the same locus are identical by descent (IBD).

The *DeGiorgio and Rosenberg* (2009) estimator is useful for autosomal markers in samples from diploid organisms that contain related or inbred individuals. However, in studying gene diversity among related individuals in non-diploid cases (e.g., *Buteler et al.* (1999)) or in cases of mixed ploidy, such as in the analysis of sex chromosomes (e.g., *Reiland et al.* (2002)), unbiasedness for this estimator has not been demonstrated. Here, we extend the *DeGiorgio and Rosenberg* (2009) estimator of gene diversity to account for situations in which known related and inbred individuals are included in a sample and in which the sample contains an arbitrary mixture of individuals of different ploidy. We use a more general method to obtain the estimator than the method used for diploids by *DeGiorgio and Rosenberg* (2009), and we show that the general estimator reduces to the *DeGiorgio and Rosenberg* (2009) estimator in the diploid case. We also derive a formula for the variance of our estimator,  $\tilde{H}$ , to facilitate evaluation of the statistical properties of the estimator. This variance formula, which is a function of identity states among individuals, includes terms that involve identity by descent among two, three, and four individuals, and among pairs of pairs of individuals. Our variance function is convenient because extensive work on IBD probabilities among individuals (*Cotterman*, 1940; *Harris*, 1964; *Gillois*, 1965; *Cockerham*, 1971; *Jacquard*, 1974; *Thompson*, 1974; *Lange*, 2002) has provided a framework for calculating the quantities incorporated in the formula.

Using the variance formula, we examine the performance of our estimator in scenarios involving the human X chromosome, for which males and females, who might both be included in a typical sample, differ in ploidy. In our evaluations, we first show that the exact theoretical values of the variance, which are obtained

from a quite complex formula, are closely matched by simulations. We also validate that when each sampled individual is related to at most one other individual in the sample, the exact theoretical variance can be approximated well by a simpler formula. Using the variance approximation and simulations, we compare the behavior of our estimator to that of the *Nei and Roychoudhury* (1974) estimator, which does not account for relatives. We then analyze human SNPs from the X chromosome and find that  $\tilde{H}$  also performs well in practice.

### 3.2 Theory

Consider a sample of  $g$  groups, each with different ploidy (e.g., haploid males and diploid females on the human X chromosome). Suppose that the sample from group  $b$  contains  $n_b$   $m_b$ -ploid individuals,  $b = 1, 2, \dots, g$ . Further, let  $(b, k)$ ,  $k = 1, 2, \dots, n_b$ , denote individual  $k$  from group  $b$ . The number of copies of allelic type  $i$  in individual  $k$  from group  $b$  is

$$X_{(b,k)}^{(i)} = \sum_{\ell=1}^{m_b} A_{(b,k),\ell}^{(i)}, \quad (3.1)$$

where  $A_{(b,k),\ell}^{(i)}$  is an indicator random variable that takes on the value 1 if the  $\ell$ th allele in individual  $(b, k)$  has type  $i$  and that equals 0 otherwise.

Note that  $\mathbb{E}\left[A_{(b,k),\ell}^{(i)}\right] = p_i$ , where  $p_i$  is the frequency of allelic type  $i$  in the population. We can then define an unbiased estimator for the frequency of allele  $i$  as

$$\hat{p}_i = \frac{1}{\sum_{b=1}^g n_b m_b} \sum_{b=1}^g \sum_{k=1}^{n_b} X_{(b,k)}^{(i)}. \quad (3.2)$$

Rewriting the estimator of *Nei and Roychoudhury* (1974) for the mixed-ploidy case, if no inbred or related individuals are included in the sample, then an unbiased estimator of gene diversity is

$$\hat{H} = \frac{\sum_{b=1}^g n_b m_b}{(\sum_{b=1}^g n_b m_b) - 1} \left( 1 - \sum_{i=1}^I \hat{p}_i^2 \right). \quad (3.3)$$

If inbred or related individuals are included in the sample, then  $\hat{H}$  is a biased estimator of  $H = 1 - \sum_{i=1}^I p_i^2$ . We follow the approach of *DeGiorgio and Rosenberg* (2009), correcting for this bias by first obtaining the variance of sample allele frequencies. However, we use a different method here for obtaining the variance of sample allele frequencies, determining the bias correction for diploids as a special case of a more general computation.

### 3.2.1 An unbiased estimator

Suppose we have four possibly, but not necessarily, distinct individuals  $(a, j)$ ,  $(b, k)$ ,  $(a', j')$ , and  $(b', k')$ . Define  $\Phi_{(a,j)(b,k)}$  as the probability that two alleles randomly chosen, one from individual  $(a, j)$  and the other from individual  $(b, k)$ , are IBD. Similarly, define  $\Phi_{(a,j)(b,k)(a',j')}$  as the probability that three alleles randomly chosen, one from  $(a, j)$ , one from  $(b, k)$ , and one from  $(a', j')$ , are IBD. Define  $\Phi_{(a,j)(b,k)(a',j')(b',k')}$  as the probability that four alleles randomly chosen, one from  $(a, j)$ , one from  $(b, k)$ , one from  $(a', j')$ , and one from  $(b', k')$ , are IBD. Finally, define  $\Phi_{(a,j)(b,k),(a',j')(b',k')}$  as the joint probability that two alleles randomly chosen, one from  $(a, j)$  and the other from  $(b, k)$ , are IBD, and two alleles randomly chosen, one from  $(a', j')$  and the other from  $(b', k')$ , are IBD. These four types of probability of identity by descent are identical to the  $\theta$ ,  $\gamma$ ,  $\delta$ , and  $\Delta$  coefficients of *Cockerham* (1971), respectively. We can then define

$$\bar{\Phi}_2 = \sum_{a=1}^g \sum_{b=1}^g \sum_{j=1}^{n_a} \sum_{k=1}^{n_b} w_a w_b \bar{\Phi}_{(a,j)(b,k)} \quad (3.4)$$

$$\bar{\Phi}_3 = \sum_{a=1}^g \sum_{b=1}^g \sum_{a'=1}^g \sum_{j=1}^{n_a} \sum_{k=1}^{n_b} \sum_{j'=1}^{n_{a'}} w_a w_b w_{a'} \bar{\Phi}_{(a,j)(b,k)(a',j')} \quad (3.5)$$

$$\bar{\Phi}_4 = \sum_{a=1}^g \sum_{b=1}^g \sum_{a'=1}^g \sum_{b'=1}^g \sum_{j=1}^{n_a} \sum_{k=1}^{n_b} \sum_{j'=1}^{n_{a'}} \sum_{k'=1}^{n_{b'}} w_a w_b w_{a'} w_{b'} \bar{\Phi}_{(a,j)(b,k)(a',j')(b',k')} \quad (3.6)$$

$$\bar{\Phi}_{2,2} = \sum_{a=1}^g \sum_{b=1}^g \sum_{a'=1}^g \sum_{b'=1}^g \sum_{j=1}^{n_a} \sum_{k=1}^{n_b} \sum_{j'=1}^{n_{a'}} \sum_{k'=1}^{n_{b'}} w_a w_b w_{a'} w_{b'} \bar{\Phi}_{(a,j)(b,k)(a',j')(b',k')} , \quad (3.7)$$

as weighted mean kinship coefficients across all sets of pairs, triples, quartets, and pairs of pairs of individuals. The weight associated with an individual in group  $x$ ,  $w_x = m_x / \sum_{b=1}^g n_b m_b$ , is proportional to the ploidy associated with the group. Define the inbreeding coefficient for individual  $(b, k)$ , denoted by  $f_{(b,k)}$ , as the probability that two alleles randomly chosen without replacement from individual  $(b, k)$  are IBD, and let  $\bar{f}_b = (1/n_b) \sum_{k=1}^{n_b} f_{(b,k)}$  be the mean inbreeding coefficient across individuals in group  $b$ . This definition reduces to the standard definition for the diploid case.

In this section we first present two equations (eqs. 3.8 and 3.9) that aid in the development of a generalized estimator of gene diversity (Theorem III.1). This general estimator, the main result of this section, corrects the bias created by the inclusion of related and inbred individuals in a sample consisting of individuals with any mixture of ploidy. Using this estimator, we provide generalizations of results presented by *DeGiorgio and Rosenberg (2009)* for diploids to the case of arbitrary ploidy (eqs. 3.13 and 3.14) and we show how these generalizations can be reduced to the diploid case.

Consider a locus with  $I$  distinct alleles, allele frequencies  $p_i \in [0, 1]$ , and  $\sum_{i=1}^I p_i = 1$ . Suppose a sample from a population has  $g$  groups, each with different ploidy, and  $n_b$   $m_b$ -ploid individuals in group  $b$ ,  $b = 1, 2, \dots, g$ , each of whom is possibly inbred

and related to other individuals in the sample. Consider the  $\ell$ th allele of individual  $(a, j)$  and the  $t$ th allele of individual  $(b, k)$ . By definition of expected value, we have

$$\begin{aligned}\mathbb{E}\left[A_{(a,j),\ell}^{(i)}A_{(b,k),t}^{(i)}\right] &= \mathbb{P}\left[A_{(a,j),\ell}^{(i)} = 1, A_{(b,k),t}^{(i)} = 1\right] \\ &= \Phi_{(a,j)(b,k)}p_i + (1 - \Phi_{(a,j)(b,k)})p_i^2 \\ &= \Phi_{(a,j)(b,k)}p_i(1 - p_i) + p_i^2.\end{aligned}\tag{3.8}$$

In taking the expected value of our estimator of gene diversity, we will need to evaluate the quantity  $\mathbb{E}[\widehat{p}_i^2]$ . Using eq. 3.8, we show in Appendix A that

$$\mathbb{E}[\widehat{p}_i^2] = \overline{\Phi}_2 p_i(1 - p_i) + p_i^2.\tag{3.9}$$

Plugging eqs. 3.8 and 3.9 into  $Var[\widehat{p}_i] = \mathbb{E}[\widehat{p}_i^2] - (\mathbb{E}[\widehat{p}_i])^2$  yields  $Var[\widehat{p}_i] = \overline{\Phi}_2 p_i(1 - p_i)$ , which reduces to the result presented for the diploid case in eq. 7 of *DeGiorgio and Rosenberg* (2009), by reduction of the definition of  $\overline{\Phi}_2$  for the diploid case. The following theorem provides a generalized unbiased estimator of gene diversity when a sample with any mixture of ploidy contains related or inbred individuals.

**Theorem III.1.** *Consider a locus with  $I$  distinct alleles, allele frequencies  $p_i \in [0, 1]$  and  $\sum_{i=1}^I p_i = 1$ . Suppose a sample from a population has  $g$  groups, each with different ploidy, and  $n_b$   $m_b$ -ploid individuals in group  $b$ ,  $b = 1, 2, \dots, g$ , each of whom is possibly inbred and related to other individuals in the sample. Then*

$$\widetilde{H} = \frac{1}{1 - \overline{\Phi}_2} \left(1 - \sum_{i=1}^I \widehat{p}_i^2\right)\tag{3.10}$$

*is an unbiased estimator for gene diversity.*

The proof that  $\widetilde{H}$  is unbiased follows that of Proposition 1 in *DeGiorgio and Rosenberg* (2009), substituting the more general  $\overline{\Phi}_2$  in place of the corresponding



mean kinship coefficient in the earlier proof.

When reducing the definition of  $\bar{\Phi}_2$  for the diploid case studied by *DeGiorgio and Rosenberg (2009)*, the result in Theorem III.1 is identical to the result presented for this case in Proposition 1 of *DeGiorgio and Rosenberg (2009)*. One interesting consequence of Theorem III.1 is that  $\tilde{H}$  has a simple representation in terms of the sample probability of identity-by-state and the probability of identity-by-descent computed based on assumed levels of inbreeding and relationship. This representation is

$$\tilde{H} = \frac{1 - \hat{\mathbb{P}}[\text{IBS}]}{1 - \mathbb{P}[\text{IBD}]}, \quad (3.11)$$

where  $\hat{\mathbb{P}}[\text{IBS}]$  is the probability that two alleles in the sample, chosen uniformly at random with replacement, are identical by state, and  $\mathbb{P}[\text{IBD}]$  is the probability that two alleles in the sample, chosen uniformly at random with replacement, are identical by descent. A proof that eq. 3.11 is a consequence of eq. 3.10 is provided in Appendix A. Note that eqs. 3.10 and 3.11 have a connection to estimators of relatedness in a context in which relatedness is unknown. Such estimators essentially invert equations similar to eq. 3.11 to get estimators of  $\bar{\Phi}_2$  (*Ritland, 1996; Rousset, 2002*).

We next seek to transform the estimator in eq. 3.10 into one that is more convenient for data analysis. Let  $\mathcal{G}_{a,b}$ ,  $a, b = 1, 2, \dots, g$ , be the set of distinct types of relative pairs for pairs of distinct individuals in a sample, one from group  $a$  and one from group  $b$ . Let  $\eta_R$  be the number of pairs of individuals with relationship type  $R$  in  $\mathcal{G}_{a,b}$ , and let  $\Phi_R$  be the kinship coefficient for each of these pairs. Then, as shown in Appendix A, we can write  $\bar{\Phi}_2$  as

$$\bar{\Phi}_2 = \frac{1}{\left(\sum_{b=1}^g n_b m_b\right)^2} \left[ \sum_{b=1}^g n_b m_b + \sum_{b=1}^g n_b m_b (m_b - 1) \bar{f}_b + 2 \sum_{b=1}^g \sum_{R \in \mathcal{G}_{b,b}} m_b^2 \eta_R \Phi_R + 2 \sum_{a=1}^{g-1} \sum_{b=a+1}^g \sum_{R \in \mathcal{G}_{a,b}} m_a m_b \eta_R \Phi_R \right]. \quad (3.12)$$

This version of  $\bar{\Phi}_2$  is convenient for computation. To obtain a formula for  $\tilde{H}$  that is convenient for computation and that is a generalized version of an analogous quantity for the diploid case in eq. 9 of *DeGiorgio and Rosenberg* (2009), we can substitute eqs. 3.3 and 3.12 into eq. 3.10 to get

$$\tilde{H} = \frac{\left(\sum_{b=1}^g n_b m_b\right) \left(\sum_{b=1}^g n_b m_b - 1\right)}{D} \hat{H}, \quad (3.13)$$

where

$$\begin{aligned} D = & \left(\sum_{b=1}^g n_b m_b\right) \left(\sum_{b=1}^g n_b m_b - 1\right) - \sum_{b=1}^g n_b m_b (m_b - 1) \bar{f}_b - 2 \sum_{b=1}^g \sum_{R \in \mathcal{G}_{b,b}} m_b^2 \eta_R \Phi_R \\ & - 2 \sum_{a=1}^{g-1} \sum_{b=a+1}^g \sum_{R \in \mathcal{G}_{a,b}} m_a m_b \eta_R \Phi_R. \end{aligned}$$

A proof of eq. 3.13 is provided in Appendix A. We note that by using  $g = 1$ ,  $n_1 = n$ , and  $m_1 = 2$  in eq. 3.13, we obtain eq. 9 of *DeGiorgio and Rosenberg* (2009).

Note that  $\tilde{H} = c\hat{H}$ , where

$$c = \frac{\left(\sum_{b=1}^g n_b m_b\right) \left(\sum_{b=1}^g n_b m_b - 1\right)}{D}.$$

By rearranging and taking the expected value, we get  $\mathbb{E}[\hat{H}] = \mathbb{E}[\tilde{H}]/c = H/c$ . Therefore,

$$\begin{aligned}
bias(\hat{H}) &= \frac{1-c}{c}H \\
&= -\frac{1}{\left(\sum_{b=1}^g n_b m_b\right)\left(\sum_{b=1}^g n_b m_b - 1\right)} \left[ \sum_{b=1}^g n_b m_b (m_b - 1) \bar{f}_b \right. \\
&\quad + 2 \sum_{b=1}^g \sum_{R \in \mathcal{G}_{b,b}} m_b^2 \eta_R \Phi_R \\
&\quad \left. + 2 \sum_{a=1}^{g-1} \sum_{b=a+1}^g \sum_{R \in \mathcal{G}_{a,b}} m_a m_b \eta_R \Phi_R \right] H.
\end{aligned} \tag{3.14}$$

Equation 3.14 is a generalized version of the bias formula in the diploid case, in eq. 11 of *DeGiorgio and Rosenberg (2009)*. The bias is always negative and it has a magnitude that increases linearly with respect to  $H$ . Using  $g = 1$ ,  $n_1 = n$ , and  $m_1 = 2$  in eq. 3.14, we obtain eq. 11 of *DeGiorgio and Rosenberg (2009)*.

### 3.2.2 Variance of the estimator

In the previous section, we derived an unbiased estimator  $\tilde{H}$  of gene diversity in a sample of arbitrary ploidy. It is useful to determine the variance of the estimator, a quantity that in the diploid case *DeGiorgio and Rosenberg (2009)* obtained only by simulation. The following theorem provides a formula for the variance of the generalized estimator of gene diversity in samples with any mixture of ploidy.

**Theorem III.2.** *Consider a locus with  $I$  distinct alleles, allele frequencies  $p_i \in [0, 1]$  and  $\sum_{i=1}^I p_i = 1$ . Suppose a sample from a population has  $g$  groups, each with different ploidy, and  $n_b$   $m_b$ -ploid individuals in group  $b$ ,  $b = 1, 2, \dots, g$ , each of whom is possibly inbred and related to other individuals in the sample. Then the variances of the  $\tilde{H}$  and  $\hat{H}$  estimators of gene diversity are*

$$\text{Var}[\tilde{H}] = \frac{1}{(1 - \bar{\Phi}_2)^2} \text{Var} \left[ 1 - \sum_{i=1}^I \hat{p}_i^2 \right] \quad (3.15)$$

and

$$\text{Var}[\hat{H}] = \left[ \frac{\sum_{b=1}^g n_b m_b}{(\sum_{b=1}^g n_b m_b) - 1} \right]^2 \text{Var} \left[ 1 - \sum_{i=1}^I \hat{p}_i^2 \right], \quad (3.16)$$

where

$$\begin{aligned} \text{Var} \left[ 1 - \sum_{i=1}^I \hat{p}_i^2 \right] &= \bar{\Phi}_{2,2} - \bar{\Phi}_2^2 + 2[\bar{\Phi}_2^2 - \bar{\Phi}_4] \sum_{i=1}^I p_i^2 + 4[2\bar{\Phi}_4 + \bar{\Phi}_2 - 2\bar{\Phi}_3 - \bar{\Phi}_{2,2}] \sum_{i=1}^I p_i^3 \\ &\quad + [3\bar{\Phi}_{2,2} + 8\bar{\Phi}_3 - 6\bar{\Phi}_4 - 4\bar{\Phi}_2 - \bar{\Phi}_2^2] \left( \sum_{i=1}^I p_i^2 \right)^2. \end{aligned} \quad (3.17)$$

The proof of Theorem III.2 is long and is provided in Appendix B.

We next derive an approximate formula that in our calculations below, we use in place of eq. 3.17 inside of eqs. 3.15 and 3.16. The approximation is based only on pairwise kinship coefficients, and is useful in cases in which the number of relatives in a sample is small enough that no individual is related to more than one other sampled individual. In such cases, the only nonzero terms included in  $\bar{\Phi}_3$ ,  $\bar{\Phi}_4$ , and  $\bar{\Phi}_{2,2}$  all involve sampling the same individual or pairs of individuals more than once. Thus, the  $\bar{\Phi}_3$ ,  $\bar{\Phi}_4$ , and  $\bar{\Phi}_{2,2}$  terms, along with  $\bar{\Phi}_2^2$ , are ignored, as they are likely to be much smaller than  $\bar{\Phi}_2$  in cases in which the number of relationships in the sample is small.

In addition to the assumptions listed in Theorem III.2, suppose that each individual in the sample is related to no more than one other individual in the sample. If we ignore terms involving  $(\sum_{b=1}^g m_b n_b)^{-k}$ ,  $k > 1$ , then terms involving  $\bar{\Phi}_2^2$ ,  $\bar{\Phi}_3$ ,  $\bar{\Phi}_4$ , and  $\bar{\Phi}_{2,2}$  in eq. 3.17 can be ignored. The only terms in eq. 3.17 that we retain are those of order  $(\sum_{b=1}^g m_b n_b)^0$  and  $(\sum_{b=1}^g m_b n_b)^{-1}$ .  $\bar{\Phi}_2$  is of order  $(\sum_{b=1}^g m_b n_b)^{-1}$ . Therefore, reducing eq. 3.17 leads to

$$\text{Var} \left[ 1 - \sum_{i=1}^I \widehat{p}_i^2 \right] \approx 4\overline{\Phi}_2 \left[ \sum_{i=1}^I p_i^3 - \left( \sum_{i=1}^I p_i^2 \right)^2 \right]. \quad (3.18)$$

This formula is an approximation to eq. 3.17 when the number of relatives in a sample is small enough that no individual is related to more than one other sampled individual.

We now show that when no related individuals are included in a sample of diploids, the variance in eq. 3.18 is exactly the formula given by *Weir* (1989). Suppose a sample from a diploid population consists of  $n$  unrelated, but possibly inbred, individuals and further suppose that we ignore terms involving  $n^{-k}$ ,  $k > 1$ . Then  $\Phi_{kk} = (1/2)(1 + f_k)$ , where  $f_k$  is the inbreeding coefficient for individual  $k$ . We can write the mean pairwise kinship coefficient as

$$\overline{\Phi}_2 = \frac{1}{n^2} \sum_{k=1}^n \Phi_{kk} = \frac{1}{n^2} \sum_{k=1}^n \frac{1}{2}(1 + f_k) = \frac{1}{2n}(1 + \bar{f}),$$

where  $\bar{f} = (1/n) \sum_{k=1}^n f_k$  is the mean inbreeding coefficient across individuals. Plugging  $\overline{\Phi}_2 = (1 + \bar{f})/(2n)$  into eq. 3.18, we get

$$\text{Var} \left[ 1 - \sum_{i=1}^I \widehat{p}_i^2 \right] \approx \frac{2}{n}(1 + \bar{f}) \left[ \sum_{i=1}^I p_i^3 - \left( \sum_{i=1}^I p_i^2 \right)^2 \right]. \quad (3.19)$$

### 3.2.3 The X chromosome case

A common situation in which data of mixed ploidy arise is on sex chromosomes, for which members of one sex have two copies of a specific sex chromosome and members of the other sex have one copy. Later, we examine data on the human X chromosome, for which females have two copies and males have one. Thus, we now utilize eq. 3.13 to derive an unbiased estimator of gene diversity in samples from the X chromosome.

Consider an X-linked locus with  $I$  distinct alleles, allele frequencies  $p_i \in [0, 1]$  and

$\sum_{i=1}^I p_i = 1$ . Suppose a sample from a population has  $n_F$  females and  $n_M$  males, each of whom is possibly inbred and related to other sampled individuals. Let  $\mathcal{M}$ ,  $\mathcal{F}$ , and  $\mathcal{U}$  be the sets of distinct types of male-male, female-female, and male-female relative pairs in the sample, respectively. Further, let  $\eta_R$  be the number of pairs of individuals with relationship type  $R$  and let  $\Phi_R$  be the kinship coefficient for each of these pairs. Let males be group 1 and let females be group 2. Plugging  $g = 2$ ,  $n_1 = n_M$ ,  $n_2 = n_F$ ,  $m_1 = 1$ , and  $m_2 = 2$  into eq. 3.13, we obtain an unbiased estimator for gene diversity at an X-linked locus as

$$\tilde{H} = \frac{(n_M + 2n_F)(n_M + 2n_F - 1)}{D} \hat{H}, \quad (3.20)$$

where

$$D = (n_M + 2n_F)(n_M + 2n_F - 1) - 2n_F \bar{f}_F - 2 \sum_{R \in \mathcal{M}} \eta_R \Phi_R - 8 \sum_{R \in \mathcal{F}} \eta_R \Phi_R - 4 \sum_{R \in \mathcal{U}} \eta_R \Phi_R,$$

$\bar{f}_F = (1/n_F) \sum_{k=1}^{n_F} f_k$  is the mean inbreeding coefficient across female individuals, and  $f_k$  is the inbreeding coefficient for female  $k$ .

The following special case of eq. 3.20 will be useful for the examples we consider in subsequent sections. It makes use of Table 3.1, which shows the various types of relationships possible for the X chromosome in pairs of individuals. Suppose a non-inbred sample from a population has  $n_F$  females and  $n_M$  males, among which  $\eta_k$  pairs of relationship type  $k$  are included. Let  $\Phi_k$  be the kinship coefficient for each of these pairs. Because the sample is not inbred, the mean inbreeding coefficient across female individuals is  $\bar{f}_F = 0$ . Plugging  $\bar{f}_F$  as well as  $\eta_k$  and  $\Phi_k$  for each relationship type  $k$  (Table 3.1) into eq. 3.20, we obtain

$$\tilde{H} = \frac{(n_M + 2n_F)(n_M + 2n_F - 1)}{(n_M + 2n_F)(n_M + 2n_F - 1) - 2 \sum_{k=1}^4 \eta_k \Phi_k - 8 \sum_{k=5}^{12} \eta_k \Phi_k - 4 \sum_{k=13}^{20} \eta_k \Phi_k} \hat{H}. \quad (3.21)$$

### 3.3 Data analysis

#### 3.3.1 Data

We investigated the properties of  $\tilde{H}$  on mixed-ploidy data using analytical computations of bias, variance, and mean squared error, simulations, and analysis of data from human populations. Our choices for simulation parameters were designed based on values in the data. In our analytical computations and simulations, we based our assumed true allele frequencies on sample allele frequencies at 36 X-chromosomal loci typed in 950 unrelated individuals, 624 males and 326 females, from the Human Genome Diversity Panel (HGDP-CEPH) microsatellite dataset of 1048 individuals (*Ramachandran et al.*, 2008). Individuals 127 and 139 from the *Ramachandran et al.* (2008) dataset were not included in our analyses. The 950 individuals were assumed to have no first- or second-degree relationships, based on the *Rosenberg* (2006) analysis of the full HGDP-CEPH panel.

Our data analysis was performed on a dataset of 13,052 X-chromosomal single nucleotide polymorphism (SNP) loci genotyped in 485 individuals from 29 populations in the HGDP-CEPH panel (*Jakobsson et al.*, 2008). We also removed individuals related through the X chromosome, yielding a dataset of 446 unrelated individuals. Unlike the *Jakobsson et al.* (2008) dataset of 443 unrelated individuals, our set of 446 individuals did not retain individuals 866, 1046, or 1049, which are not in the H952 subset of the HGDP-CEPH panel. However, individuals 292, 451, 477, 983, 988, and 1089 were included in the dataset of non-relatives because they were all involved exclusively in male-male parent-offspring relationships, and were therefore unrelated to other sampled individuals through the X chromosome.

### 3.3.2 Data analysis methods

We used simulations and analytical calculations to evaluate the behavior of the estimator  $\tilde{H}$  for X-chromosomal loci under conditions of varying heterozygosities, sample sizes, and relationships of sampled individuals. We compared the relative performance of  $\tilde{H}$  and  $\hat{H}$  by applying  $\tilde{H}$  and  $\hat{H}$  to samples containing related individuals and  $\hat{H}$  to samples in which relatives were removed so that no relative pairs remained. True allele frequencies were based on microsatellite sample allele frequencies (see “Data”). In the simulations, individuals of a relative pair were generated by randomly choosing the allele(s) of the first individual based on the empirical allele frequency distribution from the dataset. For a given type of relative pair, we then simulated the allele(s) of the second individual by copying alleles from the first individual using the probabilities of sharing zero, one, and two alleles IBD for that type of pair. Table 3.2 depicts these probabilities, as well as the symbols used here to denote the various classes of relative pairs. If only one allele was shared, then it was copied in the second individual from the first allele of the first (independently generated) individual. In cases of male-female relative pairs, the male was generated first and the second allele of the female was always chosen independently from the allele frequency distribution.

To create a reduced dataset of unrelated individuals, the second (possibly dependent) individual was not included for same-sex pairs, whereas for male-female pairs, the male relative was removed. Thus, because each individual in our simulation was included in exactly one relative pair, the number of individuals used to calculate  $\hat{H}$  for the unrelated sample was always half of that used for the other two estimators. Removing the male in male-female pairs results in the loss of 1/3 of the alleles, compared to a loss of 1/2 of the alleles for removal of an individual from a same-sex pair. Thus, compared to removing females, removing males from male-female pairs generates a larger sample of alleles while still ensuring that no individuals are related.



The value assumed for the true heterozygosity,  $H$ , of a specific locus, was calculated from the assumed true allele frequencies based on the genotypic data of the 950 unrelated individuals. In each simulated scenario, for each of the three estimators, this true heterozygosity was compared to the mean of the estimates produced by the estimator in 100,000 replicate simulations. The subscript *full* is used to denote cases in which an estimator was applied to the entire sample, whereas the subscript *reduced* indicates that relatives were removed from the sample. The bias of each estimator for a scenario was found by subtracting  $H$  from the mean value of the estimates for that estimator. Variance was calculated as the squared mean of the estimates across simulations subtracted from the mean across simulations of the squares of the estimates. Mean squared error (MSE) was then calculated as the sum of bias squared and variance.

### 3.3.2.1 Approximate variance

Because each of our analyses was performed on samples that contained only pairs of related individuals, the assumptions that underlie the derivation of the approximate variance (eq. 3.18) apply. We compared the exact, the approximate, and the simulated variance for  $\tilde{H}$  and  $\hat{H}$  in a series of cases that included only full-sib pairs. We chose nine representative cases of the various parameters that can affect estimator performance. Three of these cases considered an equal mix of male-male, female-female, and male-female full-sib pairs at the ATCT003 ( $H = 0.7794$ ), DXS1068 ( $H = 0.7344$ ), and GATA48H04 ( $H = 0.6476$ ) loci, chosen to represent high, intermediate, and low heterozygosity, respectively. Additionally, we considered cases at the intermediate heterozygosity locus involving 20 male-male, 80 male-male, 20 female-female, 80 female-female, 20 male-female, and 80 male-female pairs, to examine the effects of sample size and the sexes of the individuals. In each of our evaluations, we calculated the exact variances (eq. 3.15 and eq. 3.16),

approximate variances (eq. 3.18 plugged into eq. 3.15 and eq. 3.16), and simulation variances obtained from 100,000 replicate simulations.

As Table 3.3 shows, in all cases examined, the exact, approximate, and simulated variances are similar, with the approximate variance slightly underestimating the exact variance. Because of the complexity of the formula for the exact variance, the difference between approximate and exact variance does not have a simple dependence on heterozygosity or sample size. However, it can be observed in Table 3.3 that for both  $\tilde{H}$  and  $\hat{H}$ , the relative difference between the approximate and exact variances is smallest at low heterozygosity and large sample size, typically near  $\sim 2\%$ . In cases of high heterozygosity and small sample size, the relative difference remains at most  $\sim 10\%$ . We note that the same approximation to the variance of  $1 - \sum_{i=1}^I \hat{p}_i^2$  in eq. 3.18 is applied in obtaining the approximate variances of both  $\tilde{H}$  and  $\hat{H}$ . Thus, because the approximation is generally reasonably accurate and because it treats  $\tilde{H}$  and  $\hat{H}$  in the same way, our use of the approximation is sensible in our subsequent comparisons of the mean squared errors of  $\tilde{H}$  and  $\hat{H}$ .

### 3.3.3 Effect of parameters on the estimators

Several factors can potentially affect the performance of the estimators. These factors include the true value of heterozygosity itself, the sample size, the type of relative pair represented in the sample, and, if multiple types of relative pairs are included, the combination of particular types of relative pairs. We now examine each of these factors in sequence.

#### 3.3.3.1 Varying heterozygosity

To investigate the influence of varying heterozygosity on the estimator, we evaluated the scenario of 60 related individuals in 10  $t_1$  pairs, 10  $u_2$  pairs, and 10  $v_2$  pairs (see Table 3.2) for each of the 36 X-linked microsatellite loci. This

scheme incorporates 30 full-sib pairs, considering equally many males and females and utilizing three distinct kinship coefficients: 1/2 for male-male pairs ( $t_1$ ), 1/4 for male-female pairs ( $u_2$ ), and 3/8 for female-female pairs ( $v_2$ ). The 36 loci represent a spread of assumed true heterozygosities ranging from 0.4008 to 0.8599. For each locus, we calculated  $\tilde{H}_{full}$  (eq. 3.21), as well as  $\hat{H}_{full}$  and  $\hat{H}_{reduced}$  (Nei and Roychoudhury, 1974).

Figure 3.1 displays the properties of the three estimators,  $\tilde{H}_{full}$ ,  $\hat{H}_{full}$  and  $\hat{H}_{reduced}$ , based on application of analytical computations of bias (eq. 3.14 for  $\hat{H}_{full}$ ) and the variance approximation (eq. 3.18 plugged into equations 3.15 and 3.16) to each of the 36 loci.  $\tilde{H}_{full}$  and  $\hat{H}_{reduced}$  are unbiased estimators and therefore have zero bias, whereas  $\hat{H}_{full}$  exhibits increasing bias squared as heterozygosity increases. The bias squared for  $\hat{H}_{full}$  as a function of heterozygosity is plotted using the theoretical prediction based on eq. 3.14:  $[bias(\hat{H})]^2 = \left( -\frac{2(10 \times \frac{1}{2}) + 8(10 \times \frac{3}{8}) + 4(10 \times \frac{1}{4})}{(30+2 \times 30) \times (30+2 \times 30-1)} H \right)^2 = (3.897 \times 10^{-5})H^2$ . Generally, over the space of heterozygosities defined by the 36 microsatellite loci, the MSE and variance of all three estimators decrease with increasing heterozygosity.

### 3.3.3.2 Varying sample size and type of relative pair

We next applied the estimators to scenarios of varying sample size. The ATCT003 ( $H = 0.7794$ ), DXS1068 ( $H = 0.7344$ ), and GATA48H04 ( $H = 0.6476$ ) loci were chosen from the dataset to represent high, intermediate, and low heterozygosities, respectively. Only the data for the intermediate heterozygosity locus DXS1068 are shown; the other two loci yield similar results. For each locus and for each of the ten types of relative pairs in Table 3.2, we varied the sample size from 2 to 100 pairs. We considered a sample size of at least 2 pairs, as no information is available for the computation of  $\hat{H}_{reduced}$  from a single pair of male-male relatives. For all three loci, analytical calculations were performed using the variance approximation (eq. 3.18

plugged into equations 3.15 and 3.16).

Figure 3.2 shows that as sample size increases, MSE decreases for all three estimators, and it is always comparable for  $\tilde{H}_{full}$  and  $\hat{H}_{full}$  ( $\tilde{H}_{full}$  mostly overlaps  $\hat{H}_{full}$  in the figure). Usually, we expect MSE in a reduced sample to be highest due to greater variance. However, although the results conformed to this prediction for most types of relative pairs, for male-female relative pairs for which there was probability greater than or equal to 3/4 for sharing exactly one allele IBD (types  $u_1$  and  $u_4$ ), the MSE of  $\hat{H}_{reduced}$  was actually lower than the MSE for  $\tilde{H}_{full}$  and  $\hat{H}_{full}$ . The same result was also detected in our simulations (data not shown). Investigating further, we found that in male-male and female-female pairs, cases with high probabilities for sharing one or two alleles IBD had MSEs for  $\tilde{H}_{full}$  and  $\hat{H}_{full}$  that were closer to the  $\hat{H}_{reduced}$  MSE values, compared with the higher MSE for  $\hat{H}_{reduced}$  observed in other cases. The MSE of  $\hat{H}_{reduced}$  is smaller relative to that of the other estimators for  $u_1$  and  $u_4$  male-female pairs because when only 1/3 of the sample is removed in creating the unrelated set of individuals (removal of males), the increase in variance due to the relatively small decrease in sample size in  $\hat{H}_{reduced}$  is comparable to the increased variance caused by the high IBD probabilities for  $u_1$  and  $u_4$  pairs in  $\tilde{H}_{full}$  and  $\hat{H}_{full}$ , unlike in other cases. When females, instead of males, are removed from male-female pairs, decreasing the sample by 2/3 rather than 1/3, the estimators behave more intuitively (Figure 3.3), with  $\hat{H}_{reduced}$  yielding the highest MSE.

### 3.3.3.3 Varying combinations of relative pairs

Finally, we studied the effect of relative pair combinations in a sample, using allele frequencies at the ATCT003, DXS1068, and GATA48H04 loci. Only the results for the highest heterozygosity locus, ATCT003, are shown; as was true in the previous section, each locus yielded similar results. For each locus, we examined each of the 231 possible divisions of exactly 20 full-sib pairs into male-male ( $t_1$ ), male-female ( $u_2$ ),

and female-female ( $v_2$ ) pairs. Figure 3.4 displays the MSE, variance, and bias squared of the three estimators, calculated analytically using the variance approximation (eq. 3.18), for various combinations of  $t_1$ ,  $u_2$ , and  $v_2$  pairs for the ATCT003 locus. Variance was highest for  $\hat{H}_{reduced}$ , because it had the smallest sample of alleles. For all estimators, variance was highest where the configuration of full-sibs had mostly male-male pairs, again due to the smaller sample of alleles.  $\tilde{H}_{full}$  and  $\hat{H}_{reduced}$  were unbiased across the space of possible combinations.  $\hat{H}_{full}$  showed a trend in bias squared in which configurations with a greater proportion of males had higher bias squared, as is predicted analytically from the smaller sample size (eq. 3.14). For all configurations, the bias squared of  $\hat{H}_{full}$  was greater than that for the other estimators. Among the three estimators, MSE was highest for  $\hat{H}_{reduced}$ . Similarly to the observation for variance, MSE was greatest for configurations with a high proportion of male-male pairs. Although  $\tilde{H}_{full}$  performed slightly poorer in having a greater variance when compared to  $\hat{H}_{full}$ , it had a slightly lower MSE due to its lower bias. More generally, although  $\tilde{H}_{full}$  performed better in the setting of Figure 3.4, the exact formula can be used to determine which estimator has lowest MSE for a given scenario.

### 3.3.4 Application to data

We next investigated the behavior of our estimator using X-chromosomal SNP datasets of 485 individuals and 446 unrelated individuals (see “Data”). Table 3.4 displays the relative pairs in the sample of 485 individuals. Because we analyzed the estimators separately by population, the subscripts of 485 and 446 refer to whether or not relatives were included in a calculation, not to the actual numbers of individuals in that calculation. In the same manner as in *DeGiorgio and Rosenberg (2009)*, we took  $\hat{H}_{446}$  for each population to be a proxy for true heterozygosity, because this quantity provided an unbiased estimate when no relatives were included in the sample. Note

that removed individuals belonged only to pairs related through the X chromosome; individuals related only autosomally (such as male-male parent-offspring pairs) were included in the reduced sample. In our analysis, we compared the means of  $\widehat{H}_{485}$  and  $\widehat{H}_{446}$  across the 13,052 loci to the corresponding mean of  $\widehat{H}_{446}$ .

Figure 3.5 compares the difference between the mean of  $\widehat{H}_{485}$  across loci ( $\overline{\widehat{H}}_{485}$ ) and the mean of  $\widehat{H}_{446}$  ( $\overline{\widehat{H}}_{446}$ ) with the difference between the mean of  $\widetilde{H}_{485}$  ( $\overline{\widetilde{H}}_{485}$ ) and the mean of  $\widehat{H}_{446}$  ( $\overline{\widehat{H}}_{446}$ ). As Figure 3.5A shows,  $\overline{\widetilde{H}}_{485}$  generally yields a lower heterozygosity estimate than  $\overline{\widehat{H}}_{446}$  due to the downward bias caused by related individuals. Applying  $\overline{\widetilde{H}}_{485}$  reduces the magnitude of the difference between the estimate of heterozygosity in sets with and without relatives (Figure 3.5B), and  $\overline{\widetilde{H}}_{485}$  yields values that are not consistently lower than those of  $\overline{\widehat{H}}_{446}$ . It is important to note that because 15 of 45 of the relative pairs in the data have an uncertain second-degree relationship ( $t_3$ ,  $u_5$ , or  $v_5$ ),  $\overline{\widetilde{H}}_{485}$  might have overcorrected bias in cases in which the individuals were not related via the X chromosome and undercorrected bias in cases in which the individuals actually were related on the X-chromosome.

A Wilcoxon signed rank test was used to evaluate the differences between  $\overline{\widetilde{H}}_{485}$  and  $\overline{\widehat{H}}_{446}$  applied to the 13 populations that contained relatives (see Table 3.4). This test yielded a  $p$ -value of 0.0024, indicating that the inclusion of relatives had a significant impact on the estimation of heterozygosity using  $\widehat{H}$ . In contrast, the Wilcoxon signed rank comparison of  $\overline{\widetilde{H}}_{485}$  and  $\overline{\widehat{H}}_{446}$  yielded a  $p$ -value of 0.6355, indicating that the inclusion of relatives did not significantly alter the estimation of heterozygosity when  $\widetilde{H}$  was used. The mean difference  $\overline{\widetilde{H}}_{485} - \overline{\widehat{H}}_{446}$  ( $-8.0493 \times 10^{-5}$ ) and the mean absolute difference  $|\overline{\widetilde{H}}_{485} - \overline{\widehat{H}}_{446}|$  ( $6.3159 \times 10^{-4}$ ) were smaller across the 13 populations than the mean difference ( $-1.9393 \times 10^{-3}$ ) and mean absolute difference ( $1.9849 \times 10^{-3}$ ),  $\overline{\widehat{H}}_{446} - \overline{\widehat{H}}_{485}$  and  $|\overline{\widehat{H}}_{446} - \overline{\widehat{H}}_{485}|$ , respectively.

We also investigated the behavior of  $\widetilde{H}$  and  $\widehat{H}$  with regard to variance for the 13 populations that contained relatives. We compared  $\overline{\widetilde{H}}_{485} - \overline{\widehat{H}}_{446}$  and  $\overline{\widetilde{H}}_{485} - \overline{\widehat{H}}_{446}$ ,

which we used as proxies for bias, following the methods of *DeGiorgio and Rosenberg* (2009), and the standard deviations of the two estimators applied with relatives included. From Figure 3.6, we observe that while there was a sizeable difference in the bias proxy between  $\widehat{H}_{485}$  and  $\widetilde{H}_{485}$ , there was only a small difference in standard deviation. This result is compatible with the results from our analytical computations, which suggest that  $\widetilde{H}$  corrects bias without substantially increasing variance.

### 3.4 Discussion

Our estimator,  $\widetilde{H}$ , is an effective tool for assessing the gene diversity of a sample of arbitrary ploidy containing related or inbred individuals. It can be used to provide unbiased estimates of expected heterozygosity when the inbreeding and kinship coefficients of sampled individuals are known. We have found that the unbiasedness of the diploid estimator of *DeGiorgio and Rosenberg* (2009) extends to a much more general set of scenarios, provided that kinship coefficients are appropriately weighted by ploidy in the computation.

Here, we have evaluated the properties of  $\widetilde{H}$  in the specific case of the human X chromosome. Through our analytical calculations, we have shown that, similarly to the *DeGiorgio and Rosenberg* (2009) estimator in the diploid case, the performance of  $\widetilde{H}$  is generally superior to that of  $\widehat{H}$  when the sample to which the estimators are applied contains relatives.  $\widetilde{H}$  accounts for the bias introduced by relatedness while simultaneously maintaining comparable MSE and variance to  $\widehat{H}$ . Our estimator also performs well compared to  $\widehat{H}$  when applied to data from human populations. While the true heterozygosity of each population is not known, when we compared  $\widetilde{H}$  and  $\widehat{H}$  to an approximation of true heterozygosity,  $\widehat{H}$  applied to the dataset with no related individuals, we found that the difference between the estimate when relatives were included and when relatives were not included was significantly smaller for  $\widetilde{H}$ . Because the reduction in this proxy for bias is accompanied by only a small increase

in standard deviation, we argue that  $\tilde{H}$  should often be preferred over  $\hat{H}$  in the estimation of gene diversity in a sample containing relatives.

In addition to developing the  $\tilde{H}$  estimator for gene diversity, we also determined the analytical variance of our estimator, allowing us to theoretically evaluate the properties of  $\tilde{H}$ . We also developed an approximation for variance (eq. 3.18) that is simpler to compute and that is applicable when each individual has at most one relative in the sample. Knowledge of the theoretical variance can further allow investigators to evaluate the circumstances under which  $\tilde{H}$  applied to a full sample, including relatives, is superior to using  $\hat{H}$  with a reduced sample in which members of relative pairs have been removed. For example, Figure 3.2 indicates that removing relatives will provide a lower MSE of the heterozygosity estimate in some cases. However, Figure 3.4 suggests that  $\tilde{H}_{full}$  yields a lower MSE than  $\hat{H}_{reduced}$  except in the small fraction of relative-pair combinations that contain large numbers of  $u_1$  pairs. Thus, we propose that in most cases, the use of  $\tilde{H}$  on a sample set that includes related individuals affords a better estimate of gene diversity than applying  $\hat{H}$  on a sample that contains no relatives, and that investigators can use the theoretical variance of  $\tilde{H}$  to determine whether a given situation is likely to be among the exceptions.

### 3.5 Acknowledgments

We thank Laurent Excoffier and three anonymous reviewers for their valuable comments. This work was supported by National Institute of Health (NIH) grant R01 GM081441, NIH training grant T32 GM070449, a University of Michigan Rackham Merit Fellowship, and grants from the Burroughs Wellcome Fund and the Alfred P. Sloan Foundation.



Table 3.1: Relationship types with corresponding X-linked kinship coefficients

Relationship number ( $k$ )	Relationship type	Symbol for relationship class	Sexes of the pair	$\Phi$
1	Full-sibs	$t_1$	male-male	1/2
2	Half-sibs (female parent)	$t_1$	male-male	1/2
3	Uncle-Nephew (female parent)	$t_2$	male-male	1/4
4	Grandfather-Grandson (female parent)	$t_1$	male-male	1/2
5	Parent-Offspring	$v_1$	female-female	1/4
6	Full-sibs	$v_2$	female-female	3/8
7	Half-sibs (male parent)	$v_1$	female-female	1/4
8	Half-sibs (female parent)	$v_3$	female-female	1/8
9	Aunt-Niece (male parent)	$v_3$	female-female	1/8
10	Aunt-Niece (female parent)	$v_4$	female-female	3/16
11	Grandmother-Granddaughter (male parent)	$v_1$	female-female	1/4
12	Grandmother-Granddaughter (female parent)	$v_3$	female-female	1/8
13	Parent-Offspring	$u_1$	male-female	1/2
14	Full-sibs	$u_2$	male-female	1/4
15	Half-sibs (female parent)	$u_2$	male-female	1/4
16	Uncle-Niece (male parent)	$u_2$	male-female	1/4
17	Uncle-Niece (female parent)	$u_3$	male-female	1/8
18	Aunt-Nephew (female parent)	$u_4$	male-female	3/8
19	Grandfather-Granddaughter (female parent)	$u_2$	male-female	1/4
20	Grandmother-Grandson (female parent)	$u_2$	male-female	1/4

Relationship types with X-linked kinship coefficient of zero are not shown. These include the male-male relationships of parent-offspring, half-sibs (through male), uncle-nephew (through male), and grandfather-grandson (through male) as well as the male-female relationships of half-sibs (through male), aunt-nephew (through male), grandfather-granddaughter (through male), and grandmother-grandson (through male).

Table 3.2: Symbols used for relative pair types.

Symbol	$\Upsilon_0$	$\Upsilon_1$	$\Upsilon_2$	$\Phi$	Sex	Relative types
$t_1$	1/2	1/2	0	1/2	male-male	full-sib, half-sib (female parent), grandfather-grandson (female parent)
$t_2$	3/4	1/4	0	1/4	male-male	uncle-nephew (female parent)
$v_1$	0	1	0	1/4	female-female	parent-offspring, half-sib (male parent), grandmother-granddaughter (male parent)
$v_2$	0	1/2	1/2	3/8	female-female	full-sib
$v_3$	1/2	1/2	0	1/8	female-female	half-sib (female parent), aunt-niece (male parent), grandmother-granddaughter (female parent)
$v_4$	1/4	3/4	0	3/16	female-female	aunt-niece (female parent)
$u_1$	0	1	0	1/2	male-female	parent-offspring
$u_2$	1/2	1/2	0	1/4	male-female	full-sib, half-sib (female parent), uncle-niece (male parent), grandfather-granddaughter (female parent), grandmother-grandson (female parent)
$u_3$	3/4	1/4	0	1/8	male-female	uncle-niece (female parent)
$u_4$	1/4	3/4	0	3/8	male-female	aunt-nephew (female parent)
$t_3$	—	—	—	5/24	male-male	uncertain second degree relative
$v_5$	—	—	—	17/96	female-female	uncertain second degree relative
$u_5$	—	—	—	3/20	male-female	uncertain second degree relative

$\Upsilon_0$ ,  $\Upsilon_1$  and  $\Upsilon_2$  designate the probabilities that individuals share 0, 1, and 2 alleles IBD at an X-linked locus, respectively. All types of relative pairs denoted by the same symbol have the same kinship coefficient, sexes, and probabilities of sharing 0, 1, and 2 alleles IBD.  $\Phi$  can be calculated from  $\Upsilon_1$  and  $\Upsilon_2$  using  $\Phi_{ij} = \Upsilon_1$  if  $i$  and  $j$  are both male,  $\Phi_{ij} = \frac{1}{4}\Upsilon_1 + \frac{1}{2}\Upsilon_2$  if  $i$  and  $j$  are both female, and  $\Phi_{ij} = \frac{1}{2}\Upsilon_1 + \Upsilon_2$  if  $i$  is male and  $j$  is female. For each possible pair of sexes (male-male, female-female, male-female), the kinship coefficient for second-degree relatives of an uncertain type was found by averaging the kinship coefficients for all second-degree relationships in Table 3.1 with that pair of sexes, assuming that all were equally likely. Second-degree relationships include half-sib, grandparent-grandchild and avuncular pairs. For male-male pairs,  $t_3 = (2 \times \frac{1}{2} + 1 \times \frac{1}{4} + 3 \times 0) / 6 = 5/24$ . For female-female pairs,  $v_5 = (2 \times \frac{1}{4} + 3 \times \frac{1}{8} + 1 \times \frac{3}{16}) / 6 = 17/96$ . For male-female pairs,  $u_5 = (4 \times \frac{1}{4} + 1 \times \frac{1}{8} + 1 \times \frac{3}{8} + 4 \times 0) / 10 = 3/20$ . The divisor in each of the previous equations describes the total number of possible second-degree relatives for that sex pair (e.g. grandmother-grandson, aunt-nephew, etc. for the male-female case). This number includes second-degree relatives that are not related on the X chromosome, because the assignment of relationships in the dataset was based on autosomal data. The kinship coefficients for  $t_3$ ,  $v_5$ , and  $u_5$  were used only for analysis of population data, and they were not used in our investigations of the effects on the estimators of varying the parameters.

Table 3.3: Comparison of exact, approximate, and simulation variances

Estimator	Locus	Relative pairs	Exact variance	Approximate variance	Simulation variance	Relative difference of approximation (percent)	
$\tilde{H}$	ATCT003	10 $t_1$ , 10 $u_2$ , 10 $v_2$	$7.55 \times 10^{-4}$	$6.82 \times 10^{-4}$	$7.48 \times 10^{-4}$	9.59	
	DXS1068	10 $t_1$ , 10 $u_2$ , 10 $v_2$	$1.54 \times 10^{-3}$	$1.47 \times 10^{-3}$	$1.50 \times 10^{-3}$	4.82	
	GATA48H04	10 $t_1$ , 10 $u_2$ , 10 $v_2$	$1.50 \times 10^{-3}$	$1.48 \times 10^{-3}$	$1.54 \times 10^{-3}$	1.52	
	DXS1068	20 $t_1$	$3.62 \times 10^{-3}$	$3.31 \times 10^{-3}$	$3.49 \times 10^{-3}$	8.55	
	DXS1068	80 $t_1$	$8.08 \times 10^{-4}$	$7.82 \times 10^{-4}$	$7.93 \times 10^{-4}$	3.16	
	DXS1068	20 $u_2$	$1.97 \times 10^{-3}$	$1.90 \times 10^{-3}$	$1.97 \times 10^{-3}$	3.29	
	DXS1068	80 $u_2$	$4.68 \times 10^{-4}$	$4.60 \times 10^{-4}$	$4.61 \times 10^{-4}$	1.71	
	DXS1068	20 $v_2$	$1.99 \times 10^{-3}$	$1.87 \times 10^{-3}$	$1.95 \times 10^{-3}$	5.87	
	DXS1068	80 $v_2$	$4.64 \times 10^{-4}$	$4.53 \times 10^{-4}$	$4.56 \times 10^{-4}$	2.37	
	$\hat{H}$	ATCT003	10 $t_1$ , 10 $u_2$ , 10 $v_2$	$7.45 \times 10^{-4}$	$6.74 \times 10^{-4}$	$7.38 \times 10^{-4}$	9.59
		DXS1068	10 $t_1$ , 10 $u_2$ , 10 $v_2$	$1.52 \times 10^{-3}$	$1.45 \times 10^{-3}$	$1.49 \times 10^{-3}$	4.82
		GATA48H04	10 $t_1$ , 10 $u_2$ , 10 $v_2$	$1.49 \times 10^{-3}$	$1.46 \times 10^{-3}$	$1.53 \times 10^{-3}$	1.52
DXS1068		20 $t_1$	$3.53 \times 10^{-3}$	$3.23 \times 10^{-3}$	$3.40 \times 10^{-3}$	8.55	
DXS1068		80 $t_1$	$8.03 \times 10^{-4}$	$7.77 \times 10^{-4}$	$7.88 \times 10^{-4}$	3.16	
DXS1068		20 $u_2$	$1.95 \times 10^{-3}$	$1.88 \times 10^{-3}$	$1.94 \times 10^{-3}$	3.29	
DXS1068		80 $u_2$	$4.67 \times 10^{-4}$	$4.59 \times 10^{-4}$	$4.60 \times 10^{-4}$	1.71	
DXS1068		20 $v_2$	$1.95 \times 10^{-3}$	$1.84 \times 10^{-3}$	$1.91 \times 10^{-3}$	5.87	
DXS1068		80 $v_2$	$4.61 \times 10^{-4}$	$4.51 \times 10^{-4}$	$4.54 \times 10^{-4}$	2.37	

The exact (eq. 3.15 and eq. 3.16), approximate (eq. 3.18), and simulation variances calculated for the combination of 10 male-male ( $t_1$ ), 10 male-female ( $u_2$ ), and 10 female-female ( $v_2$ ) full-sib pairs at the ATCT003 ( $H = 0.7794$ ), DXS1068 ( $H = 0.7344$ ), and GATA48H04 ( $H = 0.6476$ ) loci as well as for sets of 20 and 80 pairs of each full-sib pair type at DXS1068. Simulation variances were calculated over 100,000 replicates. The relative difference of the approximation was computed as  $100 \times |\text{approximate variance} - \text{exact variance}| / (\text{exact variance})$ .

Table 3.4: Types of relative pairs in populations from the dataset of 485 individuals reported by *Jakobsson et al. (2008)*

	$t_1$	$t_2$	$t_3$	$u_1$	$u_2$	$u_3$	$u_4$	$u_5$	$v_1$	$v_2$	$v_3$	$v_4$	$v_5$
Bantu (Kenya)	<b>1</b>												
Bedouin				<b>1</b>				<b>1</b>					
Biaka Pygmy	<b>1</b>		<b>2</b>		<b>1</b>			<b>2</b>					
Druze					<b>2</b>				<b>1</b>				<b>2</b>
Kalash			<b>1</b>										
Mandenka			<b>1</b>										<b>1</b>
Maya					<b>1</b>				<b>1</b>				<b>2</b>
Mbuti Pygmy			<b>1</b>	<b>1</b>									
Melanesian	<b>1</b>			<b>3</b>					<b>2</b>	<b>2</b>			<b>1</b>
Mozabite										<b>1</b>			
Palestinian										<b>1</b>			<b>1</b>
Pima	<b>1</b>	<b>1</b>		<b>1</b>	<b>1</b>	<b>1</b>						<b>1</b>	
Yoruba				<b>1</b>	<b>1</b>				<b>1</b>	<b>1</b>			
Total	4	1	5	7	6	1	0	3	5	5	0	1	7

Symbols for the types of relative pairs appear in Table 3.2.

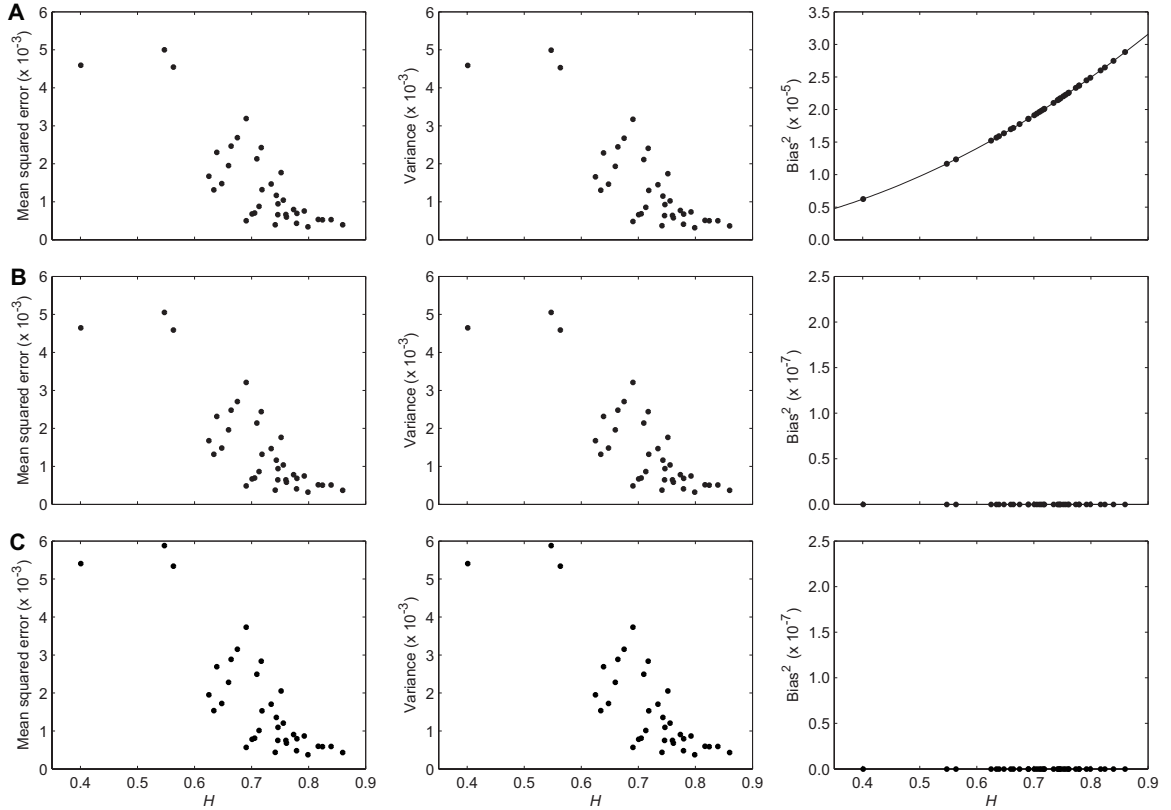


Figure 3.1: Mean squared error, variance, and bias squared for each estimator, obtained analytically using the variance approximation (eq. 3.18), as a function of heterozygosity for 36 loci. The scheme considered included 60 individuals in 10  $t_1$  pairs ( $\Phi = 1/2$ ), 10  $u_2$  pairs ( $\Phi = 1/4$ ), and 10  $v_2$  pairs ( $\Phi = 3/8$ ). **A.**  $\hat{H}_{full}$ . The curve through the points in the third column is described by eq. 3.14. **B.**  $\tilde{H}_{full}$ . **C.**  $\hat{H}_{reduced}$ .

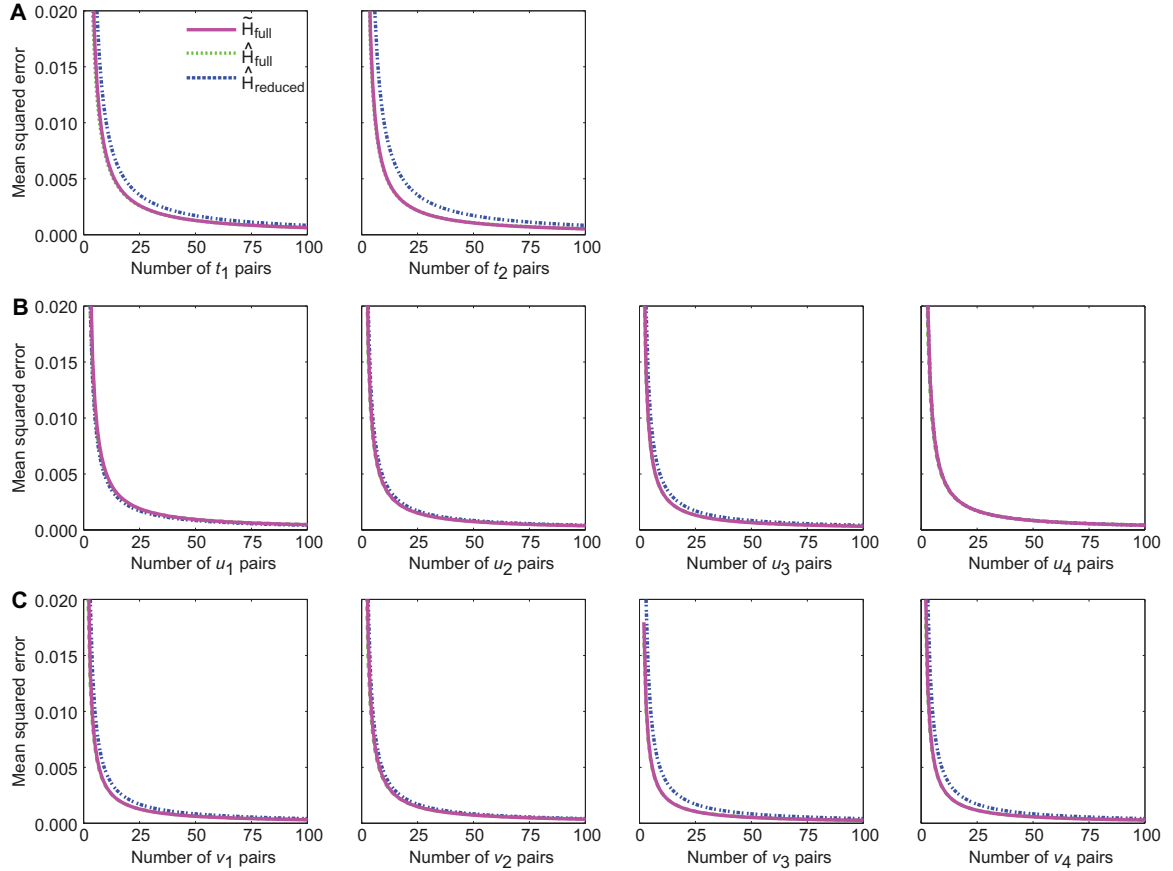


Figure 3.2: Mean squared error as a function of sample size (number of pairs = number of individuals / 2) calculated analytically using the variance approximation (eq. 3.18) based on allele frequencies at the DXS1068 locus ( $H = 0.7344$ ). Each plot considers different sample sizes for one type of relative pair (Table 3.2). The range of each plot is truncated at 0.020 and the graph of  $\tilde{H}_{full}$  covers that of  $\hat{H}_{full}$ . **A.** Male-male relative pairs. **B.** Male-female relative pairs. **C.** Female-female relative pairs. Note that the  $\hat{H}_{reduced}$  line in the graph of mean squared error as a function of the number of  $u_4$  pairs is behind the other two lines.

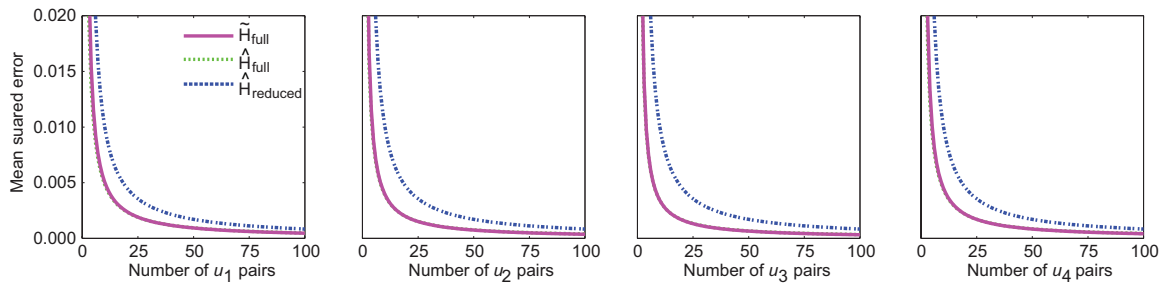


Figure 3.3: Mean squared error as a function of sample size (number of pairs = number of individuals / 2) calculated analytically using the variance approximation (eq. 3.18), based on allele frequencies at the DXS1068 locus ( $H = 0.7344$ ) for male-female relative pairs in which the females were removed to calculate  $\hat{H}_{reduced}$ . The range of each plot is truncated at 0.020. The graph of  $\tilde{H}_{full}$  covers that of  $\hat{H}_{full}$ .

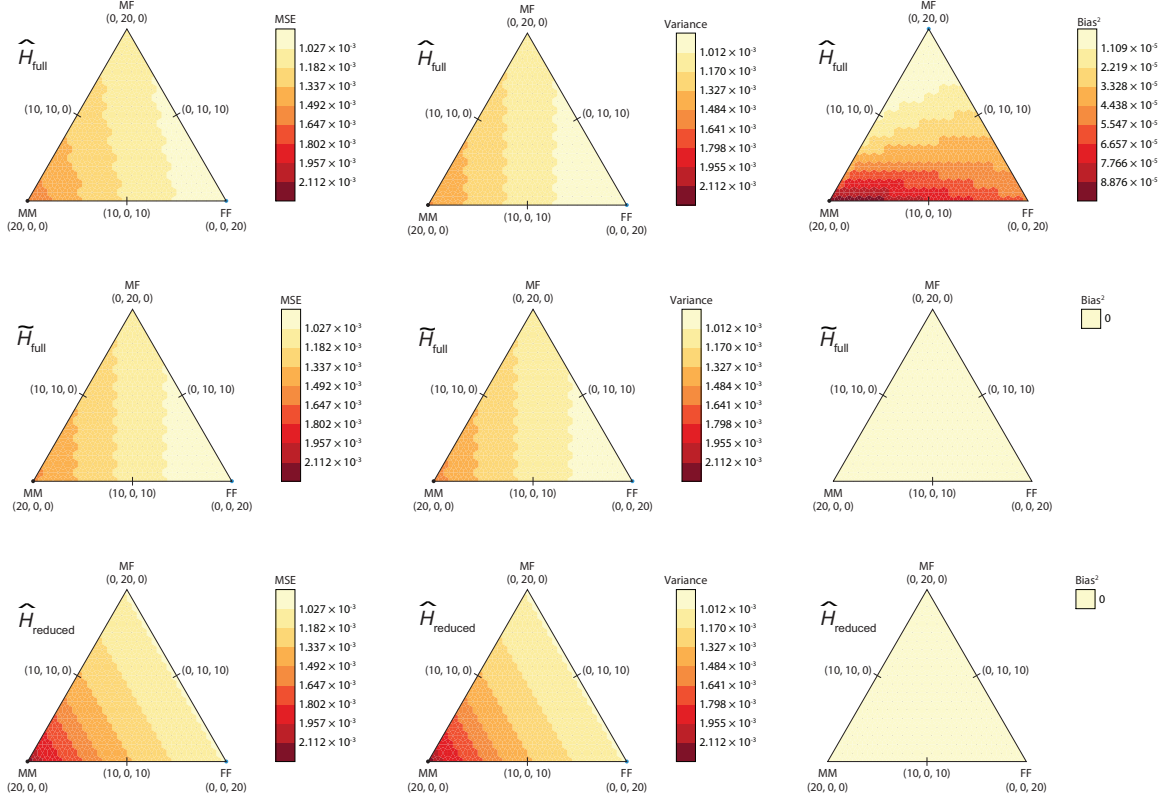


Figure 3.4: Mean squared error (MSE), variance, and bias squared of  $\widehat{H}_{full}$ ,  $\widetilde{H}_{full}$ , and  $\widehat{H}_{reduced}$ , calculated analytically using the variance approximation (eq. 3.18), as functions of the configuration of  $t_1$  male-male ( $\Phi = 1/2$ ),  $u_1$  male-female ( $\Phi = 1/2$ ), and  $v_2$  female-female ( $\Phi = 3/8$ ) pairs in 20 total relative pairs, based on allele frequencies at the ATCT003 locus ( $H = 0.7794$ ). Each row displays a different estimator and each column displays a different statistic. The three vertices of each triangle represent 20 male-male, 20 male-female, and 20 female-female full-sib pairs. The numbers on the scale indicate the cutoff values for colors. Note that unlike for the other two estimators, the scale for bias squared of  $\widehat{H}_{full}$  includes nonzero values. The black dot on each graph (except the bias squared graphs for  $\widetilde{H}_{full}$  and  $\widehat{H}_{reduced}$ ) represents the largest value in that triangle, and the blue dot represents the smallest value.



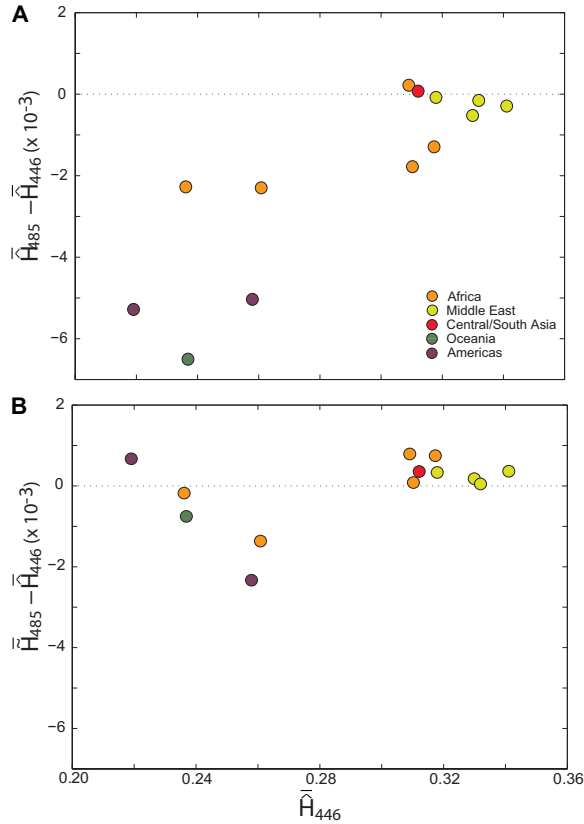


Figure 3.5: Comparison of the difference between the mean of  $\widehat{H}_{485}$  across loci and the mean of  $\widehat{H}_{446}$  with the difference between the mean of  $\widetilde{H}_{485}$  and the mean of  $\widetilde{H}_{446}$ . **A.** The difference between the mean of  $\widehat{H}_{485}$  and the mean of  $\widehat{H}_{446}$  for each of the 13 populations containing relatives (Table 3.4). **B.** The difference between the mean of  $\widetilde{H}_{485}$  and the mean of  $\widetilde{H}_{446}$  for each of the 13 populations. The estimators were applied to a dataset of 13,052 SNP loci with 485 individuals belonging to 29 populations, and the results for the 13 populations with relatives are shown. Included in the set of 485 individuals was a subset of 446 individuals that contained no relatives. The subscripts of 485 and 446 refer to whether or not relatives were included, not to the actual number of individuals in the calculation. Each data point represents one population, with color indicating the geographic region of that population. The dotted line indicates a difference of zero.

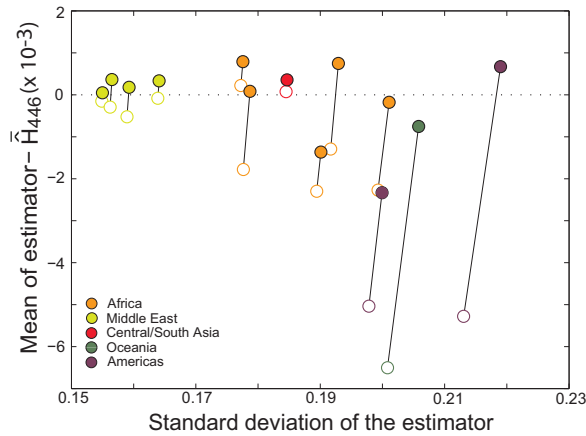


Figure 3.6: Comparison of the difference between the mean of the estimator and the mean of  $\hat{H}_{446}$  and standard deviation of the estimator, for the estimators  $\hat{H}_{485}$  and  $\hat{H}_{485}$ . These estimators were applied to a full dataset of 13,052 X-chromosome SNP loci with 485 individuals belonging to 29 populations, whereas 446 individuals were included in the reduced dataset that contained no relatives. Only the 13 populations containing relatives are shown. The subscripts 485 and 446 refer to whether or not relatives were included, not to the actual number of individuals in the calculation. Open and closed points represent the estimates for  $\hat{H}_{485}$  and  $\hat{H}_{446}$ , respectively. The dotted line indicates a difference of zero. Lines connect data points representing the same population, with each population colored by geographic region.

### 3.6 Appendix A

In this section, we present proofs for eqs. 3.9, 3.11, 3.12, and 3.13.

*Proof of eq. 3.9.* Applying the definition of  $\widehat{p}_i$  and using eq. 3.8, we have

$$\begin{aligned}
\mathbb{E}[\widehat{p}_i^2] &= \frac{1}{\left(\sum_{b=1}^g n_b m_b\right)^2} \sum_{a=1}^g \sum_{b=1}^g \sum_{j=1}^{n_a} \sum_{k=1}^{n_b} \mathbb{E}\left[X_{(a,j)}^{(i)} X_{(b,k)}^{(i)}\right] \\
&= \frac{1}{\left(\sum_{b=1}^g n_b m_b\right)^2} \sum_{a=1}^g \sum_{b=1}^g \sum_{j=1}^{n_a} \sum_{k=1}^{n_b} \sum_{\ell=1}^{m_a} \sum_{t=1}^{m_b} \mathbb{E}\left[A_{(a,j),\ell}^{(i)} A_{(b,k),t}^{(i)}\right] \\
&= \frac{1}{\left(\sum_{b=1}^g n_b m_b\right)^2} \sum_{a=1}^g \sum_{b=1}^g \sum_{j=1}^{n_a} \sum_{k=1}^{n_b} \sum_{\ell=1}^{m_a} \sum_{t=1}^{m_b} \left(\Phi_{(a,j)(b,k)} p_i (1 - p_i) + p_i^2\right) \\
&= \frac{1}{\left(\sum_{b=1}^g n_b m_b\right)^2} \sum_{a=1}^g \sum_{b=1}^g \sum_{j=1}^{n_a} \sum_{k=1}^{n_b} m_a m_b \left(\Phi_{(a,j)(b,k)} p_i (1 - p_i) + p_i^2\right) \\
&= \frac{\left(\sum_{b=1}^g n_b m_b\right)^2}{\left(\sum_{b=1}^g n_b m_b\right)^2} \overline{\Phi}_2 p_i (1 - p_i) + \frac{\left(\sum_{b=1}^g n_b m_b\right)^2}{\left(\sum_{b=1}^g n_b m_b\right)^2} p_i^2 \\
&= \overline{\Phi}_2 p_i (1 - p_i) + p_i^2. \quad \square
\end{aligned}$$

*Proof of eq. 3.11.*  $\widehat{\mathbb{P}}[\text{IBS}] = \sum_{i=1}^I \widehat{p}_i^2$ . We need only show that  $\mathbb{P}[\text{IBD}] = \overline{\Phi}_2$ . Note that while we write  $\widehat{\mathbb{P}}[\text{IBS}]$  as an estimate,  $\mathbb{P}[\text{IBD}]$  depends only on quantities that are treated as known with certainty and we write it as a known quantity itself. Consider two alleles from the sample (that are not necessarily distinct). Let  $C_{(a,j)(b,k)}$  denote the event that the first of the two alleles is from individual  $(a, j)$  and the second is from individual  $(b, k)$ , where  $(a, j)$  and  $(b, k)$  are not necessarily distinct. Supposing that the two alleles are drawn uniformly at random from the sample, with replacement, let  $\mathbb{P}[C_{(a,j)(b,k)}]$  denote the probability of event  $C_{(a,j)(b,k)}$ . Let  $\mathbb{P}[\text{IBD}|C_{(a,j)(b,k)}]$  be the probability that two alleles are IBD given that the first allele is chosen from individual  $(a, j)$  and the second is chosen from individual  $(b, k)$ . Then

$$\begin{aligned}
\mathbb{P}[\text{IBD}] &= \sum_{b=1}^g \left\{ \sum_{k=1}^{n_b} \mathbb{P}[\text{IBD}|C_{(b,k)(b,k)}] \mathbb{P}[C_{(b,k)(b,k)}] \right. \\
&\quad \left. + \sum_{j=1}^{n_b} \sum_{\substack{k=1 \\ k \neq j}}^{n_b} \mathbb{P}[\text{IBD}|C_{(b,j)(b,k)}] \mathbb{P}[C_{(b,j)(b,k)}] \right\} \\
&\quad + \sum_{a=1}^g \sum_{\substack{b=1 \\ b \neq a}}^g \sum_{j=1}^{n_a} \sum_{k=1}^{n_b} \mathbb{P}[\text{IBD}|C_{(a,j)(b,k)}] \mathbb{P}[C_{(a,j)(b,k)}].
\end{aligned}$$

Note that, for individuals  $(a, j)$  and  $(b, k)$ , which are not necessarily distinct,

$$\mathbb{P}[C_{(a,j)(b,k)}] = \left( \frac{m_a}{\sum_{c=1}^g n_c m_c} \right) \left( \frac{m_b}{\sum_{c=1}^g n_c m_c} \right) = \frac{m_a m_b}{(\sum_{c=1}^g n_c m_c)^2}$$

$$\mathbb{P}[\text{IBD}|C_{(a,j)(b,k)}] = \Phi_{(a,j)(b,k)}.$$

It follows that

$$\begin{aligned}
\mathbb{P}[\text{IBD}] &= \sum_{b=1}^g \left\{ \sum_{k=1}^{n_b} \Phi_{(b,k)(b,k)} \frac{m_b^2}{(\sum_{c=1}^g n_c m_c)^2} + \sum_{j=1}^{n_b} \sum_{\substack{k=1 \\ k \neq j}}^{n_b} \Phi_{(b,j)(b,k)} \frac{m_b^2}{(\sum_{c=1}^g n_c m_c)^2} \right\} \\
&\quad + \sum_{a=1}^g \sum_{\substack{b=1 \\ b \neq a}}^g \sum_{j=1}^{n_a} \sum_{k=1}^{n_b} \Phi_{(a,j)(b,k)} \frac{m_a m_b}{(\sum_{c=1}^g n_c m_c)^2} \\
&= \frac{1}{(\sum_{b=1}^g n_b m_b)^2} \sum_{a=1}^g \sum_{b=1}^g \sum_{j=1}^{n_a} \sum_{k=1}^{n_b} m_a m_b \Phi_{(a,j)(b,k)} \\
&= \bar{\Phi}_2. \quad \square
\end{aligned}$$

*Proof of eq. 3.12.* For an  $m_b$ -ploid individual  $k$ ,  $\Phi_{(b,k)(b,k)} = 1/m_b + (1 - 1/m_b)f_{(b,k)} = (1/m_b)[1 + (m_b - 1)f_{(b,k)}]$ . Note that  $\Phi_{(a,j)(b,k)} = 0$  if individuals  $(a, j)$  and  $(b, k)$  are unrelated. We can then break  $\bar{\Phi}_2$  into three components, considering three different types of pairs of individuals: same group-same individual, same group-different individual, and different group. Therefore

$$\begin{aligned}
\bar{\Phi}_2 &= \frac{1}{\left(\sum_{b=1}^g n_b m_b\right)^2} \sum_{a=1}^g \sum_{b=1}^g \sum_{j=1}^{n_a} \sum_{k=1}^{n_b} m_a m_b \Phi_{(a,j)(b,k)} \\
&= \frac{1}{\left(\sum_{b=1}^g n_b m_b\right)^2} \left[ \sum_{b=1}^g \sum_{k=1}^{n_b} m_b^2 \Phi_{(b,k)(b,k)} + 2 \sum_{b=1}^g \sum_{j=1}^{n_b-1} \sum_{k=j+1}^{n_b} m_b^2 \Phi_{(b,j)(b,k)} \right. \\
&\quad \left. + 2 \sum_{a=1}^{g-1} \sum_{b=a+1}^g \sum_{j=1}^{n_a} \sum_{k=1}^{n_b} m_a m_b \Phi_{(a,j)(b,k)} \right] \\
&= \frac{1}{\left(\sum_{b=1}^g n_b m_b\right)^2} \left[ \sum_{b=1}^g \sum_{k=1}^{n_b} m_b^2 \frac{1}{m_b} [1 + (m_b - 1) f_{(b,k)}] + 2 \sum_{b=1}^g \sum_{R \in \mathcal{G}_{b,b}} m_b^2 \eta_R \Phi_R \right. \\
&\quad \left. + 2 \sum_{a=1}^{g-1} \sum_{b=a+1}^g \sum_{R \in \mathcal{G}_{a,b}} m_a m_b \eta_R \Phi_R \right] \\
&= \frac{1}{\left(\sum_{b=1}^g n_b m_b\right)^2} \left[ \sum_{b=1}^g n_b m_b + \sum_{b=1}^g n_b m_b (m_b - 1) \bar{f}_b + 2 \sum_{b=1}^g \sum_{R \in \mathcal{G}_{b,b}} m_b^2 \eta_R \Phi_R \right. \\
&\quad \left. + 2 \sum_{a=1}^{g-1} \sum_{b=a+1}^g \sum_{R \in \mathcal{G}_{a,b}} m_a m_b \eta_R \Phi_R \right]. \quad \square
\end{aligned}$$

*Proof of eq. 3.13.* First we note that

$$1 - \bar{\Phi}_2 = \frac{D}{\left(\sum_{b=1}^g n_b m_b\right)^2}.$$

Substituting  $1 - \bar{\Phi}_2$  into  $\tilde{H}$  (eq. 3.10) gives

$$\tilde{H} = \frac{\left(\sum_{b=1}^g n_b m_b\right)^2}{D} \left(1 - \sum_{i=1}^I \hat{p}_i^2\right).$$

Rearranging eq. 3.3 we get

$$1 - \sum_{i=1}^I \hat{p}_i^2 = \frac{\sum_{b=1}^g n_b m_b - 1}{\sum_{b=1}^g n_b m_b} \hat{H},$$

from which

$$\begin{aligned}\tilde{H} &= \frac{\left(\sum_{b=1}^g n_b m_b\right)^2}{D} \left( \frac{\sum_{b=1}^g n_b m_b - 1}{\sum_{b=1}^g n_b m_b} \hat{H} \right) \\ &= \frac{\left(\sum_{b=1}^g n_b m_b\right) \left(\sum_{b=1}^g n_b m_b - 1\right)}{D} \hat{H}.\end{aligned}\quad \square$$

### 3.7 Appendix B

In this section, we present results that aid in the derivation of the variance of our gene diversity estimator. Lemma III.3 derives certain expectations involving four alleles. These expectations are used to calculate the variance and covariance of squared allele frequency estimates in Lemma III.4. Lemma III.4 is then used to prove the variance formula in Theorem III.2 when related and inbred individuals are included in a sample.

**Lemma III.3.** *Consider a locus with  $I$  distinct alleles, allele frequencies  $p_i \in [0, 1]$  and  $\sum_{i=1}^I p_i = 1$ . Suppose a sample from a population has  $g$  groups, each with different ploidy, and  $n_b$   $m_b$ -ploid individuals in group  $b$ ,  $b = 1, 2, \dots, g$ , each of whom is possibly inbred and related to other individuals in the sample. Consider the  $l$ th allele of individual  $(a, j)$ , the  $t$ th allele of individual  $(b, k)$ , the  $l'$ th allele of individual  $(a', j')$ , and the  $t'$ th allele of individual  $(b', k')$ . For clarity, let  $w = (a, j)$ ,  $x = (b, k)$ ,  $y = (a', j')$ , and  $z = (b', k')$ . Then for allelic types  $i$  and  $i' \neq i$ ,*

$$\begin{aligned}
\mathbb{E}\left[A_{w,\ell}^{(i)}A_{x,t}^{(i)}A_{y,\ell'}^{(i)}A_{z,t'}^{(i)}\right] &= \Phi_{wxyz}p_i + [\Phi_{wxy} + \Phi_{wxz} + \Phi_{wyz} + \Phi_{xyz} + \Phi_{wx,yz} \\
&\quad + \Phi_{wy,xz} + \Phi_{wz,xy} - 7\Phi_{wxyz}]p_i^2 \\
&+ [12\Phi_{wxyz} + (\Phi_{wx} + \Phi_{wy} + \Phi_{wz} + \Phi_{xy} + \Phi_{xz} + \Phi_{yz}) \\
&\quad - 3(\Phi_{wxy} + \Phi_{wxz} + \Phi_{wyz} + \Phi_{xyz}) \\
&\quad - 2(\Phi_{wx,yz} + \Phi_{wy,xz} + \Phi_{wz,xy})]p_i^3 \\
&+ [1 + (\Phi_{wx,yz} + \Phi_{wy,xz} + \Phi_{wz,xy}) \\
&\quad + 2(\Phi_{wxy} + \Phi_{wxz} + \Phi_{wyz} + \Phi_{xyz}) - 6\Phi_{wxyz} \\
&\quad - (\Phi_{wx} + \Phi_{wy} + \Phi_{wz} + \Phi_{xy} + \Phi_{xz} + \Phi_{yz})]p_i^4 \quad (3.22)
\end{aligned}$$

$$\begin{aligned}
\mathbb{E}\left[A_{w,\ell}^{(i)}A_{x,t}^{(i)}A_{y,\ell'}^{(i')}A_{z,t'}^{(i')}\right] &= [\Phi_{wx,yz} - \Phi_{wxyz}]p_i p_{i'} \\
&+ [2\Phi_{wxyz} + \Phi_{wx} - (\Phi_{wxy} + \Phi_{wxz}) - \Phi_{wx,yz}]p_i p_{i'}^2 \\
&+ [2\Phi_{wxyz} + \Phi_{yz} - (\Phi_{wyz} + \Phi_{xyz}) - \Phi_{wx,yz}]p_i^2 p_{i'} \\
&+ [1 + \Phi_{wx,yz} + \Phi_{wy,xz} + \Phi_{wz,xy} \\
&\quad + 2(\Phi_{wxy} + \Phi_{wxz} + \Phi_{wyz} + \Phi_{xyz}) - 6\Phi_{wxyz} \\
&\quad - (\Phi_{wx} + \Phi_{wy} + \Phi_{wz} + \Phi_{xy} + \Phi_{xz} + \Phi_{yz})]p_i^2 p_{i'}^2. \quad (3.23)
\end{aligned}$$

*Proof.* We need to evaluate

$$\mathbb{E}\left[A_{w,\ell}^{(i)}A_{x,t}^{(i)}A_{y,\ell'}^{(i')}A_{z,t'}^{(i')}\right] = \sum_{s=1}^{15} \Delta_s \mathbb{P}\left[A_{w,\ell}^{(i)} = 1, A_{x,t}^{(i)} = 1, A_{y,\ell'}^{(i')} = 1, A_{z,t'}^{(i')} = 1 \mid S = s\right],$$

where  $S$  represents one of the 15 identity states in Figure 3.7 for four alleles—one from  $w$ , one from  $x$ , one from  $y$ , and one from  $z$ —and  $\Delta_s$  is the identity coefficient, the probability of observing state  $S = s$  for four alleles randomly chosen, one from  $w$ , one from  $x$ , one from  $y$ , and one from  $z$ . We can rewrite the identity coefficients in terms of kinship coefficients by using the following relationships:

$$\begin{aligned}
\sum_{s=1}^{15} \Delta_s &= 1 \\
\Phi_{wxyz} &= \Delta_1 \\
\Phi_{wxy} &= \Delta_1 + \Delta_2 \\
\Phi_{wxz} &= \Delta_1 + \Delta_3 \\
\Phi_{wyz} &= \Delta_1 + \Delta_4 \\
\Phi_{xyz} &= \Delta_1 + \Delta_5 \\
\Phi_{wx,yz} &= \Delta_1 + \Delta_6 \\
\Phi_{wy,xz} &= \Delta_1 + \Delta_9 \\
\Phi_{wz,xy} &= \Delta_1 + \Delta_{12} \\
\Phi_{wx} &= \Delta_1 + \Delta_2 + \Delta_3 + \Delta_6 + \Delta_7 \\
\Phi_{wy} &= \Delta_1 + \Delta_2 + \Delta_4 + \Delta_9 + \Delta_{10} \\
\Phi_{wz} &= \Delta_1 + \Delta_3 + \Delta_4 + \Delta_{12} + \Delta_{13} \\
\Phi_{xy} &= \Delta_1 + \Delta_2 + \Delta_5 + \Delta_{12} + \Delta_{14} \\
\Phi_{xz} &= \Delta_1 + \Delta_3 + \Delta_5 + \Delta_9 + \Delta_{11} \\
\Phi_{yz} &= \Delta_1 + \Delta_4 + \Delta_5 + \Delta_6 + \Delta_8.
\end{aligned} \tag{3.24}$$

Note that the  $\Delta$  coefficients above are identical to the  $\delta$  coefficients in *Cockerham* (1971). Also, the  $\Phi$  coefficients involving two individuals, three individuals, and pairs of pairs of individuals are identical to Cockerham's  $\theta$ ,  $\gamma$ , and  $\Delta$  coefficients, respectively (*Cockerham*, 1971). If  $i' = i$ , we get

$$\mathbb{E} \left[ A_{w,\ell}^{(i)} A_{x,t}^{(i)} A_{y,\ell'}^{(i)} A_{z,t'}^{(i)} \right] = \Delta_1 p_i + (\Delta_2 + \Delta_3 + \Delta_4 + \Delta_5 + \Delta_6 + \Delta_9 + \Delta_{12}) p_i^2. \tag{3.25}$$

If  $i \neq i'$ , we get



$$\mathbb{E} \left[ A_{w,\ell}^{(i)} A_{x,t}^{(i)} A_{y,\ell'}^{(i')} A_{z,t'}^{(i')} \right] = \Delta_6 p_i p_{i'} + \Delta_7 p_i p_{i'}^2 + \Delta_8 p_i^2 p_{i'} + \Delta_{15} p_i^2 p_{i'}^2. \quad (3.26)$$

The desired result follows by substituting equations (3.24) into equations (3.25) and (3.26).  $\square$

Note that expressions mathematically identical to eqs. 3.22 and 3.23 except with different notation appear in Table 1 of *Cockerham* (1971). However, a slight conceptual difference is that our formulas involve an expectation of a product among four arbitrary alleles, not necessarily four alleles in two pairs of diploid genotypes. We now use Lemma III.3 to derive  $Var[\hat{p}_i^2]$  and  $Cov(\hat{p}_i^2, \hat{p}_{i'}^2)$ .

**Lemma III.4.** *Consider a locus with  $I$  distinct alleles, allele frequencies  $p_i \in [0, 1]$  and  $\sum_{i=1}^I p_i = 1$ . Suppose a sample from a population has  $g$  groups, each with different ploidy, and  $n_b$   $m_b$ -ploid individuals in group  $b$ ,  $b = 1, 2, \dots, g$ , each of whom is possibly inbred and related to other individuals in the sample. Then for allelic types  $i$  and  $i' \neq i$ ,*

$$\begin{aligned} \mathbb{E}[\hat{p}_i^4] &= \bar{\Phi}_4 p_i + [4\bar{\Phi}_3 + 3\bar{\Phi}_{2,2} - 7\bar{\Phi}_4] p_i^2 + [12\bar{\Phi}_4 + 6\bar{\Phi}_2 - 12\bar{\Phi}_3 - 6\bar{\Phi}_{2,2}] p_i^3 \\ &\quad + [1 + 3\bar{\Phi}_{2,2} + 8\bar{\Phi}_3 - 6\bar{\Phi}_4 - 6\bar{\Phi}_2] p_i^4 \end{aligned} \quad (3.27)$$

$$\begin{aligned} \mathbb{E}[\hat{p}_i^2 \hat{p}_{i'}^2] &= [\bar{\Phi}_{2,2} - \bar{\Phi}_4] p_i p_{i'} + [2\bar{\Phi}_4 + \bar{\Phi}_2 - 2\bar{\Phi}_3 - \bar{\Phi}_{2,2}] p_i p_{i'}^2 \\ &\quad + [2\bar{\Phi}_4 + \bar{\Phi}_2 - 2\bar{\Phi}_3 - \bar{\Phi}_{2,2}] p_i^2 p_{i'} \\ &\quad + [1 + 3\bar{\Phi}_{2,2} + 8\bar{\Phi}_3 - 6\bar{\Phi}_4 - 6\bar{\Phi}_2] p_i^2 p_{i'}^2 \end{aligned} \quad (3.28)$$

and therefore

$$\begin{aligned}
Var[\hat{p}_i^2] &= \bar{\Phi}_4 p_i + [4\bar{\Phi}_3 + 3\bar{\Phi}_{2,2} - 7\bar{\Phi}_4 - \bar{\Phi}_2^2] p_i^2 \\
&\quad + [12\bar{\Phi}_4 + 4\bar{\Phi}_2 + 2\bar{\Phi}_2^2 - 12\bar{\Phi}_3 - 6\bar{\Phi}_{2,2}] p_i^3 \\
&\quad + [3\bar{\Phi}_{2,2} + 8\bar{\Phi}_3 - 6\bar{\Phi}_4 - 4\bar{\Phi}_2 - \bar{\Phi}_2^2] p_i^4
\end{aligned} \tag{3.29}$$

$$\begin{aligned}
Cov(\hat{p}_i^2, \hat{p}_{i'}^2) &= [\bar{\Phi}_{2,2} - \bar{\Phi}_4 - \bar{\Phi}_2^2] p_i p_{i'} + [2\bar{\Phi}_4 + \bar{\Phi}_2^2 - 2\bar{\Phi}_3 - \bar{\Phi}_{2,2}] p_i p_{i'}^2 \\
&\quad + [2\bar{\Phi}_4 + \bar{\Phi}_2^2 - 2\bar{\Phi}_3 - \bar{\Phi}_{2,2}] p_i^2 p_{i'} \\
&\quad + [3\bar{\Phi}_{2,2} + 8\bar{\Phi}_3 - 6\bar{\Phi}_4 - 4\bar{\Phi}_2 - \bar{\Phi}_2^2] p_i^2 p_{i'}^2.
\end{aligned} \tag{3.30}$$

*Proof.* Applying the definition of  $\hat{p}_i$ , we have

$$\begin{aligned}
\mathbb{E}[\widehat{p}_i^4] &= \frac{1}{\left(\sum_{b=1}^g n_b m_b\right)^4} \sum_{a=1}^g \sum_{b=1}^g \sum_{a'=1}^g \sum_{b'=1}^g \sum_{j=1}^{n_a} \sum_{k=1}^{n_b} \sum_{j'=1}^{n_{a'}} \sum_{k'=1}^{n_{b'}} \sum_{\ell=1}^{m_a} \sum_{t=1}^{m_b} \sum_{\ell'=1}^{m_{a'}} \sum_{t'=1}^{m_{b'}} \\
&\quad \times \mathbb{E}\left[A_{(a,j),\ell}^{(i)} A_{(b,k),t}^{(i)} A_{(a',j'),\ell'}^{(i)} A_{(b',k'),t'}^{(i)}\right] \\
&= \frac{1}{\left(\sum_{b=1}^g n_b m_b\right)^4} \sum_{a=1}^g \sum_{b=1}^g \sum_{a'=1}^g \sum_{b'=1}^g \sum_{j=1}^{n_a} \sum_{k=1}^{n_b} \sum_{j'=1}^{n_{a'}} \sum_{k'=1}^{n_{b'}} m_a m_b m_{a'} m_{b'} \\
&\quad \times \left\{ \Phi_{(a,j)(b,k)(a',j')(b',k')} p_i \right. \\
&\quad + [\Phi_{(a,j)(b,k)(a',j')} + \Phi_{(a,j)(b,k)(b',k')} + \Phi_{(a,j)(a',j')(b',k')} + \Phi_{(b,k)(a',j')(b',k')} \\
&\quad + \Phi_{(a,j)(b,k),(a',j')(b',k')} + \Phi_{(a,j)(a',j'),(b,k)(b',k')} + \Phi_{(a,j)(b',k'),(b,k)(a',j')} \\
&\quad \left. - 7\Phi_{(a,j)(b,k)(a',j')(b',k')} \right] p_i^2 \\
&\quad + [12\Phi_{(a,j)(b,k)(a',j')(b',k')} + \Phi_{(a,j)(b,k)} + \Phi_{(a,j)(a',j')} + \Phi_{(a,j)(b',k')} \\
&\quad + \Phi_{(b,k)(a',j')} + \Phi_{(b,k)(b',k')} + \Phi_{(a',j')(b',k')} \\
&\quad - 3(\Phi_{(a,j)(b,k)(a',j')} + \Phi_{(a,j)(b,k)(b',k')} + \Phi_{(a,j)(a',j')(b',k')} + \Phi_{(b,k)(a',j')(b',k')}) \\
&\quad - 2(\Phi_{(a,j)(b,k),(a',j')(b',k')} + \Phi_{(a,j)(a',j'),(b,k)(b',k')} + \Phi_{(a,j)(b',k'),(b,k)(a',j')}] p_i^3 \\
&\quad + [1 + \Phi_{(a,j)(b,k),(a',j')(b',k')} + \Phi_{(a,j)(a',j'),(b,k)(b',k')} + \Phi_{(a,j)(b',k'),(b,k)(a',j')} \\
&\quad + 2(\Phi_{(a,j)(b,k)(a',j')} + \Phi_{(a,j)(b,k)(b',k')} + \Phi_{(a,j)(a',j')(b',k')} + \Phi_{(b,k)(a',j')(b',k')}) \\
&\quad - 6\Phi_{(a,j)(b,k)(a',j')(b',k')} \\
&\quad - (\Phi_{(a,j)(b,k)} + \Phi_{(a,j)(a',j')} + \Phi_{(a,j)(b',k')} + \Phi_{(b,k)(a',j')} + \Phi_{(b,k)(b',k')} \\
&\quad \left. + \Phi_{(a',j')(b',k')}) \right] p_i^4 \Big\} \\
&= \overline{\Phi}_4 p_i + [4\overline{\Phi}_3 + 3\overline{\Phi}_{2,2} - 7\overline{\Phi}_4] p_i^2 + [12\overline{\Phi}_4 + 6\overline{\Phi}_2 - 12\overline{\Phi}_3 - 6\overline{\Phi}_{2,2}] p_i^3 \\
&\quad + [1 + 3\overline{\Phi}_{2,2} + 8\overline{\Phi}_3 - 6\overline{\Phi}_4 - 6\overline{\Phi}_2] p_i^4.
\end{aligned}$$

For the case with alleles  $i$  and  $i' \neq i$ , we have

$$\begin{aligned}
\mathbb{E}[\widehat{p}_i^2 \widehat{p}_{i'}^2] &= \frac{1}{\left(\sum_{b=1}^g n_b m_b\right)^4} \sum_{a=1}^g \sum_{b=1}^g \sum_{a'=1}^g \sum_{b'=1}^g \sum_{j=1}^{n_a} \sum_{k=1}^{n_b} \sum_{j'=1}^{n_{a'}} \sum_{k'=1}^{n_{b'}} \sum_{\ell=1}^{m_a} \sum_{t=1}^{m_b} \sum_{\ell'=1}^{m_{a'}} \sum_{t'=1}^{m_{b'}} \\
&\quad \times \mathbb{E}\left[A_{(a,j),\ell}^{(i)} A_{(b,k),t}^{(i)} A_{(a',j'),\ell'}^{(i')} A_{(b',k'),t'}^{(i')}\right] \\
&= \frac{1}{\left(\sum_{b=1}^g n_b m_b\right)^4} \sum_{a=1}^g \sum_{b=1}^g \sum_{a'=1}^g \sum_{b'=1}^g \sum_{j=1}^{n_a} \sum_{k=1}^{n_b} \sum_{j'=1}^{n_{a'}} \sum_{k'=1}^{n_{b'}} m_a m_b m_{a'} m_{b'} \\
&\quad \times \left\{ [\Phi_{(a,j)(b,k),(a',j')(b',k')} - \bar{\Phi}_{(a,j)(b,k)(a',j')(b',k')}] p_i p_{i'} \right. \\
&\quad + [2\Phi_{(a,j)(b,k)(a',j')(b',k')} + \Phi_{(a,j)(b,k)} \\
&\quad \quad - (\Phi_{(a,j)(b,k)(a',j')} + \Phi_{(a,j)(b,k)(b',k')}) - \Phi_{(a,j)(b,k),(a',j')(b',k')}] p_i p_{i'}^2 \\
&\quad + [2\Phi_{(a,j)(b,k)(a',j')(b',k')} + \Phi_{(a',j')(b',k')} \\
&\quad \quad - (\Phi_{(a,j)(a',j')(b',k')} + \Phi_{(b,k)(a',j')(b',k')}) - \Phi_{(a,j)(b,k),(a',j')(b',k')}] p_i^2 p_{i'} \\
&\quad + [1 + \Phi_{(a,j)(b,k),(a',j')(b',k')} + \Phi_{(a,j)(a',j')(b,k)(b',k')} + \Phi_{(a,j)(b',k')(b,k)(a',j')} \\
&\quad \quad + 2(\Phi_{(a,j)(b,k)(a',j')} + \Phi_{(a,j)(b,k)(b',k')} + \Phi_{(a,j)(a',j')(b',k')} + \Phi_{(b,k)(a',j')(b',k')}) \\
&\quad \quad - 6\Phi_{(a,j)(b,k)(a',j')(b',k')} \\
&\quad \quad - (\Phi_{(a,j)(b,k)} + \Phi_{(a,j)(a',j')} + \Phi_{(a,j)(b',k')} + \Phi_{(b,k)(a',j')} + \Phi_{(b,k)(b',k')} \\
&\quad \quad \left. + \Phi_{(a',j')(b',k')})] p_i^2 p_{i'}^2 \right\} \\
&= [\bar{\Phi}_{2,2} - \bar{\Phi}_4] p_i p_{i'} + [2\bar{\Phi}_4 + \bar{\Phi}_2 - 2\bar{\Phi}_3 - \bar{\Phi}_{2,2}] p_i p_{i'}^2 \\
&\quad + [2\bar{\Phi}_4 + \bar{\Phi}_2 - 2\bar{\Phi}_3 - \bar{\Phi}_{2,2}] p_i^2 p_{i'} \\
&\quad + [1 + 3\bar{\Phi}_{2,2} + 8\bar{\Phi}_3 - 6\bar{\Phi}_4 - 6\bar{\Phi}_2] p_i^2 p_{i'}^2.
\end{aligned}$$

Applying the definition of variance, we have

$$\begin{aligned}
\text{Var}[\widehat{p}_i^2] &= \mathbb{E}[\widehat{p}_i^4] - (\mathbb{E}[\widehat{p}_i^2])^2 \\
&= \overline{\Phi}_4 p_i + [4\overline{\Phi}_3 + 3\overline{\Phi}_{2,2} - 7\overline{\Phi}_4] p_i^2 + [12\overline{\Phi}_4 + 6\overline{\Phi}_2 - 12\overline{\Phi}_3 - 6\overline{\Phi}_{2,2}] p_i^3 \\
&\quad + [1 + 3\overline{\Phi}_{2,2} + 8\overline{\Phi}_3 - 6\overline{\Phi}_4 - 6\overline{\Phi}_2] p_i^4 - [\overline{\Phi}_2 p_i (1 - p_i) + p_i^2]^2 \\
&= \overline{\Phi}_4 p_i + [4\overline{\Phi}_3 + 3\overline{\Phi}_{2,2} - 7\overline{\Phi}_4 - \overline{\Phi}_2^2] p_i^2 + [12\overline{\Phi}_4 + 4\overline{\Phi}_2 + 2\overline{\Phi}_2^2 - 12\overline{\Phi}_3 - 6\overline{\Phi}_{2,2}] p_i^3 \\
&\quad + [3\overline{\Phi}_{2,2} + 8\overline{\Phi}_3 - 6\overline{\Phi}_4 - 4\overline{\Phi}_2 - \overline{\Phi}_2^2] p_i^4.
\end{aligned}$$

Applying the definition of covariance, we have

$$\begin{aligned}
\text{Cov}(\widehat{p}_i^2, \widehat{p}_{i'}^2) &= \mathbb{E}[\widehat{p}_i^2 \widehat{p}_{i'}^2] - \mathbb{E}[\widehat{p}_i^2] \mathbb{E}[\widehat{p}_{i'}^2] \\
&= [\overline{\Phi}_{2,2} - \overline{\Phi}_4] p_i p_{i'} + [2\overline{\Phi}_4 + \overline{\Phi}_2 - 2\overline{\Phi}_3 - \overline{\Phi}_{2,2}] p_i p_{i'}^2 \\
&\quad + [2\overline{\Phi}_4 + \overline{\Phi}_2 - 2\overline{\Phi}_3 - \overline{\Phi}_{2,2}] p_i^2 p_{i'} \\
&\quad + [1 + 3\overline{\Phi}_{2,2} + 8\overline{\Phi}_3 - 6\overline{\Phi}_4 - 6\overline{\Phi}_2] p_i^2 p_{i'}^2 \\
&\quad - [\overline{\Phi}_2 p_i (1 - p_i) + p_i^2] [\overline{\Phi}_2 p_{i'} (1 - p_{i'}) + p_{i'}^2] \\
&= [\overline{\Phi}_{2,2} - \overline{\Phi}_4 - \overline{\Phi}_2^2] p_i p_{i'} + [2\overline{\Phi}_4 + \overline{\Phi}_2^2 - 2\overline{\Phi}_3 - \overline{\Phi}_{2,2}] p_i p_{i'}^2 \\
&\quad + [2\overline{\Phi}_4 + \overline{\Phi}_2^2 - 2\overline{\Phi}_3 - \overline{\Phi}_{2,2}] p_i^2 p_{i'} \\
&\quad + [3\overline{\Phi}_{2,2} + 8\overline{\Phi}_3 - 6\overline{\Phi}_4 - 4\overline{\Phi}_2 - \overline{\Phi}_2^2] p_i^2 p_{i'}^2. \quad \square
\end{aligned}$$

We now utilize Lemma III.4 to prove Theorem III.2.

*Proof of Theorem III.2.* Applying the definition of variance, we have

$$\begin{aligned}
\text{Var} \left[ 1 - \sum_{i=1}^I \widehat{p}_i^2 \right] &= \sum_{i=1}^I \sum_{i'=1}^I \text{Cov}(\widehat{p}_i^2, \widehat{p}_{i'}^2) \\
&= \sum_{i=1}^I \text{Var}[\widehat{p}_i^4] + 2 \sum_{i=1}^{I-1} \sum_{i'=i+1}^I \text{Cov}(\widehat{p}_i^2, \widehat{p}_{i'}^2) \\
&= \sum_{i=1}^I \left\{ \bar{\Phi}_4 p_i + [4\bar{\Phi}_3 + 3\bar{\Phi}_{2,2} - 7\bar{\Phi}_4 - \bar{\Phi}_2^2] p_i^2 \right. \\
&\quad + [12\bar{\Phi}_4 + 4\bar{\Phi}_2 + 2\bar{\Phi}_2^2 - 12\bar{\Phi}_3 - 6\bar{\Phi}_{2,2}] p_i^3 \\
&\quad \left. + [3\bar{\Phi}_{2,2} + 8\bar{\Phi}_3 - 6\bar{\Phi}_4 - 4\bar{\Phi}_2 - \bar{\Phi}_2^2] p_i^4 \right\} \\
&\quad + 2 \sum_{i=1}^{I-1} \sum_{i'=i+1}^I \left\{ [\bar{\Phi}_{2,2} - \bar{\Phi}_4 - \bar{\Phi}_2^2] p_i p_{i'} + [2\bar{\Phi}_4 + \bar{\Phi}_2^2 - 2\bar{\Phi}_3 - \bar{\Phi}_{2,2}] p_i p_{i'}^2 \right. \\
&\quad + [2\bar{\Phi}_4 + \bar{\Phi}_2^2 - 2\bar{\Phi}_3 - \bar{\Phi}_{2,2}] p_i^2 p_{i'} \\
&\quad \left. + [3\bar{\Phi}_{2,2} + 8\bar{\Phi}_3 - 6\bar{\Phi}_4 - 4\bar{\Phi}_2 - \bar{\Phi}_2^2] p_i^2 p_{i'}^2 \right\} \\
&= \bar{\Phi}_4 + [4\bar{\Phi}_3 + 3\bar{\Phi}_{2,2} - 7\bar{\Phi}_4 - \bar{\Phi}_2^2] \sum_{i=1}^I p_i^2 \\
&\quad + [12\bar{\Phi}_4 + 4\bar{\Phi}_2 + 2\bar{\Phi}_2^2 - 12\bar{\Phi}_3 - 6\bar{\Phi}_{2,2}] \sum_{i=1}^I p_i^3 \\
&\quad + [3\bar{\Phi}_{2,2} + 8\bar{\Phi}_3 - 6\bar{\Phi}_4 - 4\bar{\Phi}_2 - \bar{\Phi}_2^2] \sum_{i=1}^I p_i^4 \\
&\quad + 2[\bar{\Phi}_{2,2} - \bar{\Phi}_4 - \bar{\Phi}_2^2] \sum_{i=1}^{I-1} \sum_{i'=i+1}^I p_i p_{i'} \\
&\quad + 2[2\bar{\Phi}_4 + \bar{\Phi}_2^2 - 2\bar{\Phi}_3 - \bar{\Phi}_{2,2}] \sum_{i=1}^{I-1} \sum_{i'=i+1}^I p_i p_{i'}^2 \\
&\quad + 2[2\bar{\Phi}_4 + \bar{\Phi}_2^2 - 2\bar{\Phi}_3 - \bar{\Phi}_{2,2}] \sum_{i=1}^{I-1} \sum_{i'=i+1}^I p_i^2 p_{i'} \\
&\quad + 2[3\bar{\Phi}_{2,2} + 8\bar{\Phi}_3 - 6\bar{\Phi}_4 - 4\bar{\Phi}_2 - \bar{\Phi}_2^2] \sum_{i=1}^{I-1} \sum_{i'=i+1}^I p_i^2 p_{i'}^2.
\end{aligned}$$

Simplifying, we get

$$\begin{aligned}
\text{Var} \left[ 1 - \sum_{i=1}^I \widehat{p}_i^2 \right] &= \bar{\Phi}_4 + 2[2\bar{\Phi}_3 + \bar{\Phi}_{2,2} - 3\bar{\Phi}_4] \sum_{i=1}^I p_i^2 + 4[2\bar{\Phi}_4 + \bar{\Phi}_2 - 2\bar{\Phi}_3 - \bar{\Phi}_{2,2}] \sum_{i=1}^I p_i^3 \\
&\quad + [\bar{\Phi}_{2,2} - \bar{\Phi}_4 - \bar{\Phi}_2^2] \left( \sum_{i=1}^I p_i^2 + 2 \sum_{i=1}^{I-1} \sum_{i'=i+1}^I p_i p_{i'} \right) \\
&\quad + [2\bar{\Phi}_4 + \bar{\Phi}_2^2 - 2\bar{\Phi}_3 - \bar{\Phi}_{2,2}] \left( 2 \sum_{i=1}^I p_i^3 + 2 \sum_{i=1}^{I-1} \sum_{i'=i+1}^I [p_i^2 p_{i'} + p_i p_{i'}^2] \right) \\
&\quad + [3\bar{\Phi}_{2,2} + 8\bar{\Phi}_3 - 6\bar{\Phi}_4 - 4\bar{\Phi}_2 - \bar{\Phi}_2^2] \left( \sum_{i=1}^I p_i^4 + 2 \sum_{i=1}^{I-1} \sum_{i'=i+1}^I p_i^2 p_{i'}^2 \right) \\
&= \bar{\Phi}_4 + 2[2\bar{\Phi}_3 + \bar{\Phi}_{2,2} - 3\bar{\Phi}_4] \sum_{i=1}^I p_i^2 + 4[2\bar{\Phi}_4 + \bar{\Phi}_2 - 2\bar{\Phi}_3 - \bar{\Phi}_{2,2}] \sum_{i=1}^I p_i^3 \\
&\quad + [\bar{\Phi}_{2,2} - \bar{\Phi}_4 - \bar{\Phi}_2^2] \sum_{i=1}^I \sum_{i'=1}^I p_i p_{i'} \\
&\quad + 2[2\bar{\Phi}_4 + \bar{\Phi}_2^2 - 2\bar{\Phi}_3 - \bar{\Phi}_{2,2}] \sum_{i=1}^I \sum_{i'=1}^I p_i^2 p_{i'} \\
&\quad + [3\bar{\Phi}_{2,2} + 8\bar{\Phi}_3 - 6\bar{\Phi}_4 - 4\bar{\Phi}_2 - \bar{\Phi}_2^2] \sum_{i=1}^I \sum_{i'=1}^I p_i^2 p_{i'}^2 \\
&= \bar{\Phi}_4 + 2[2\bar{\Phi}_3 + \bar{\Phi}_{2,2} - 3\bar{\Phi}_4] \sum_{i=1}^I p_i^2 + 4[2\bar{\Phi}_4 + \bar{\Phi}_2 - 2\bar{\Phi}_3 - \bar{\Phi}_{2,2}] \sum_{i=1}^I p_i^3 \\
&\quad + [\bar{\Phi}_{2,2} - \bar{\Phi}_4 - \bar{\Phi}_2^2] \left( \sum_{i=1}^I p_i \right)^2 \\
&\quad + 2[2\bar{\Phi}_4 + \bar{\Phi}_2^2 - 2\bar{\Phi}_3 - \bar{\Phi}_{2,2}] \left( \sum_{i=1}^I p_i^2 \right) \left( \sum_{i=1}^I p_i \right) \\
&\quad + [3\bar{\Phi}_{2,2} + 8\bar{\Phi}_3 - 6\bar{\Phi}_4 - 4\bar{\Phi}_2 - \bar{\Phi}_2^2] \left( \sum_{i=1}^I p_i^2 \right)^2 \\
&= \bar{\Phi}_{2,2} - \bar{\Phi}_2^2 + 2[\bar{\Phi}_2^2 - \bar{\Phi}_4] \sum_{i=1}^I p_i^2 + 4[2\bar{\Phi}_4 + \bar{\Phi}_2 - 2\bar{\Phi}_3 - \bar{\Phi}_{2,2}] \sum_{i=1}^I p_i^3 \\
&\quad + [3\bar{\Phi}_{2,2} + 8\bar{\Phi}_3 - 6\bar{\Phi}_4 - 4\bar{\Phi}_2 - \bar{\Phi}_2^2] \left( \sum_{i=1}^I p_i^2 \right)^2.
\end{aligned}$$

Applying the identity  $\text{Var}[(1 - \sum_{i=1}^I \widehat{p}_i^2)/(1 - \bar{\Phi}_2)] = \text{Var}[1 - \sum_{i=1}^I \widehat{p}_i^2]/(1 - \bar{\Phi}_2)^2$  gives

eq. 3.15.

□

It is interesting (and convenient) that although the derivation requires the use of all 15  $\Delta$  coefficients, the only coefficients required in the variance formula are  $\bar{\Phi}_2$ ,  $\bar{\Phi}_3$ ,  $\bar{\Phi}_4$ , and  $\bar{\Phi}_{2,2}$ . The 15  $\Delta$  coefficients in Figure 3.7 completely specify the 14  $\Phi$  coefficients in eq. 3.24 (along with the 15th  $\Phi$  coefficient equal to  $\Delta_{15}$ ). Through symmetry of the six  $\Phi$  coefficients involving two individuals, symmetry of the four  $\Phi$  coefficients involving three individuals, and symmetry of the three  $\Phi$  coefficients involving pairs of pairs of individuals, by averaging over sets of individuals, the variance of gene diversity becomes a function of only four average  $\Phi$  coefficients.



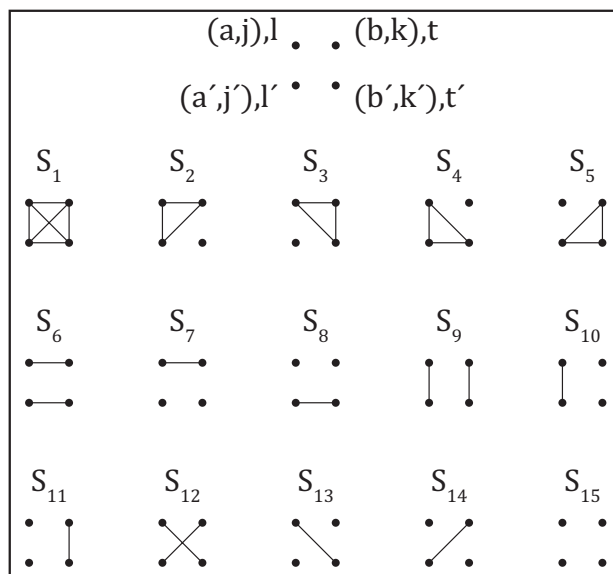


Figure 3.7: Identity states. Two alleles (dots) are identical by descent if and only if there is a line connecting them. This figure is similar to Figure 6.2 of *Jacquard* (1974) and has been reproduced here for convenience.

## CHAPTER IV

# Explaining worldwide patterns of human genetic variation using a coalescent-based serial founder model of migration outward from Africa

### 4.1 Introduction

The nature of the origin and geographic spread of anatomically modern humans has been the focus of much recent interest in anthropology and genetics (*Wolpoff et al.*, 2000; *Stringer*, 2002; *Cavalli-Sforza and Feldman*, 2003; *Klein*, 2008; *Relethford*, 2008; *Weaver and Roseman*, 2008), with considerable effort having been centered on the potential contribution of archaic hominids to the modern human gene pool (*Serre et al.*, 2004; *Garrigan and Hammer*, 2006; *Green et al.*, 2006; *Noonan et al.*, 2006; *Plagnol and Wall*, 2006; *Herrera et al.*, 2009). Within this context, population-genetic studies have examined a variety of aspects of worldwide human variation, identifying several striking geographical patterns in statistics that describe human genetic diversity (Fig. 4.1). First, the level of genetic variation, as measured by heterozygosity, exhibits a linear decline as a function of geographic distance from Africa (*Prugnolle et al.*, 2005; *Ramachandran et al.*, 2005; *Li et al.*, 2008). Second, LD increases linearly as a function of geographic distance from Africa (*Jakobsson et al.*, 2008). Third, the ancestral allele frequency spectrum “flattens” with increasing geographic distance

from Africa, indicating that derived alleles tend to be more frequent in populations at a greater distance away from Africa (*Li et al.*, 2008).

These three patterns point to an important role for Africa in the history of human genetic variation. Thus, many models involving migrations outward from Africa have been proposed for providing simulation-based explanations of geographical patterns in human genetic data. This collection of models includes coalescent-based migration models that proceed retrospectively in time and that are easily simulated, but that involve relatively few populations, each of which typically represents a large geographic region (*Plagnol and Wall*, 2006; *Takahata et al.*, 2001; *Schaffner et al.*, 2005; *Fagundes et al.*, 2007). It also includes models that permit complex phenomena and multiple populations per continent through a prospective approach, but that are often limited in terms of computation time and applicability to statistical inference (*Ramachandran et al.*, 2005; *Eswaran*, 2002; *Liu et al.*, 2006; *Deshpande et al.*, 2009).

One model that has performed well in explaining the decline of heterozygosity with increasing distance from Africa is a model of serial founder events beginning from an African origin (*Ramachandran et al.*, 2005; *Deshpande et al.*, 2009; *Hunley et al.*, 2009). In this model, starting with a single source population, a new population is formed from a subset of the individuals in the founding population. The new population experiences a bottleneck, in that it is founded by a small group. It grows to a larger size, after which a subset of the population becomes the founding group for a third population. The founding process is then iterated (Fig. 7.1A). Simulations of the serial founder model in a prospective framework produce a decrease of heterozygosity in each subsequent group, so that heterozygosity appears to decline linearly with the number of colonization steps from the source population. Intuitively, when a new colony is founded, it carries only a subset of the diversity from the previous colony, and therefore, a heterozygosity decrease occurs. Thus, it has been shown that if the source is placed in Africa, then the prediction of serial founder models matches

the observed pattern of heterozygosity (*Ramachandran et al.*, 2005; *Deshpande et al.*, 2009; *Hunley et al.*, 2009). It has also been suggested that the serial founder model can explain worldwide patterns in LD and the ancestral allele frequency spectrum (*Li et al.*, 2008; *Jakobsson et al.*, 2008), although these claims have not yet been verified in simulations of the model.

Here, we develop a retrospective coalescent approach that enables a generalization of the serial founder model. As few models of human range expansions have considered linked loci (*Hellenthal et al.*, 2008), our approach makes it possible to examine a broader variety of patterns than have been studied in most out-of-Africa models. Rather than performing formal statistical inference under our new general model, we aim to determine whether the model qualitatively accords with worldwide trends in human genetic variation. We indeed find that the new model provides explanations not only of geographic patterns of heterozygosity, but also of patterns of LD and the ancestral allele frequency spectrum. The model accommodates migration between neighboring colonies and admixture between modern and archaic populations; through the introduction of two additional models, an archaic persistence model and an instantaneous divergence model, we discuss the extent to which these phenomena are compatible with worldwide variation patterns.

## 4.2 Results

### 4.2.1 Overview of models

Our serial founder model is a special case of a more general model (Fig. 7.1A). In our serial founder model, each of  $K$  populations, numbered with increasing distance from a founding group (population 1), has present population size  $N$  diploid individuals. The divergence time of populations 1 and 2,  $t_D$  generations ago, represents the time of formation of a second modern human population. The

model proceeds as a series of founding events in which a group of individuals migrates from the most recently founded colony to form a new colony. Because each founding group is small compared to its source, when a new colony  $k$  is founded, it undergoes a bottleneck of size  $N_b < N$  individuals lasting  $L_b$  generations. It then immediately expands to size  $N$ . After  $L$  generations, a group of individuals migrates from colony  $k$  to found population  $k + 1$ . Population divergence times are arranged such that founding events occur at intervals of  $t_D/(K - 1)$  generations. Thus,  $L + L_b = t_D/(K - 1)$ .

To include migration between neighboring populations, as in *Deshpande et al.* (2009), we add symmetric migration between neighbors at rate  $M = 4Nm$ , where  $m$  is the per-generation fraction of a population consisting of new migrants. Backward in time, population  $k$  sends migrants to populations  $k - 1$  and  $k + 1$ , each with rate  $M$ , and populations  $k - 1$  and  $k + 1$  send migrants to population  $k$ , each with rate  $M$ . Migration only involves populations that have already been founded, so that during the stage when population  $k$  is the newest population, it only experiences migration with one colony instead of two. Populations 1 and  $K$  never experience migration with two populations during the entire time of their existence.

In our general model, an archaic population diverges at time  $t_D^A$  generations ( $t_D^A > t_D$ ) to form a population of constant diploid size  $N_A$  individuals. After a period of isolation, the archaic population admixes with a single modern population  $k^*$  at rate  $\gamma$  so that at time  $t_{Admix}$  generations, the probability that a lineage from population  $k^*$  enters the archaic population is  $\gamma$  going back in time. Admixture occurs  $L/2$  generations after population  $k^*$  expands to size  $N$ .

### 4.2.2 Simulations

Sets of  $K$  populations under the basic serial founder model, the migration model, and the archaic admixture model were simulated using the coalescent simulator MS

(Hudson, 2002). For each model, parameter values that produced representative phenomena were selected within plausible ranges. Each population sample consisted of  $n$  100 kb chromosomes, randomly paired to create  $n/2$  diploid individuals. We used a 25-year generation time, a sequence length  $S_L = 10^5$  bases, a per-base mutation rate  $\mu_s = 2.5 \times 10^{-9}$ , a per-base recombination rate  $r_s = 2.50025 \times 10^{-9}$ , and a population size  $N = 10,000$ . These values produce a population mutation rate  $\theta = 4N\mu = 10$ , where  $\mu = S_L\mu_s$ , and a population recombination rate  $\rho = 4Nr = 10$ , where  $r = (S_L - 1)r_s$ . For each model, we simulated 5,000 datasets of  $K = 100$  populations, each with a sample of size  $n = 50$ . Heterozygosity, LD, and the slope of the ancestral allele frequency spectrum were calculated for each dataset, and weighted averages were taken over replicate simulations to produce final values of the statistics (see Methods).

### 4.2.3 Basic model

We first examined a basic serial founder model with  $t_D = 2,079$  (51.975 kya), with no migration between neighbors and no archaic admixture. We used a bottleneck size of  $N_b = 250$ , a bottleneck length of  $L_b = 2$ , and a time length between a population expansion and the founding of a new colony of  $L = 19$ . These choices were largely designed to mimic values used in past simulations (Ramachandran *et al.*, 2005; Liu *et al.*, 2006).

Under this model, Fig. 4.4A displays a linear decline in heterozygosity with increasing colony number. The heterozygosities are small, because they are means over all segregating sites, and many sites are monomorphic within a given population sample. However, the qualitative pattern matches that seen in data (Fig. 4.1A) and in forward simulations (Ramachandran *et al.*, 2005; Liu *et al.*, 2006; Deshpande *et al.*, 2009). The LD decay with increasing distance along a chromosome has a pattern in which populations far from the parental colony have the highest LD (Fig. 4.4B).

Focusing on LD at 10 kb, a linear LD increase with increasing colony number is apparent (Fig. 4.4C). We can also observe a flattening of the ancestral allele frequency spectrum with increasing colony number, as reflected in a decline in the regression slope of this spectrum on ancestral allele frequency (Fig. 4.4D). Thus, the serial founder model produces LD patterns and ancestral allele frequency spectra that match those observed in data (Fig. 4.1).

## Migration

We next added symmetric migration to our basic model, with rate  $M = 4Nm$  between neighboring colonies, holding all other parameters the same. To represent higher and lower migration rates, we considered  $M = 40$  and  $M = 1$ .

Figs. 4.5 and 4.6 show, as was observed in the case of no migration, that as colony number increases there is a decline in heterozygosity, an increase in LD, and a decline in the slope of the ancestral allele frequency spectrum. These results suggest that the migration parameter does not have a major effect on the qualitative patterns, and that inclusion of migration only slightly alters the patterns observed with bottlenecks alone.

One possible reason for a stronger influence of bottlenecks compared to migration is that in the time scale of the model—with recent bottlenecks followed by short periods of migration—migration between neighbors might not move ancestral lineages very far from their original locations. Instead of being located during the bottleneck at the founding of the population from which two lineages are sampled, the common ancestor for a pair of lineages might be located during a bottleneck only a few steps earlier in the serial expansion. Although migration increases the coalescence time of a random pair of lineages from a population relative to the corresponding time in the model without migration, the extra time to coalescence caused by migration might typically be quite small.

A heterozygosity peak visible in the first few populations (Fig. 4.5A) is likely to result from edge effects. Central populations receive more diverse migrants than edge populations, because migration brings in distant lineages from both sides. In a similar model in which a linearly-arrayed population has persisted for a long time (*Wilkins and Wakeley, 2002*), diversity is greatest at the center, because the ancestors of lineages in the center typically wander over a greater range before coalescing. In our model, the fact that central populations originate more recently than populations near the source lessens this effect, because lineages from groups near the source have been through fewer bottlenecks than lineages in the middle and might therefore have deeper coalescence times. As a result of the competing effects of bottlenecks and migration, the highest diversity occurs in populations located between the source and the center.

#### 4.2.4 Archaic admixture

We next added archaic admixture to the basic model, using  $N_A = 1,000$  for the size of the archaic population,  $t_D^A = 16,000$  (400 kya) for the splitting time of the modern and archaic populations, and  $t_{Admix} = 1584.5$  (39.6125 kya) for the time of admixture. Archaic admixture occurred in modern population  $k^* = 25$ , with fraction  $\gamma$  of this population instantaneously taken from the archaic population at time  $t_{Admix}$ . The parameter choices reflect a model of admixture of a European modern population and Neanderthals, with admixture occurring halfway between the end of one founding event and the beginning of the next founding event. The time and extent of admixture were chosen to have similar values to those employed in previous models (*Noonan et al., 2006; Plagnol and Wall, 2006*).

In Fig. 4.7, admixture with  $\gamma = 0.05$  leads to patterns in heterozygosity, LD, and the slope of the ancestral allele frequency spectrum similar to those observed in the basic serial founder model. However, archaic admixture causes an increase



in heterozygosity and a decrease in LD that occur at population 25 and that are carried into subsequent populations. Heterozygosity increases at population 25 because admixture brings in archaic lineages distinct from the modern lineages in that population. The LD decrease at population 25 results from the way in which bottlenecks and admixture interact in our model. Bottlenecks increase the genetic drift experienced by a population, and genetic drift increases short-range LD (*Slatkin, 2008*). Admixture can also inflate LD, particularly long-range LD (*Plagnol and Wall, 2006*), as allelic correlations at distant loci can arise via the separate sets of haplotypes contained in the distinct groups ancestral to a population. If the effect of bottlenecks in producing LD is stronger than the effect of admixture, then including admixture in the model causes short-range LD to be smaller in the first population that experiences the admixture than in the previous population in the series.

Increasing the admixture fraction to  $\gamma = 0.1$  further increases heterozygosity and decreases short-range LD at population 25 (Fig. 4.8). The slope of the ancestral allele frequency spectrum increases at population 25, causing a discontinuity that was less visible at population 25 in the case of  $\gamma = 0.05$ . This jump occurs because population 25 receives an influx of ancestral haplotypes from the archaic population, thereby increasing both the frequencies of ancestral alleles and the slope of the ancestral allele frequency spectrum. With larger  $\gamma$ , the amount of LD in population 25 at long physical distances is larger (Fig. 4.9), compatible with the greater effect of admixture on long-range LD compared to that of bottlenecks.

#### **4.2.5 Archaic persistence model**

To examine if an admixture model with substantially greater contributions from archaic populations can explain patterns in heterozygosity, LD, and the ancestral allele frequency spectrum, we developed an “archaic persistence model” to reflect a scenario in which modern humans originate from one archaic population and then

expand into a collection of preexisting archaic populations.

In this model (Fig. 7.1B),  $K$  populations, each with size  $N$  diploid individuals, diverge  $t_D$  generations ago, and each experiences subsequent migration with its immediate neighbors at rate  $M$  in each direction. At  $t_1$  generations in the past, looking forward in time, population 1 sends a large wave of migrants to population 2 over a series of  $L_w$  generations. Backward in time, this wave corresponds to a change from  $M$  to  $W$  in the backward migration rate from population 2 to population 1, so that a fraction  $W/(4N)$  of population 2 is drawn from population 1 in each of the  $L_w$  generations. For each  $k$ , population  $k$  sends a similar (forward) wave to population  $k + 1$  at  $t_k$  generations in the past.

Using MS, we simulated 5,000 datasets with  $K = 100$ ,  $N = 1,000$ ,  $n = 50$ ,  $M = 0.1$ ,  $t_D = 40,000$ , and  $t_k = 2,079 - 21(k - 1)$ . The value for  $t_k$  matches the founding time for population  $k + 1$  in our basic serial founder simulations. Each wave lasts  $L_w = 2$  generations, matching our serial founder bottleneck length, and sends 250 migrants per generation ( $W = 1,000$ ). The parameter choices reflect a scenario in which archaic humans spread  $\sim 1$  million years ago and modern humans arose via admixture of archaic populations with descendants of a recent expansion out of Africa.

In contrast to the basic serial founder model, the archaic persistence model produces patterns opposite to those observed in human data (Fig. 4.10). Heterozygosity *increases*, LD *decreases*, and the ancestral allele frequency spectrum slope *increases* with increasing colony number. These results can be understood from the fact that in the long time since the initial divergence, the  $K$  archaic populations have enough time to develop distinctive localized variants. As the migration wave travels through them, it accumulates diversity, gathering new variants from each population through which it passes. Thus, heterozygosity increases with increasing colony number in the same way that it increases in the archaic admixture model at the

population in which admixture occurs. The difference between models lies in the fact that in the archaic persistence model, archaic admixture occurs in *every* population, so that heterozygosity increases at each step rather than at a single location. This occurrence of archaic admixture at each step also explains the decrease in LD and increase in the slope of the ancestral allele frequency spectrum that occur at each step. Deviations in the initial and final colonies from the general patterns are likely due to edge effects; the linear arrangement of populations prevents edge populations from accumulating the same level of diversity prior to the migration wave as that accumulated in central populations.

#### 4.2.6 Instantaneous divergence model

A key feature of the serial founder model is that compared with an earlier colony in the series, a subsequent colony has fewer ancestors over a longer period of time in its history. Thus, to assess if a decline in effective size can explain patterns in heterozygosity, LD, and the ancestral allele frequency spectrum, we devised an instantaneous divergence model that captures this effective size reduction without explicitly modeling bottlenecks. This model, which itself is implausible as a description of human migrations, can help illuminate the features of the serial founder model that allow it to explain observed patterns.

Our instantaneous divergence model (Fig. 7.1C) has  $K$  populations each with a different constant size, chosen so that the total elapsed coalescent time for population  $k$  since the divergence equals that elapsed for population  $k$  since initial divergence in the basic serial founder model. The cumulative coalescent intensity from the present back  $t_D$  generations of a variable-sized population whose size was  $N(s)$  at  $s$  generations in the past is  $\int_0^{t_D} 1/N(s) ds$  (*Sjodin et al.*, 2005). The corresponding cumulative intensity from the present back  $t_D$  generations of a population of constant size  $N_k$  is  $\int_0^{t_D} 1/N_k ds = t_D/N_k$ . Setting the intensities of

the variable-sized and constant-sized populations equal, a population of constant size  $N_k = t_D / \int_0^{t_D} 1/N(s) ds$  experiences the same length in coalescent time as a variable-sized population with size function  $N(s)$ . In our basic serial founder model, bottlenecks last  $L_b$  generations, and during a bottleneck, a population has constant size  $N_b$ . Therefore, for each bottleneck, the elapsed coalescent time is  $L_b/N_b$ . Because population  $k$  experiences  $k - 1$  bottlenecks, the cumulative coalescent time elapsed during bottlenecks is  $(k - 1)L_b/N_b$ . Similarly, the cumulative coalescent time outside of bottlenecks is  $[t_D - (k - 1)L_b]/N$ . Thus, we assign population  $k$  in the instantaneous divergence model size

$$N_k = \frac{t_D}{[t_D - (k - 1)L_b]/N + (k - 1)L_b/N_b}, \quad (4.1)$$

where  $N$ ,  $N_b$ , and  $L_b$  are as in the serial founder model.

For this model, using MS, we simulated 5,000 datasets with  $K = 100$  and  $n = 50$ . The divergence between the  $K$  populations occurred at  $t_D = 2,079$  (51.975 kya). The ancestral population had size  $N = 10,000$  diploid individuals, and for each  $k$ , population  $k$  had size  $N_k$  (eq. 4.1).

Comparing Figs. 4.4 and 4.11, the serial founder and instantaneous divergence models display nearly identical patterns and ranges of values for heterozygosity, LD, and the slope of the ancestral allele frequency spectrum. This concordance has the interpretation that the worldwide genetic patterns observed in human populations can be explained by a decrease in the cumulative number of ancestors of a population—that is, an increase in genetic drift and in total elapsed coalescent time—with increasing distance from the source. Thus, the utility of this instantaneous divergence model is that it provides an explanation for the success of the more realistic serial founder model in describing worldwide patterns of variation.

### 4.3 Discussion

In this paper we have developed a general coalescent-based serial founder model that incorporates linked loci, providing a versatile tool for generating and testing hypotheses about features of human population-genetic data. Using several cases of the model, we examined heterozygosity, LD, and the ancestral allele frequency spectrum, mimicking computations performed in past data analyses. If the source population is placed in Africa, then the serial founder model explains three patterns observed in data: a decrease in heterozygosity, increase in LD, and decrease in the slope of the ancestral allele frequency spectrum with increasing distance from Africa. Our use of an instantaneous divergence model suggests that the patterns observed in the data—and the success of the serial founder model—are due to an increase in genetic drift and a corresponding increase in elapsed coalescence time with increasing distance from Africa. Unlike the serial founder model, an archaic persistence model, in which a migration wave of modern humans into preexisting archaic populations has the effect of increasing the diversity of the ancestors for populations at a greater distance from Africa, does not produce increasing drift with increasing distance from Africa, and does not explain observed patterns.

We considered a variant of the basic serial founder model that included migration between neighboring populations, finding that migration did not have a large impact on the decrease in heterozygosity, increase in LD, and decrease in the slope of the ancestral allele frequency spectrum that were observed from the basic model. However, an increased migration rate caused a peak in the level of heterozygosity to appear in populations near the founding colony rather than in the founding colony itself. This result suggests that when using patterns of diversity to pinpoint the origin of an expansion in a serial founder framework (*Ramachandran et al.*, 2005; *Tishkoff et al.*, 2009), the site of origin might reside in a neighboring population to the highest-diversity population, rather than in the highest-diversity population itself.

We examined the effect of limited archaic admixture on the serial founder model and found that it increased heterozygosity, decreased short-range LD, and increased the slope of the ancestral allele frequency spectrum, starting at the admixed population. The LD decrease contrasts with the results of *Plagnol and Wall* (2006), who found that archaic admixture was needed to inflate the level of LD to match that observed in Europeans at intermediate physical distances. Note, however, that whereas we considered the standard  $r^2$  LD statistic using all loci with minor allele frequency  $\geq 0.05$ , Plagnol & Wall focused on a statistic specifically designed to be sensitive to archaic admixture, and applied it to a different restricted class of SNPs. Differences in results might also have arisen from modeling differences, such as in the values used for the time of admixture, population size, and bottleneck size. As the effect we observed for archaic admixture on LD was relatively weak, our LD summaries might not be informative enough to empirically detect archaic admixture.

More generally, the relative similarity of predictions of the basic serial founder, migration, archaic admixture, and instantaneous divergence models suggests that it is difficult to distinguish these models solely using the summary statistics that we have considered. Thus, although a serial founder model is supported by the analysis, many alternatives cannot be excluded. However, the archaic persistence model, whose predictions disagree with the patterns in the data, is not in this collection. Because a migration wave of modern humans in this model carries an increasing diversity of archaic contributions into subsequent populations, this model does not possess the essential feature that permits other models to explain observed patterns, namely an increase in genetic drift with distance from the source. Use of unequal sizes for persisting archaic populations, however, might have produced patterns with greater similarity to those produced by a serial founder model (*Weaver and Roseman*, 2008; *Relethford*, 1998). If archaic population sizes had instead decreased with increasing colony number, via an *archaic* serial founder process, then the production by archaic

persistence of patterns opposite to those in the data might have been offset by an archaic serial founder increase in genetic drift with increasing distance from a founding archaic colony.

Such a scenario is likely implausible, as an archaic serial founder process is not expected to have similar behavior to the modern analog that we have analyzed. Let  $t_D$  grow in a serial founder model with migration while holding  $L_b$  and  $L$  constant. We expect lineages from each population to find common ancestors before any population split or bottleneck is reached. The model would then approximate the finite linear population model of *Wilkins and Wakeley (2002)*, for which predictions differ substantially from those of the serial founder model with migration. The finite linear model predicts that the center of the range receives diverse migrants and therefore has the highest diversity, whereas in the serial founder model with migration, populations near the founding colony have the highest diversity. Thus, although some flexibility exists in the parameters that allow the serial founder model to match observed data, and although we have only explored a small part of the parameter space, consideration of archaic persistence suggests that the model cannot be made too different and still explain the patterns in the data.

## 4.4 Materials and Methods

### 4.4.1 Heterozygosity

For the heterozygosity of a population in one simulation we used the standard unbiased estimator (*Nei and Roychoudhury, 1974*), averaged over all loci polymorphic in the set of  $K$  populations. We then calculated a weighted average across simulations of the mean heterozygosity across loci. We used the proportion of segregating sites in a simulated dataset (segregating in the whole simulated set of  $K$  populations) relative to the total number of segregating sites from all 5,000 simulated datasets as weights.

#### 4.4.2 Linkage disequilibrium

For each population we calculated the  $r^2$  LD statistic (*Slatkin, 2008*) between all distinct pairs of sites with minor allele frequency  $\geq 5\%$  in that population. For each simulation, using the distance between the two sites in a pair, we placed  $r^2$  values into 1 kb bins representing physical distances in the ranges  $[0 \text{ kb}, 1 \text{ kb})$ ,  $\dots$ ,  $[99 \text{ kb}, 100 \text{ kb})$ . In each population we obtained an average of all  $r^2$  values in each bin. We then computed an average  $r^2$  across simulations, weighting the results of a simulation by the proportion of  $r^2$  comparisons performed for that simulated dataset in that population relative to the total number of  $r^2$  comparisons in that population from all 5,000 simulations. The  $[9 \text{ kb}, 10 \text{ kb})$  bin was used to indicate LD at 10 kb.

#### 4.4.3 Ancestral allele frequency spectrum

In computing the slope of the ancestral allele frequency spectrum as a function of allele frequency, we modified the method of *Li et al. (2008)* using resampling to evade a discreteness effect in which some frequency bins contain more markers than others due to more discrete frequencies being assigned to those specific bins. We used ancestral and derived allele assignments from *Li et al. (2008)* for 407,001 autosomal markers in the data of *Jakobsson et al. (2008)*. For each population, for each locus, we computed ancestral allele frequency using 1000 random draws from the empirical allele frequency distribution. Loci were binned by ancestral frequency into 20 bins, representing  $[0/20, 1/20)$ ,  $[1/20, 2/20)$ ,  $\dots$ ,  $[19/20, 20/20]$ . For each population, bin counts were normalized by the total number of loci. The slope of the linear regression of the normalized frequency spectrum on ancestral allele frequency was then computed using bins centered at  $9/40$  to  $31/40$  (similarly to the use by *Li et al. (2008)* of frequencies  $1/5$  to  $4/5$ ).

For the corresponding computation from our simulations, we used  $n + 1$  bins, so that if a locus had the ancestral allele occurring  $i$  times, then the count for bin  $i/n$  was



incremented. For a given simulation, for each population, counts were normalized by the number of segregating sites in that simulation. For each population, the slope of the linear regression of normalized frequency spectrum on ancestral allele frequency was computed using bins 10/50 through 40/50. In each population, we calculated an average slope cross simulations, weighting the value for a simulation by the proportion of segregating sites observed in that simulated dataset relative to the total number of segregating sites from all simulated datasets.

## 4.5 Acknowledgments

We thank Jun Li for ancestral allele frequency details from *Li et al.* (2008) and three reviewers for helpful comments. Support was provided by National Institute of Health grants T32 GM070449 and R01 GM081441, and by grants from the Burroughs Wellcome Fund, the Alfred P. Sloan Foundation, and the Swedish Research Council Formas.

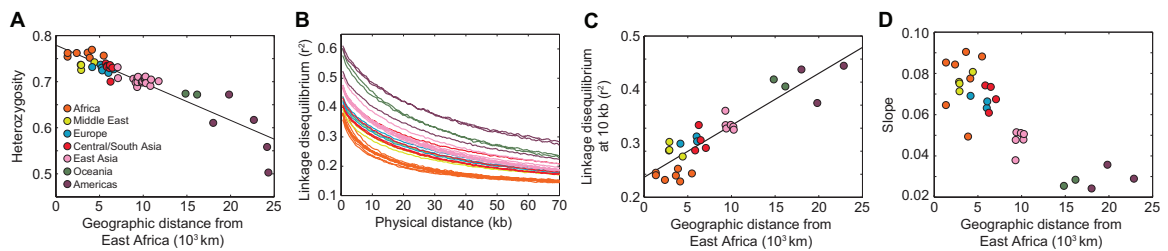


Figure 4.1: Patterns of heterozygosity, LD, and the ancestral allele frequency spectrum observed in human population-genetic data. (A) Heterozygosity as a function of distance from East Africa (redrawn from *Ramachandran et al. (2005)* as in Fig. 7C of *DeGiorgio and Rosenberg (2009)*). (B) LD measured by  $r^2$  as a function of physical distance in kb (redrawn from Fig. S4 of *Jakobsson et al. (2008)*). (C) LD at 10 kb measured by  $r^2$  as a function of distance from East Africa (based on data in Fig. S4 of *Jakobsson et al. (2008)*). (D) Slope of the ancestral allele frequency spectrum in the range of 20% to 80% ancestral allele frequency as a function of distance from East Africa (modified from Fig. 4B of *Li et al. (2008)* using a resampling technique and the allele frequencies in Fig. 4.2).

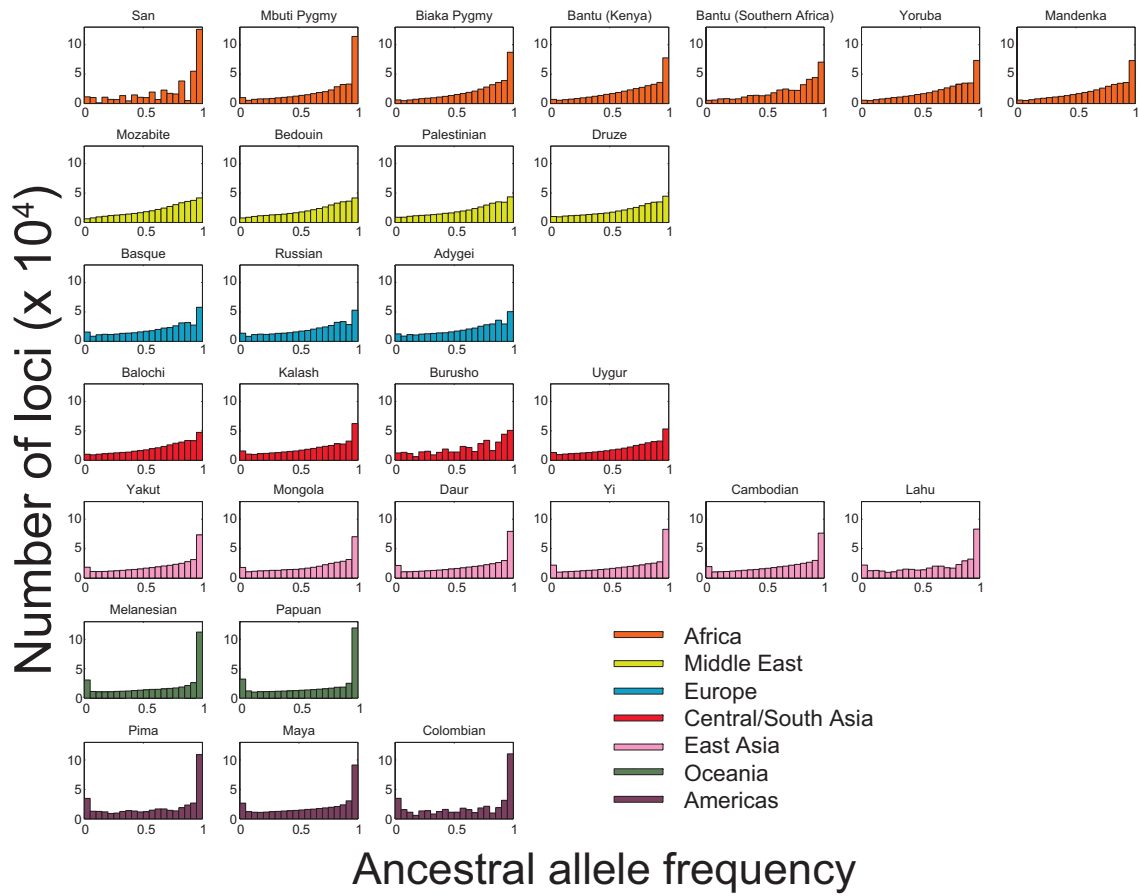


Figure 4.2: Ancestral allele frequency spectra calculated using a resampling technique applied to the data of *Jakobsson et al.* (2008).

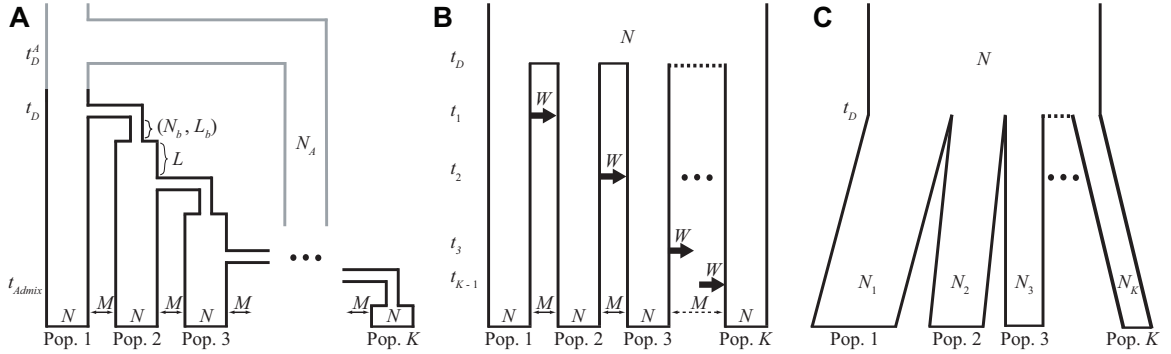


Figure 4.3: Models. (A) Serial founder model with population size  $N$  diploid individuals in each of  $K$  populations, time  $t_D$  of the first divergence from the founding population, bottleneck size  $N_b$ , bottleneck length  $L_b$ , time interval  $L$  between successive bottlenecks, and symmetric migration rate  $M$  between neighboring populations. An extension of the model that allows admixture with archaic humans has additional parameters for the population size for archaic humans ( $N_A$ ), divergence time between modern and archaic humans ( $t_D^A$ ), and time of admixture between a specific modern population and the archaic population ( $t_{Admix}$ ). (B) Archaic persistence model with population size  $N$  diploid individuals in each of  $K$  populations, time  $t_D$  of the divergence of archaic populations, symmetric migration rate  $M$  between neighboring populations, and migration rate  $W$  for the migration wave from population  $k$  to population  $k + 1$  at time  $t_k$ . (C) Instantaneous divergence model with population sizes  $N_k$  for populations  $k = 1, 2, \dots, K$ , population size  $N$  for the ancestral population, and divergence time  $t_D$ .

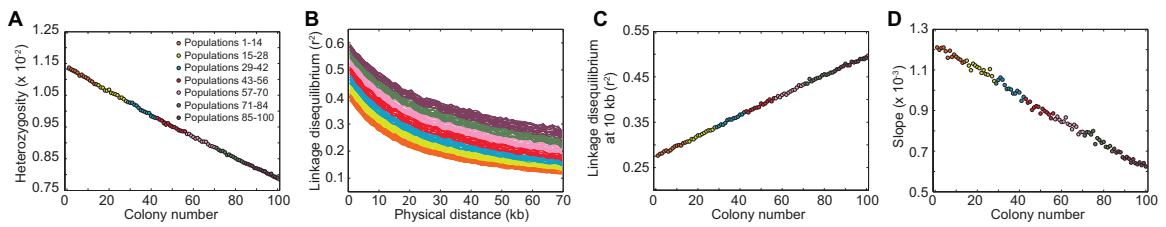


Figure 4.4: Patterns of heterozygosity, LD, and the ancestral allele frequency spectrum in simulations of the basic serial founder model. (A) Heterozygosity as a function of colony number. (B) LD measured by  $r^2$  as a function of physical distance in kb. (C) LD at 10 kb measured by  $r^2$  as a function of colony number. (D) Slope of the ancestral allele frequency spectrum in the range of 20% to 80% ancestral allele frequency as a function of colony number.

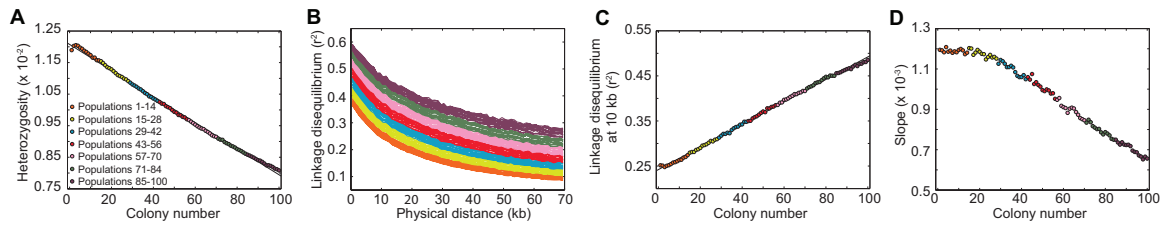


Figure 4.5: Patterns of heterozygosity, LD, and the ancestral allele frequency spectrum in simulations of the serial founder model with symmetric migration at rate  $M = 40$  between neighboring populations. All other parameters are the same as in Fig 4.4.

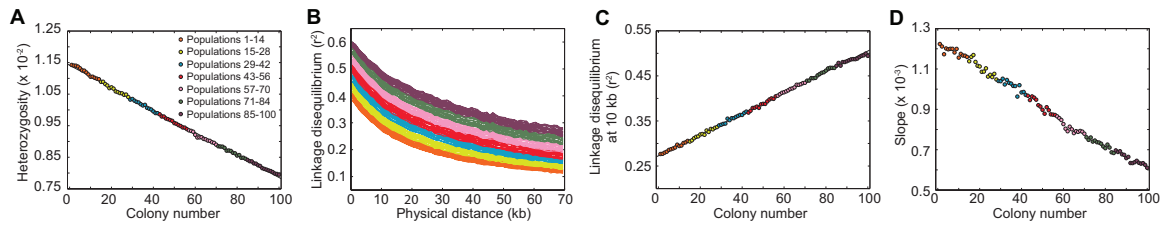


Figure 4.6: Patterns of heterozygosity, LD, and the ancestral allele frequency spectrum in simulations of the serial founder model with symmetric migration at rate  $M = 1$  between neighboring populations. All other parameters are the same as in Fig 4.4.

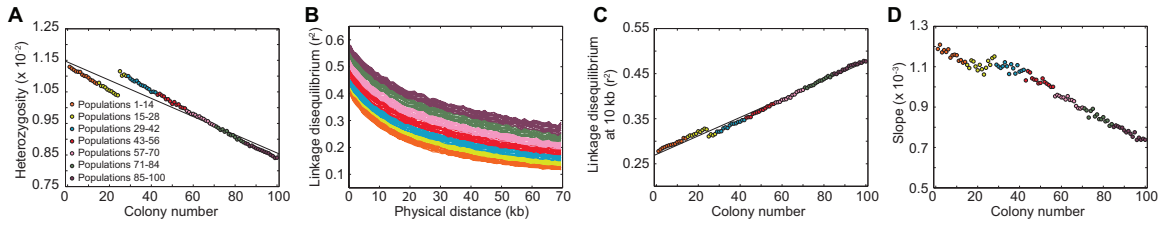


Figure 4.7: Patterns of heterozygosity, LD, and the ancestral allele frequency spectrum in simulations of the serial founder model with archaic admixture. The model incorporates archaic admixture with an admixture fraction  $\gamma = 0.05$  of population 25 deriving from the archaic population. All other parameters are the same as in Fig. 4.4.



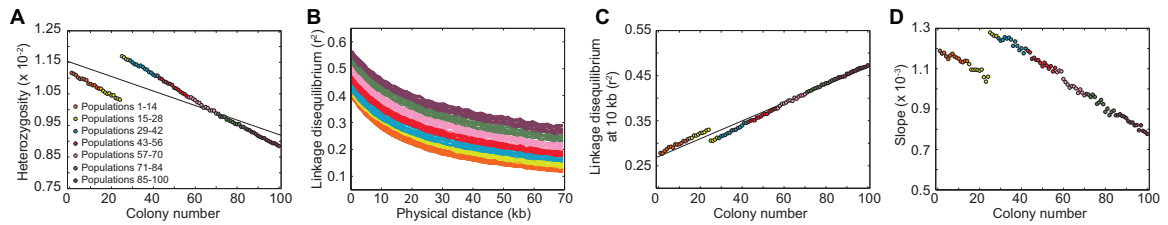


Figure 4.8: Patterns of heterozygosity, LD, and the ancestral allele frequency spectrum in simulations of the serial founder model with archaic admixture. The model incorporates archaic admixture with an admixture fraction  $\gamma = 0.1$  of population 25 deriving from the archaic population. All other parameters are the same as in Fig. 4.4.

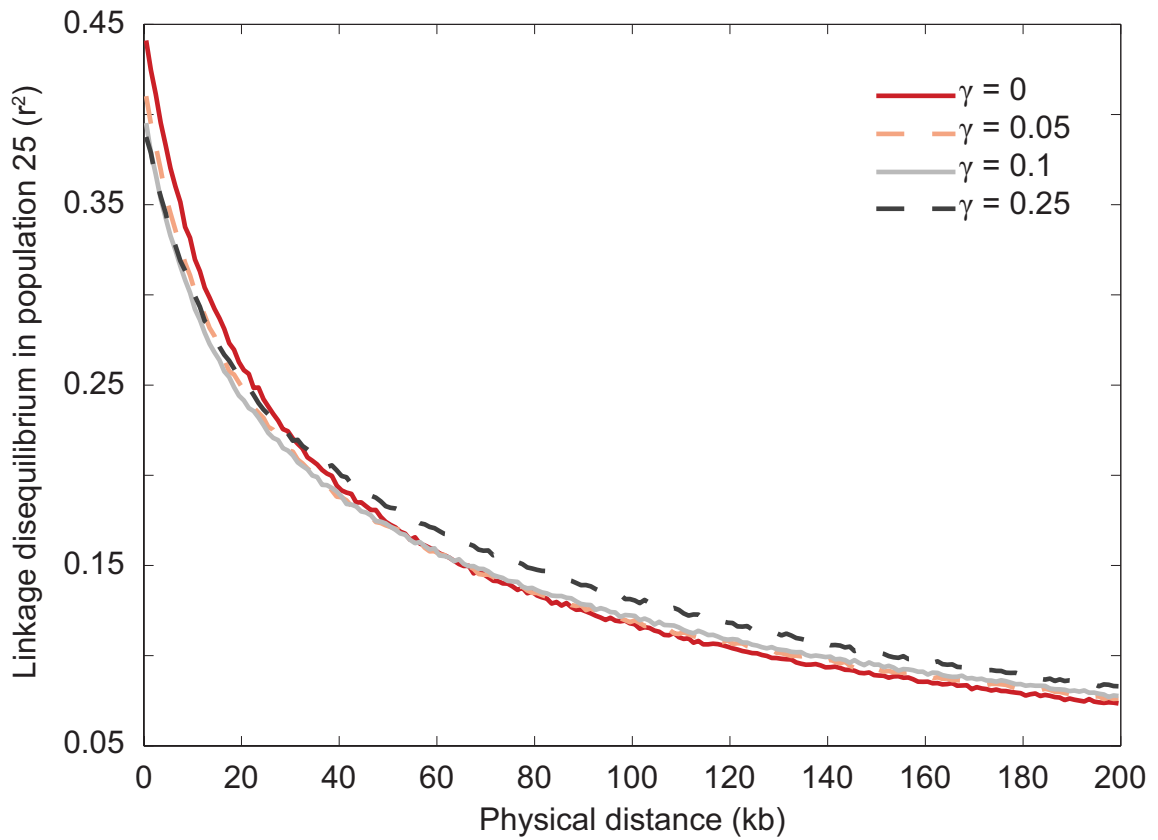


Figure 4.9: LD ( $r^2$ ) as a function of physical distance for population 25 in the serial founder model with archaic admixture at rate  $\gamma$ . The simulation proceeded in the saw as in Fig. 4.7, except that longer regions were simulated (1 Mb).

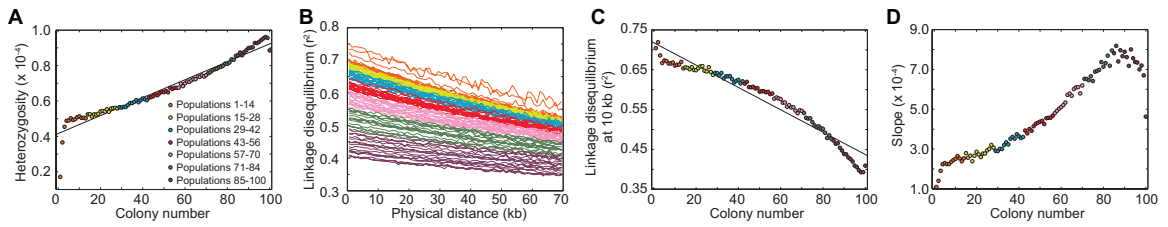


Figure 4.10: Patterns of heterozygosity, LD, and the ancestral allele frequency spectrum in simulations of the archaic persistence model. (A) Heterozygosity as a function of colony number. (B) LD measured by  $r^2$  as a function of physical distance in kb. (C) LD at 10 kb measured by  $r^2$  as a function of colony number. (D) Slope of the ancestral allele frequency spectrum in the range of 20% to 80% ancestral allele frequency as a function of colony number.

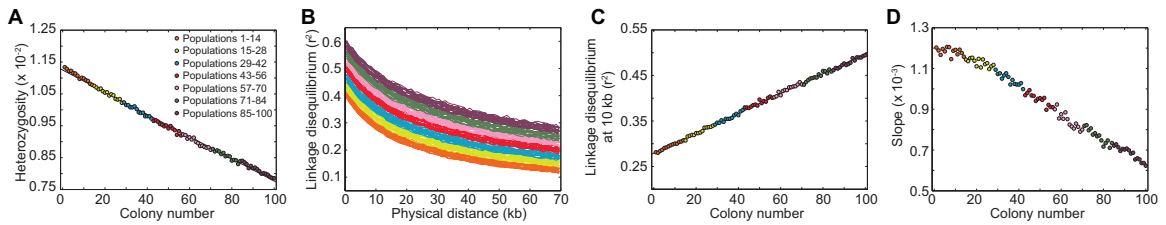


Figure 4.11: Patterns of heterozygosity, LD, and the ancestral allele frequency spectrum in simulations of the instantaneous divergence model. (A) Heterozygosity as a function of colony number. (B) LD measured by  $r^2$  as a function of physical distance in kb. (C) LD at 10 kb measured by  $r^2$  as a function of colony number. (D) Slope of the ancestral allele frequency spectrum in the range of 20% to 80% ancestral allele frequency as a function of colony number.

## CHAPTER V

# Coalescence-time distributions in a serial founder model of human evolutionary history

### 5.1 Introduction

Equilibrium population structure models in population genetics, which assume that the rules specifying the evolution of alleles within and among populations do not change with time, have achieved much success in describing genetic variation. Although equilibrium models are convenient for obtaining analytical results that can be used to test hypotheses and predict patterns of genetic variation, non-equilibrium models often provide more realistic representations of patterns that occur in real populations. Non-equilibrium models assume that the rules specifying the evolution of alleles within and among populations change as a function of time. In non-equilibrium models, however, except in a small number of cases (e.g., *Takahata et al.*, 1995; *Wakeley*, 1996b,c,a; *Jesus et al.*, 2006; *Efromovich and Kubatko*, 2008), analytical formulas have been somewhat scarce because model complexity can make them difficult to obtain.

Recently, a non-equilibrium structured population model, the “serial founder model,” has been proposed for describing the colonization of the world by modern humans (*Ramachandran et al.*, 2005). The colonization process in this model starts

with a single source population. The source population sends a subset of its individuals to migrate outward and found a new population. This newly founded population has a small size at its founding and subsequently expands to a larger size. After expanding to a larger size, it then sends out migrants to form the next population. The founding process is iterated until  $K$  populations have been founded. The appeal of this model is that using both forward (*Ramachandran et al.*, 2005; *Deshpande et al.*, 2009) and backward (coalescent) simulations (*DeGiorgio et al.*, 2009; *Hunley et al.*, 2009), it has been successful in describing observed patterns of human genetic variation, such as the decline in expected heterozygosity observed with increasing geographic distance from a putative African source location.

In addition to the initial serial founder model of *Ramachandran et al.* (2005), a variety of models that contain the geographical expansions and bottlenecks characteristic of the serial founder model have recently been investigated (*Austerlitz et al.*, 1997; *Le Corre and Kremer*, 1998; *Edmonds et al.*, 2004; *Ray et al.*, 2005; *Klopfstein et al.*, 2006; *Liu et al.*, 2006; *Deshpande et al.*, 2009; *Excoffier and Ray*, 2008; *Hallatschek and Nelson*, 2008; *DeGiorgio et al.*, 2009; *Hunley et al.*, 2009). Among formulations with a one-dimensional geographic structure, some models (e.g., *Austerlitz et al.*, 1997; *Deshpande et al.*, 2009) allow migration after the initial founding of populations and assume that once a population is founded, it logistically grows to its carrying capacity. When carrying capacity is reached (or shortly thereafter), migrants exit the population to found the next population. Other models (e.g., *DeGiorgio et al.*, 2009) do not permit migration after populations are founded and assume that population growth is instantaneous. In these models, after a population is founded, it experiences a small size for some length of time before instantaneously expanding to a larger size. For the former class of models, *Austerlitz et al.* (1997) presented recursion equations to generate the distribution of coalescence times for pairs of lineages sampled either from the same population or from different

populations. These equations can then be used to calculate geographic patterns in summary statistics such as gene diversity and  $F_{ST}$ . For the latter class, *DeGiorgio et al.* (2009) and *Hunley et al.* (2009) approached similar problems using simulations. The relative simplicity of the population growth and migration assumptions in this latter group of models, however, potentially permits explicit analytical formulas, rather than recursions or simulations, to be investigated.

Here, generalizing the coalescent-based version of the serial founder model as formulated by *DeGiorgio et al.* (2009), we provide an analytical distribution of the coalescence time for a pair of lineages at a randomly selected locus, along with corresponding expected coalescence times, expected homozygosity values, and  $F_{ST}$  values. In this non-equilibrium model, we show that the decrease in expected heterozygosity and the corresponding increase in homozygosity with increasing distance from the source population can be predicted analytically. We then provide analytical results for the expected identity for two alleles drawn randomly from a given pair of populations, and we find that the qualitative patterns produced by the formulas closely match those observed from human genetic data and the simulations of *Hunley et al.* (2009). Furthermore, we discuss how our results can be used to obtain analytical formulas for summary statistics for an archaic serial founder model, for the nested-regions model of *Hunley et al.* (2009), and for the instantaneous divergence model of *DeGiorgio et al.* (2009). Our new analytical formulas on within-population gene diversity, between-population gene identity, and pairwise  $F_{ST}$  motivate an analysis of empirical trends in these summary statistics in worldwide human genetic data. Because a serial founder process is largely consistent with worldwide patterns of human genetic variation, the analytical results presented here are useful both for generating and for testing hypotheses about human origins.

## 5.2 Serial founder model

In this section, we begin by formally defining the serial founder model. This model was used in a simulation of *DeGiorgio et al.* (2009), and in this article, we provide a more complete generalization. We then obtain the probability density of coalescence times for two lineages sampled under the model. Utilizing this density, we obtain  $m$ th moments of coalescence times,  $m$ th moments of homozygosities, and  $F_{ST}$  values between pairs of populations.

### 5.2.1 Model

We formulate the serial founder model in a coalescent setting. A diagram of the model appears in Figure 5.1A. Our generic formulation contains a sequence of bottlenecks in which bottleneck sizes, population sizes, bottleneck lengths, and the times for the population founding events are allowed to vary. The model considers  $K$  extant populations, denoted  $E_1, E_2, \dots, E_K$ . For  $i < j$ , the founding of extant population  $E_i$  took place at least as far back in time as that of extant population  $E_j$ . The model has  $2K$  ancestral populations, denoted  $A_0, A_1, \dots, A_{2K-1}$ . For  $i < j$ , the founding of ancestral population  $A_j$  took place at least as far back in time as that of ancestral population  $A_i$ .  $N_i$  denotes the size of ancestral population  $A_i$ ,  $i = 0, 1, \dots, 2K - 1$ . Note that for  $i = 1, 2, \dots, K$ , the size of extant population  $E_i$  is equal to  $N_{2(K-i)}$ , which also is the size of ancestral population  $A_{2(K-i)}$ . Time is measured in generations, and the present has time  $\tau_0 = 0$ .

Forward in time, ancestral population  $A_{2K-1}$  expands to a larger size at time  $\tau_{2K-1}$  to create ancestral population  $A_{2(K-1)}$  (the population directly ancestral to the source population  $E_1$ ). At time  $\tau_{2(K-1)}$ , ancestral population  $A_{2(K-1)}$  splits into extant population  $E_1$  and ancestral population  $A_{2(K-1)-1}$  (a newly founded population during the time in which it experiences a small size prior to expansion). At time  $\tau_{2(K-1)-1}$ , ancestral population  $A_{2(K-1)-1}$  expands to a larger size to form



ancestral population  $A_{2(K-2)}$ . At time  $\tau_{2(K-2)}$ , ancestral population  $A_{2(K-2)}$  splits to form extant population  $E_2$  and ancestral population  $A_{2(K-2)-1}$  (the next founded population during its bottleneck phase). This process is iterated until extant population  $K$  has been founded. In general, at time  $\tau_{2(K-i)}$ ,  $i = 1, 2, \dots, K - 1$ , ancestral population  $A_{2(K-i)}$  splits into extant population  $E_i$  and a newly founded ancestral population  $A_{2(K-i)-1}$ . At time  $\tau_{2(K-i)-1}$ ,  $i = 0, 1, \dots, K - 1$ , ancestral population  $A_{2(K-i)-1}$  expands to a larger size to form ancestral population  $A_{2[K-(i+1)]}$ . Note that by construction, extant population  $E_K$  and ancestral population  $A_0$  are the same population.

We note that several past studies (e.g., *Austerlitz et al.*, 1997; *Ramachandran et al.*, 2005; *Liu et al.*, 2006; *Deshpande et al.*, 2009) utilized formulations of the serial founder model that involved logistic growth of newly founded populations, migration between neighboring populations after their initial founding, or both of these model features. In contrast, for the purpose of obtaining analytical results, our model has a mathematically simpler formulation that involves an instantaneous expansion of a newly founded population to a larger size and that does not permit migration between neighboring populations after founding events.

### 5.2.2 Coalescence times

In this section, we derive the probability density of coalescence times for a pair of lineages sampled under the serial founder model. We begin by deriving the probability density function  $f_{ij}(t)$  for the coalescence time of a pair of lineages, one randomly sampled from extant population  $E_i$  and the other randomly sampled from extant population  $E_j$  (where  $j$  is not necessarily distinct from  $i$ ). This function is defined piecewise over the space of possible coalescence times  $t \in [0, \infty)$ . Using our formula for  $f_{ij}(t)$ , we derive  $m$ th moments of coalescence times, from which we obtain mean pairwise coalescence times. We use the result from coalescent theory that coalescence

times are exponentially distributed with a rate that is inversely proportional to the population size (*Kingman*, 1982; *Hudson*, 1983; *Tajima*, 1983). Also, we use the result that the number of mutations along a genealogical branch is Poisson-distributed, and because we restrict our attention to neutral loci, we separate the mutation process from the genealogical process (*Tavaré*, 1984; *Hudson*, 1990).

Let  $T_{ij}$  be a random variable that denotes the coalescence time for a pair of lineages, one randomly sampled from extant population  $E_i$  and the other randomly sampled from extant population  $E_j$ , with  $i \leq j$ . If  $i < j$ , then the two lineages cannot coalesce until they are in the same ancestral population (*i.e.*, more ancient than  $\tau_{2(K-i)}$ ). Suppose the two lineages exist in the same population during the time interval  $[\tau_h, \tau_{h+1})$ , where  $h \geq 2(K-i)$ . The probability density for coalescence at time  $t \in [\tau_h, \tau_{h+1})$  is the product of the probability that the lineages do not coalesce in the more recent time intervals,  $\exp[-\sum_{\ell=2(K-i)}^{h-1} (\tau_{\ell+1} - \tau_\ell)/N_\ell]$ , and  $(1/N_h)e^{-(t-\tau_h)/N_h}$ , the probability density for a coalescence event at time  $t$  conditional on failure to coalesce more recently than  $\tau_h$ .

If  $i = j$ , then the two lineages can also coalesce in the interval  $[\tau_0, \tau_{2(K-i)})$ . Suppose the two lineages exist in the same population during time interval  $[\tau_0, \tau_{2(K-i)})$ . The probability density for coalescence at time  $t \in [\tau_0, \tau_{2(K-i)})$  in extant population  $E_i$  is  $(1/N_{2(K-i)})e^{-(t-\tau_0)/N_{2(K-i)}}$ . The probability that the lineages do not coalesce in time interval  $[\tau_0, \tau_{2(K-i)})$  is  $e^{-(\tau_{2(K-i)}-\tau_0)/N_{2(K-i)}}$  (we write  $\tau_0$  for notational consistency, but recall  $\tau_0 = 0$ ).

For  $i \leq j$  and  $h \in \{2(K-i), 2(K-i)+1, \dots, 2K-1\}$ , denote the probability that a coalescence event has not occurred by time  $\tau_h$  between a pair of lineages, one from  $E_i$  and one from  $E_j$ , by

$$\Lambda_{ijh} = \exp \left( -\delta_{ij} \frac{\tau_{2(K-i)} - \tau_0}{N_{2(K-i)}} - \sum_{\ell=2(K-i)}^{h-1} \frac{\tau_{\ell+1} - \tau_\ell}{N_\ell} \right),$$

where  $\delta_{ij}$  is the Kronecker delta. We then arrive at the density function for the time to coalescence of a pair of lineages sampled from extant populations  $E_i$  and  $E_j$ ,  $i \leq j$ ,

$$f_{ij}(t) = \begin{cases} \delta_{ij} \frac{e^{-(t-\tau_0)/N_{2(K-i)}}}{N_{2(K-i)}} & , 0 \leq \tau_0 \leq t < \tau_{2(K-i)} \\ \Lambda_{ijh} \frac{e^{-(t-\tau_h)/N_h}}{N_h} & , \tau_h \leq t < \tau_{h+1} \\ & \text{and } h=2(K-i), \dots, 2K-1 \\ 0 & , \text{otherwise,} \end{cases} \quad (5.1)$$

where  $\tau_{2K} = \infty$ . This density for the pairwise coalescence time consists of a collection of shifted exponential distributions, each defined on a different interval.

Equipped with the density in eq. 5.1, we next derive  $m$ th moments for the distribution of coalescence times. We are interested primarily in the mean, but the derivation for arbitrary  $m$  is no more difficult than that for  $m = 1$ .

$$\begin{aligned} \mathbb{E}[T_{ij}^m] &= \int_0^{\infty} t^m f_{ij}(t) dt \\ &= \int_{\tau_0}^{\tau_{2(K-i)}} t^m \delta_{ij} \frac{e^{-(t-\tau_0)/N_{2(K-i)}}}{N_{2(K-i)}} dt + \sum_{h=2(K-i)}^{2K-1} \int_{\tau_h}^{\tau_{h+1}} t^m \Lambda_{ijh} \frac{e^{-(t-\tau_h)/N_h}}{N_h} dt \\ &= \delta_{ij} \frac{e^{\tau_0/N_{2(K-i)}}}{N_{2(K-i)}} \int_{\tau_0}^{\tau_{2(K-i)}} t^m e^{-t/N_{2(K-i)}} dt + \sum_{h=2(K-i)}^{2K-1} \Lambda_{ijh} \frac{e^{\tau_h/N_h}}{N_h} \int_{\tau_h}^{\tau_{h+1}} t^m e^{-t/N_h} dt. \end{aligned}$$

Using the result (*Gradshteyn and Ryzhik*, 2007, p. 106) that

$$\int x^m e^{ax} dx = e^{ax} \sum_{\ell=0}^m \frac{(-1)^\ell \ell! \binom{m}{\ell}}{a^{\ell+1}} x^{m-\ell}, \quad (5.2)$$

we obtain

$$\begin{aligned} \mathbb{E}[T_{ij}^m] &= \sum_{\ell=0}^m \ell! \binom{m}{\ell} \left\{ \delta_{ij} N_{2(K-i)}^\ell \left[ \tau_0^{m-\ell} - \tau_{2(K-i)}^{m-\ell} e^{-(\tau_{2(K-i)}-\tau_0)/N_{2(K-i)}} \right] \right. \\ &\quad \left. + \sum_{h=2(K-i)}^{2K-1} \Lambda_{ijh} N_h^\ell \left[ \tau_h^{m-\ell} - \tau_{h+1}^{m-\ell} e^{-(\tau_{h+1}-\tau_h)/N_h} \right] \right\}. \quad (5.3) \end{aligned}$$

Setting  $m = 1$ , the expected coalescence time is

$$\begin{aligned} \mathbb{E}[T_{ij}] = & \delta_{ij} \left[ \tau_0 + N_{2(K-i)} - \left( \tau_{2(K-i)} + N_{2(K-i)} \right) e^{-(\tau_{2(K-i)} - \tau_0)/N_{2(K-i)}} \right] \\ & + \sum_{h=2(K-i)}^{2K-1} \Lambda_{ijh} \left[ \tau_h + N_h - \left( \tau_{h+1} + N_h \right) e^{-(\tau_{h+1} - \tau_h)/N_h} \right]. \end{aligned} \quad (5.4)$$

Using the density in eq. 5.1, we can investigate how the initial divergence time and the severity of bottlenecks influence the distribution of coalescence times. Figure 5.2B displays density plots for coalescence times in the serial founder model given in Figure 5.2A. Analytical density functions closely match the histograms generated in  $10^7$  coalescent simulations using MS (*Hudson, 2002*), following the simulation method of *DeGiorgio et al. (2009)*. Figure 5.2B shows that multiple modes appear in the distributions of pairwise coalescence times, as a result of the increased rate of coalescence during bottlenecks. Coalescence-time distributions for pairs of lineages from different populations are shifted by the divergence time of the two populations, so that coalescence times for pairs of lineages from distinct populations tend to exceed those of pairs of lineages from the same population.

We can consider the effect of bottleneck size by examining the coalescence time distribution for a pair of lineages in two scenarios that are identical except that one has a smaller bottleneck size. In Figure 5.2B, considering a pair of lineages from population 4, with bottleneck size 1000 diploid individuals, most of the coalescence-time distribution accumulates early because of the strong bottleneck during the time interval  $[\tau_1, \tau_2) = [5000, 10000)$ . Much of the remainder of the distribution accumulates during the next strong bottleneck, in the time interval  $[\tau_3, \tau_4) = [15000, 20000)$ .

Increasing the bottleneck size in Figure 5.2A, from 1000 to 5000, the rate of coalescence of lineages within bottlenecks decreases. Because of this decrease, lineages are more likely to persist farther into the past without coalescing. Thus, Figure 5.2C

shows that decreasing the severity of the bottleneck by increasing the bottleneck population size reduces the probability that the lineages coalesce during the most recent bottleneck. Additionally, a fourth mode of the coalescence-time distribution becomes visible during the bottleneck in the time interval  $[\tau_5, \tau_6) = [25000, 30000)$ .

### 5.2.3 Pairwise homozygosity and heterozygosity

Two commonly used summary statistics are expected homozygosity (gene identity) and expected heterozygosity (gene diversity). Let  $J_{ij}$  be a random variable that denotes the homozygosity for a pair of lineages, one randomly sampled from extant population  $E_i$  and the other randomly sampled from extant population  $E_j$  (where  $j$  is not necessarily distinct from  $i$ ). Further, let  $H_{ij} = 1 - J_{ij}$  be a random variable that denotes the heterozygosity for a pair of lineages, one randomly sampled from  $E_i$  and the other randomly sampled from  $E_j$ . We define homozygosity as the probability that two alleles sampled at a locus are identical by descent (the definition of locus used here is flexible and can range from a single site to a haplotype). Assuming an infinite alleles mutation model and a time interval of length  $T$  generations, if mutations are Poisson-distributed, then homozygosity (or the probability that no mutation occurs on an interval of length  $T$ ) is  $e^{-2\mu T}$ , where  $\mu$  is the per-generation mutation rate (*Wakeley, 2009, p. 107*). We can therefore find  $m$ th moments of homozygosity as

$$\begin{aligned}
\mathbb{E}[J_{ij}^m] &= \int_0^\infty e^{-2m\mu t} f_{ij}(t) dt \\
&= \int_{\tau_0}^{\tau_{2(K-i)}} e^{-2m\mu t} \delta_{ij} \frac{e^{-(t-\tau_0)/N_{2(K-i)}}}{N_{2(K-i)}} dt + \sum_{h=2(K-i)}^{2K-1} \int_{\tau_h}^{\tau_{h+1}} e^{-2m\mu t} \Lambda_{ijh} \frac{e^{-(t-\tau_h)/N_h}}{N_h} dt \\
&= \frac{\delta_{ij}}{1 + 2N_{2(K-i)}m\mu} \left[ e^{-2m\mu\tau_0} - e^{-2m\mu\tau_{2(K-i)} - \frac{\tau_{2(K-i)} - \tau_0}{N_{2(K-i)}}} \right] \\
&\quad + \sum_{h=2(K-i)}^{2K-1} \frac{\Lambda_{ijh}}{1 + 2N_h m\mu} \left[ e^{-2m\mu\tau_h} - e^{-2m\mu\tau_{h+1} - \frac{\tau_{h+1} - \tau_h}{N_h}} \right]. \tag{5.5}
\end{aligned}$$

By the binomial theorem, the  $m$ th moment of heterozygosity is

$$\mathbb{E}[H_{ij}^m] = \mathbb{E}[(1 - J_{ij})^m] = \sum_{\ell=0}^m \binom{m}{\ell} (-1)^\ell \mathbb{E}[J_{ij}^\ell]. \quad (5.6)$$

Setting  $m = 1$  in eqs. 5.5 and 5.6, we get the expected homozygosity and expected heterozygosity for a pair of lineages, one randomly sampled from population  $E_i$  and the other randomly sampled from population  $E_j$ ,

$$\begin{aligned} \mathbb{E}[J_{ij}] &= \frac{\delta_{ij}}{1 + 2N_{2(K-i)}\mu} \left[ e^{-2\mu\tau_0} - e^{-2\mu\tau_{2(K-i)} - \frac{\tau_{2(K-i)} - \tau_0}{N_{2(K-i)}}} \right] \\ &\quad + \sum_{h=2(K-i)}^{2K-1} \frac{\Lambda_{ijh}}{1 + 2N_h\mu} \left[ e^{-2\mu\tau_h} - e^{-2\mu\tau_{h+1} - \frac{\tau_{h+1} - \tau_h}{N_h}} \right] \end{aligned} \quad (5.7)$$

$$\begin{aligned} \mathbb{E}[H_{ij}] &= 1 - \frac{\delta_{ij}}{1 + 2N_{2(K-i)}\mu} \left[ e^{-2\mu\tau_0} - e^{-2\mu\tau_{2(K-i)} - \frac{\tau_{2(K-i)} - \tau_0}{N_{2(K-i)}}} \right] \\ &\quad - \sum_{h=2(K-i)}^{2K-1} \frac{\Lambda_{ijh}}{1 + 2N_h\mu} \left[ e^{-2\mu\tau_h} - e^{-2\mu\tau_{h+1} - \frac{\tau_{h+1} - \tau_h}{N_h}} \right]. \end{aligned} \quad (5.8)$$

Using the model in Figure 5.2A, Figure 5.3 plots the expected heterozygosity of two lineages sampled from population 4 as a function of both bottleneck population size and bottleneck length. When the bottleneck has length zero, bottlenecks do not increase genetic drift and hence the expected heterozygosity reaches its maximum. Increasing the bottleneck length causes a monotonic decrease in expected heterozygosity. Decreasing the population size of the bottlenecks further decreases the heterozygosity. The smallest expected heterozygosity shown is reached with the combination of the smallest bottleneck population size (100 diploid individuals) and the largest bottleneck length (5000 generations).

#### 5.2.4 Pairwise $F_{ST}$

Our computation of expected coalescence times in eq. 5.4 provides a basis for obtaining the commonly used measure of genetic differentiation, pairwise  $F_{ST}$  between populations. Using the results of *Slatkin* (1991) on pairwise  $F_{ST}$  between populations at small mutation rates, we can write  $F_{ST} = (\bar{T} - \bar{T}_0)/\bar{T}$ , where  $\bar{T}_0$  is the mean coalescence time of two lineages randomly drawn from the same population and  $\bar{T}$  is the mean coalescence time of two lineages randomly drawn from any two populations (same or different). By using the expected coalescence times in our serial founder model (eq. 5.4), we can define these times for pairwise comparisons of populations  $E_i$  and  $E_j$  ( $i < j$ ) as  $\bar{T}_0 = (1/2)\mathbb{E}[T_{ii}] + (1/2)\mathbb{E}[T_{jj}]$ ,  $\bar{T}_{\text{diff}} = \mathbb{E}[T_{ij}]$  (the mean pairwise coalescence time for two lineages from different populations), and  $\bar{T} = (1/2)\bar{T}_0 + (1/2)\bar{T}_{\text{diff}}$ . Therefore, we can write  $F_{ST}$  as

$$F_{ST}^{ij} = \frac{\mathbb{E}[T_{ij}] - (1/2)\mathbb{E}[T_{ii}] - (1/2)\mathbb{E}[T_{jj}]}{\mathbb{E}[T_{ij}] + (1/2)\mathbb{E}[T_{ii}] + (1/2)\mathbb{E}[T_{jj}]}, \quad (5.9)$$

where the quantities  $\mathbb{E}[T_{ij}]$  are defined in eq. 5.4.

### 5.3 Patterns observed in human population data

In this section we describe a worldwide human population-genetic dataset and patterns in summary statistics calculated from the dataset. The summary statistics we investigate are within-population gene diversity, between-population gene identity, and pairwise  $F_{ST}$ . Analytical formulas for these summary statistics under the serial founder model are obtained through eqs. 5.7, 5.8, and 5.9. We compare patterns in these summary statistics observed in human genetic data to those predicted by specific models of human evolutionary history. Through these comparisons, we discuss which models of human history are compatible with patterns of genetic variation observed in present-day human populations. Note that only one of the three summary statistics

that we investigate (gene diversity) was discussed by *DeGiorgio et al.* (2009).

We analyzed data from the Human Genome Diversity Project-Centre d’Etude du Polymorphisme Humain (HGDP-CEPH) Cell Line Panel (*Cann et al.*, 2002; *Cavalli-Sforza*, 2005). We used a set of 783 autosomal microsatellite loci in 1048 individuals sampled from 53 worldwide populations (*Ramachandran et al.*, 2005; *Rosenberg et al.*, 2005). For a given population, gene diversity was calculated using eq. 10 of *DeGiorgio and Rosenberg* (2009), averaged across loci; the values were taken from Figure 7C of *DeGiorgio and Rosenberg* (2009). For a pair of distinct populations  $A$  and  $B$ , between-population gene identity was calculated as

$$J_{AB} = \frac{1}{L} \sum_{\ell=1}^L \sum_{i=1}^{I_{\ell}} \hat{p}_{\ell i} \hat{q}_{\ell i},$$

where  $\hat{p}_{\ell i}$  and  $\hat{q}_{\ell i}$  are the sample frequencies of the  $i$ th distinct allele at locus  $\ell$  in populations  $A$  and  $B$ , respectively, and  $I_{\ell}$  is the number of distinct alleles in the pair of populations at locus  $\ell$  (*Nei*, 1987). For a pair of distinct populations,  $F_{ST}$  was calculated using eq. 5.3 of *Weir* (1996).

Figure 5.4 displays patterns in the three summary statistics, as observed from the HGDP dataset. Figure 5.4A shows an approximate linear decline of gene diversity with increasing geographic distance from a putative East African location of modern human origins. Figure 5.4B shows a heat map of gene identity between all pairs of populations, illustrating that pairs of populations closer to Africa generally have lower between-population gene identity than pairs of populations farther from Africa. Figure 5.4C displays a heat map of pairwise  $F_{ST}$  between populations.  $F_{ST}$  is lower for pairs of populations that are close geographically than for pairs of populations that are geographically distant. Additionally,  $F_{ST}$  values between populations in the Americas are generally larger than  $F_{ST}$  values between pairs of non-American populations. In all three panels of Figure 5.4, a slight jump in the values of summary statistics is visible



at the boundaries of geographic regions. That is, separate values of gene diversity computed within populations from the same geographic region, and gene identity and  $F_{ST}$  values for pairs of populations from the same geographic region, tend to be more similar to each other than to corresponding values involving populations from different geographic regions.

Given the three patterns in summary statistics observed from the HGDP dataset, we can now compare these patterns with those predicted by models of human evolutionary history. We consider several special cases of our general serial founder model that are chosen based on previous investigations of human evolution. These cases include a modern serial founder model (*Ramachandran et al., 2005; Deshpande et al., 2009; DeGiorgio et al., 2009*), a nested regions model in which bottlenecks between continental regions are more severe than those within continental regions (*Hunley et al., 2009*), an instantaneous divergence model in which all populations diverged at the same time in the past (*DeGiorgio et al., 2009*), and an archaic serial founder model in which the founding process started distantly in the past (*DeGiorgio et al., 2009*). Using eqs. 5.7, 5.8, and 5.9, we now examine the patterns in gene diversity, between-population gene identity, and pairwise  $F_{ST}$  generated by these four special cases of the general serial founder model. We examine the extent to which each model can reproduce the patterns observed in worldwide human population-genetic data in the three statistics.

## 5.4 Modern serial founder model

### 5.4.1 Motivation and model

A modern serial founder model (Figure 5.5A) is a special case of our general formulation of the serial founder model (Figure 5.1). To obtain the *DeGiorgio et al. (2009)* serial founder model with  $K$  populations, suppose that the bottleneck length is

$L_b$  generations and that the amount of time between the end of a bottleneck and the founding of a new population is  $L$  generations. In other words, suppose  $\tau_{2h+1} - \tau_{2h} = L$  for  $h = 0, 1, \dots, K - 2$  and  $\tau_{2h} - \tau_{2h-1} = L_b$  for  $h = 1, 2, \dots, K - 1$ . Let  $\tau_0 = 0$ . Modern population 1 founds modern population 2 at time  $\tau_{2(K-1)} = \tau_{2K-1} = \tau_D$ . Each bottleneck has size  $N_b$  diploid individuals, and all other populations have size  $N$  diploid individuals. For the exact serial founder model studied by *DeGiorgio et al.* (2009), we set  $K = 100$ ,  $L_b = 2$ ,  $L = 19$ ,  $\tau_D = 2079$ ,  $N = 10000$ , and  $N_b = 250$ . These values were chosen to represent reasonable values for human populations. The value of  $\tau_D$  was chosen to lie within an estimated interval of time for the out-of-Africa migration (e.g., *Relethford*, 2008), the value of  $N$  was chosen as a commonly used value to represent the present-day effective population size of human populations (e.g., *Takahata et al.*, 1995), the value of  $N_b$  was chosen to represent a size typical for small isolated hunter-gatherer populations (*Cavalli-Sforza*, 2004), the value of  $L_b$  was chosen to represent a process in which individuals migrate in the first generation and finalize the settlement of a population during the second generation, and the value of  $L$  was chosen such that all founding events were distributed uniformly over  $\tau_D = 2079$  generations. Utilizing this parameterization and a per-generation mutation rate of  $\mu = 2.5 \times 10^{-5}$ , we examine whether the modern serial founder model can or cannot reproduce observed patterns of human genetic variation.

#### 5.4.2 Patterns generated by the model

Figure 5.6 displays patterns of genetic variation generated by the modern serial founder model. As was observed previously in simulations (*Ramachandran et al.*, 2005; *Deshpande et al.*, 2009; *DeGiorgio et al.*, 2009), the modern serial founder model reproduces the approximate linear decline in gene diversity with distance from the source population (Figure 5.6A). Figure 5.6B displays a heat map of pairwise gene identity values between all pairs of modern populations. The heat map shows

that populations close to the source population have smaller between-population gene identities than populations far from the source, as is observed in human population data (Figure 5.4*B*). Figure 5.6*C* displays a heat map of  $F_{ST}$  values between all pairs of modern populations, demonstrating that pairs of populations that are geographically distant tend to have larger  $F_{ST}$  than pairs of populations that are geographically close. The model largely recovers the pattern observed in human population data (Figure 5.4*C*); however, it also predicts small  $F_{ST}$  between pairs of populations that are far from the source population, a pattern that is not observed for human populations distant from Africa.

The pattern of decrease in gene diversity with increasing distance from a source population is due to the decrease in pairwise coalescence time within populations caused by a cumulative increase in genetic drift with increasing distance from the source. Pairs of lineages from distinct populations distant from the source have the potential to coalesce more recently than do pairs of lineages close to the source, thereby explaining the increased gene identity for pairs of populations distant from the source. However,  $F_{ST}$  between populations that are geographically distant from the source is smaller than  $F_{ST}$  between populations that are close to the source, as the effect of reduced between-population coalescence times in decreasing  $F_{ST}$  for populations distant from the source outweighs the effect of their reduced within-population coalescence times in increasing  $F_{ST}$ .

Our results show that the modern serial founder model largely recovers the patterns observed from human genetic data (Figure 5.4). Two noteworthy exceptions are that it does not predict either a peculiar pattern of small values of gene identity observed between Oceanian and non-Oceanian populations (Figure 5.4*B*) or the large values of  $F_{ST}$  observed in the Americas (Figure 5.4*C*).

## 5.5 Nested regions model

### 5.5.1 Motivation and model

One aspect of the trends in genetic diversity that was not captured by our parameterization of the modern serial founder model above is the larger difference in diversity observed between populations from different continental regions than between populations from the same continental region (Figure 5.4A). This observation motivates the nested regions model (Figure 5.5B) simulated by *Hunley et al.* (2009), in which the set of populations is distributed across several “regions” separated by barriers to migration. Examples of such regions include different continents, areas separated by mountain ranges, or islands within an archipelago. Because crossing between regions is more difficult than migration within a region, significant genetic drift might occur during the expansion into a new region. The nested regions model incorporates this increase in genetic drift during the geographic expansion through increased bottleneck severity between regions relative to bottleneck severity within regions.

We incorporate severe bottlenecks into the modern serial founder model (Figure 5.5A) by increasing the bottleneck lengths to  $L_b^r = 16$  generations instead of  $L_b = 2$  during the founding of modern populations 15, 29, 43, 57, 71, and 85. Hence, the length of time between the end of any of these bottlenecks and the founding of the next population is  $L^r = 5$  generations instead of  $L = 19$ , so that the time between founding events is still  $L_b + L = 21$  generations. These severe bottlenecks subdivide the set of  $K = 100$  modern populations into  $R = 7$  regions.

### 5.5.2 Patterns generated by the model

Figure 5.7 depicts patterns of genetic variation generated by the nested regions model. As was observed in simulations of *Hunley et al.* (2009), the nested regions

model reproduces the approximate linear decline in gene diversity with distance from the source population, with small discontinuities in genetic diversity at region boundaries (Figure 5.7A). Similarly, as was observed in the simulations of *Hunley et al.* (2009), the nested regions model reproduces the patterns of between-population gene identity observed from human data, with pairs of populations far from the source displaying larger gene identity than pairs of populations close to the source (Figure 5.7B). Also, in the nested regions model, pairs of populations that are geographically distant tend to have larger  $F_{ST}$  than pairs of populations that are geographically close (Figure 5.7C). The nested regions model predicts regional boundaries in the gene identity and  $F_{ST}$  heat maps (Figures 5.7B and C) that partly reproduce the block structure in the human population data (Figures 5.4B and C). However, as was seen with the modern serial founder model, the nested regions model predicts small  $F_{ST}$  between pairs of populations that are far from the source population, a pattern that is not observed for populations in the Americas (contrast Figure 5.4C and Figure 5.7C).

As was seen with the modern serial founder model above, the nested regions model recovers most of the patterns observed in human population-genetic data (Figure 5.4). Because of the increased bottleneck severity between regions, unlike the modern serial founder model, the nested regions model also reproduces the larger differences in values of the three summary statistics observed between regions compared to values observed within regions (Figure 5.4).

## 5.6 Instantaneous divergence model

### 5.6.1 Motivation and model

*DeGiorgio et al.* (2009) found that another model, the instantaneous divergence model, was capable of generating patterns that were compatible with observed

patterns of within-population gene diversity, linkage disequilibrium, and the ancestral allele frequency spectrum. Because we only investigated within-population summary statistics, however, it was not examined whether the gene identity and  $F_{ST}$  patterns observed in Figures 5.4*B* and *C* could also be generated by the instantaneous divergence model.

The instantaneous divergence model (Figure 5.5*C*) is a model in which all populations diverge at the same time in the past and populations that are farther from the source population have a smaller population size than those that are closer to the source. The motivation for this model is that populations that have traveled a greater distance from a source population will likely have lost alleles through genetic drift. The instantaneous divergence model allows for this increased drift for populations that are located far from the source population by assigning such populations a smaller size. An increase in genetic drift causes a decrease in gene diversity due to the random loss of alleles, as also occurs in bottlenecks. *DeGiorgio et al.* (2009) found that when the size of population  $i$  in the instantaneous divergence model was set so that the elapsed coalescent time was the same as in modern population  $i$  in the modern serial founder model, the approximate linear trend in gene diversity with distance from the source population was virtually indistinguishable from that of the modern serial founder model.

Suppose a modern serial founder model is parameterized as in Figure 5.6*A*. We obtain the instantaneous divergence model of *DeGiorgio et al.* (2009) by setting the divergence time of all  $K$  populations to  $\tau_D$ , the ancestral diploid population size to  $N$ , and the diploid size of population  $i$  to

$$N_i = \frac{\tau_D}{[\tau_D - (i - 1)L_b]/N + (i - 1)L_b/N_b}, \quad (5.10)$$

for  $i = 1, 2, \dots, K$ , where  $t_D$ ,  $N$ ,  $N_b$ ,  $L$ , and  $L_b$  are the parameters in the modern

serial founder model in the section “Modern serial founder model” (*DeGiorgio et al.*, 2009). The value of  $N_i$  is chosen so that  $\tau_D/N_i$  is the total duration in coalescent units of population  $i$ . To obtain the exact instantaneous divergence model described by *DeGiorgio et al.* (2009), we set  $\tau_D = 2079$ ,  $N = 10000$ ,  $N_b = 250$ ,  $L = 19$ , and  $L_b = 2$ . These values are the same values used for the modern serial founder model in Figure 5.6A. Using eq. 5.10 for the size of population  $i$  allows population  $i$  to experience the same level of genetic drift as modern population  $i$  in the modern serial founder model.

### 5.6.2 Patterns generated by the model

Figure 5.8 depicts patterns of genetic variation generated by the instantaneous divergence model. As was observed in the simulations of *DeGiorgio et al.* (2009), the instantaneous divergence model reproduces the approximate linear decline in gene diversity with increasing distance from the source population (Figure 5.8A). In contrast, between-population gene identity and pairwise  $F_{ST}$  yield patterns that are quite different from those observed in human data (contrast Figures 5.8B and C with Figures 5.4B and C). All off-diagonal entries of Figure 5.8B are identical, and they are the smallest values of gene identity within the heat map. Also, pairs of populations that are close to the source population have smaller  $F_{ST}$  than pairs of populations that are far from the source (Figure 5.8C).

The approximate linear decline in gene diversity produced by the instantaneous divergence model (Figure 5.8A) is caused by the loss of alleles and corresponding decrease in heterozygosity due to increased genetic drift within populations that are far from the source population (*DeGiorgio et al.*, 2009). However, the fact that all off-diagonal entries of Figure 5.8B are identical indicates that no correlation exists with geography for between-population gene identity under the instantaneous divergence model. This lack of correlation with geography for between-population

gene identity causes the pattern of pairwise  $F_{ST}$  values to be driven completely by the sizes of population pairs. Hence, population pairs far from the source location, which have smaller population sizes, and therefore smaller within-population coalescence times, have higher  $F_{ST}$  values.

Because the approximate linear decline in gene diversity (Figure 5.8A) generated by the instantaneous divergence model matches the pattern observed from human genetic data (Figure 5.4A), we can conclude that the pattern of within-population gene diversity observed from human data reflects the cumulative increase in genetic drift with increasing distance from Africa (*DeGiorgio et al.*, 2009). However, the patterns of between-population summary statistics generated by the instantaneous divergence model (Figures 5.8B and C) do not match the patterns observed from human genetic data (Figures 5.4B and C). Thus, a model that only incorporates a cumulative increase in genetic drift with increasing distance from a source is not sufficient to predict observed patterns of between-population genetic diversity.

## 5.7 Archaic serial founder model

### 5.7.1 Motivation and model

The serial founder model was motivated as a model to explain how modern humans expanded out of Africa and colonized the world. Our general serial founder model, however, does not place restrictions on the time of the first founding event. Therefore, our general model reduces to an archaic serial founder model (Figure 5.5D) when the time to the first founding event occurs distantly in the past. The archaic serial founder model, although it has an identical mathematical form to the modern serial founder model, is conceptually different in the sense that it is motivated by hypotheses regarding expansions of ancient hominids out of Africa, whereas the modern serial founder model is motivated by hypotheses of recent expansion of anatomically modern



humans out of Africa. The effect of increasing the time of the first founding event can be investigated in the serial founder model while holding all other parameters in the model constant.

In this section, we discuss how the patterns for within-population gene diversity, between-population gene identity, and pairwise  $F_{ST}$  change as the serial founding process is pushed farther into the past. To obtain an archaic serial founder model, we assume that except for divergence time  $\tau_D$ , all parameters are the same as in the modern serial founder model considered in Figure 5.6. We investigate divergence times of  $\tau_D = 5000$ , 7500, 10000, 16000, and 40000 generations ago. Divergence times  $\tau_D = 16000$  and  $\tau_D = 40000$  are of particular interest because, assuming a generation time of 25 years, they approximate estimates of the divergence of modern humans with Neanderthal (400 kya; *Green et al.*, 2006; *Noonan et al.*, 2006) and *Homo erectus* (1 mya; *Takahata*, 1993) populations, respectively.

### 5.7.2 Patterns generated by the model

For  $\tau_D = 5000$ , a decrease occurs, relative to the modern serial founder model in which  $\tau_D = 2079$ , in the magnitude of the slope of the decline of gene diversity with increasing distance from the source population (Figure 5.9A). The patterns of increased gene identity and decreased  $F_{ST}$  between populations that are far from the source population relative to between populations that are close to the source, although still present, are less distinct with the increased divergence time. Further increasing the divergence time to  $\tau_D = 7500$  (Figure 5.9B) and  $\tau_D = 10000$  (Figure 5.9C) leads to a progressive decrease in the differences among populations in values of the three summary statistics. For a serial founder model with a divergence time of  $\tau_D = 16000$ , at a putative time of the Neanderthal divergence, differences in values among populations for each of the three summary statistics are small (Figure 5.9D). For the *H. erectus* serial founder model with a divergence time of

$\tau_D = 40000$ , differences in values among populations for each of the three summary statistics are nearly negligible, displaying almost no trend (Figure 5.9E).

As  $\tau_D$  increases, the differences among populations in values of gene diversity, between-population gene identity, and  $F_{ST}$  decrease. These smaller differences result from the smaller degree of influence that ancient bottlenecks have on genetic diversity in comparison with recent bottlenecks of identical severity. This lack of influence of ancient bottlenecks on present-day gene diversity is reflected most strongly in the small difference in gene diversity between population 1 and population 100 in the *H. erectus* serial founder model (Figure 5.9E). Furthermore, with greater  $\tau_D$ , the difference between the divergence time for two populations sampled close to the source and for two populations sampled far from the source is small relative to  $\tau_D$ . This small difference in divergence times causes between-population summary statistics such as gene identity and  $F_{ST}$  to have little correlation with geography (*i.e.*, most off-diagonal entries have similar values) at large divergence times (Figure 5.9E).

These results imply that the patterns in gene diversity, gene identity, and  $F_{ST}$  observed from empirical data cannot be predicted solely by an archaic serial founder process using our parameterization; specifically, the observed patterns are not consistent with a serial founder process that occurs too far back in the past. Pushing back the time of the first founding event while holding all other parameters constant decreases the ability of the serial founder model to generate the patterns observed in Figure 5.4.

## 5.8 Discussion

In this article, we have derived pairwise coalescence-time distributions for a serial founder model. Under the model, we have provided analytical formulas for expected coalescence times, expected homozygosity, and pairwise  $F_{ST}$ . In addition, we have analytically described the trend of decreasing gene diversity with increasing distance

from the source population, and the patterns observed in between-population gene identity and pairwise  $F_{ST}$ . Using coalescence-time densities in a variety of special cases, we have found that the modern serial founder model and the nested regions model are consistent with geographic patterns of within- and between-population genetic diversity observed in human data. Our work demonstrates the utility of using theoretical computations on between-population summary statistics in conjunction with similar computations on within-population statistics to predict geographic patterns in genetic data.

One pattern that was not predicted by any of our models was the large  $F_{ST}$  observed in the Americas. Whereas the modern serial founder and the nested regions models predict small  $F_{ST}$  between populations far from the source,  $F_{ST}$  values in the Americas are large. It is possible that the models provide a poor fit to the pattern of evolution in the Americas after the initial founding of the Native American population, as they also are inconsistent with the large differences in gene diversity among populations in the Americas. During the initial migration into the Americas, small individual populations may have experienced highly variable levels of genetic drift as they spread over a large unoccupied region (*Wang et al.*, 2007; *Goebel et al.*, 2008; *Meltzer*, 2009). Such a migration process could have given rise to highly variable levels of genetic diversity across the region, as well as a somewhat irregular pattern in  $F_{ST}$ . If we were to modify our model to incorporate this variability along with stronger bottlenecks or smaller population sizes within the Americas relative to those in non-American populations, then we might be able to produce patterns that agree with the observed data. Indeed, *Hunley et al.* (2009) found that model parameters can be chosen to enable patterns of within- and between-population genetic diversity to closely match those empirically observed in the Americas.

Another pattern that was not predicted by any of our models is the small between-population gene identity observed between pairs of populations, one from

Oceania and the other not from Oceania (Figure 5.4B). This pattern could potentially be explained either by an ancient divergence of the Oceanian populations from the non-Oceanian populations through a separate migration out of Africa to Oceania (e.g., *Derricourt, 2005; Bulbeck, 2007; Field et al., 2007; Szpiech et al., 2008; Kayser, 2010*), or by admixture of the populations in Oceania with an archaic human population (e.g., *Reich et al., 2010*). A separate founding process could have generated low levels of within-population gene diversity for the Oceanian populations while simultaneously producing the low levels of between-population gene identity between Oceanian and non-Oceanian populations. Alternatively, because the increase in between-population coalescence times that would be caused by ancient admixture would result in a decrease in between-population gene identity, such admixture could potentially explain the disagreement of the data with our model predictions. Separate migrations or ancient admixture could potentially be incorporated into a more general version of our model to investigate the plausibility of these scenarios.

By increasing the time of the first founding event, we have determined that the archaic serial founder model is not able to reproduce patterns of gene diversity, between-population gene identity, and pairwise  $F_{ST}$  observed in human genetic data. However, limited archaic admixture coupled with a modern serial founder model might not be incompatible with the patterns we have examined. Recent evidence suggests that signatures of archaic admixture might exist in modern human population-genetic data (e.g., *Plagnol and Wall, 2006; Garrigan and Hammer, 2006; Green et al., 2010; Reich et al., 2010*) and as discussed above, such admixture could potentially explain anomalous observations in Oceania. However, this admixture, if it indeed occurred, must have been insufficient to generate a large signature in most of the patterns that we have investigated.

Although the patterns of gene diversity produced by the serial founder and the instantaneous divergence models are virtually indistinguishable (*DeGiorgio*

*et al.*, 2009), we have shown that these models can be differentiated using between-population gene identity and pairwise  $F_{ST}$ . Ultimately, this potential for differentiation traces to distinctive distributions of pairwise coalescence times. In the instantaneous divergence model, each population has a constant size up until time  $\tau_D$  and consequently, the coalescent process simply follows an exponential distribution until time  $\tau_D$  and then another exponential distribution with a different rate after time  $\tau_D$ . In contrast, in the serial founder model, the rate of coalescence inside a bottleneck is elevated compared to outside the bottleneck. This increased rate of coalescence causes lineages to merge within a narrow time interval. Because the serial founder model incorporates multiple bottlenecks, the distribution of coalescence times is multimodal.

Recently, many studies have found that two-dimensional spatial maps generated from principal components analysis (PCA) applied to human genetic data closely match maps of geographic sampling locations of populations (e.g., *Lao et al.*, 2008; *Novembre et al.*, 2008; *Price et al.*, 2009; *Bryc et al.*, 2010; *Wang et al.*, 2010; *Xing et al.*, 2010). *McVean* (2010) demonstrated a close link between pairwise coalescence times and PCA, in which sampled lineages can be projected onto principal components through expected coalescence times for pairs of lineages. The coalescence time distributions provided in this article can potentially be used to interpret PCA maps, so that PCA maps themselves might be used as summary statistics for testing evolutionary models.

Estimated coalescence-time distributions might also be utilized more formally for maximum likelihood estimation of parameters such as bottleneck lengths, bottleneck sizes, and divergence times (e.g., *Thomson et al.*, 2000; *Takahata et al.*, 2001; *Tang et al.*, 2002; *Rannala and Yang*, 2003; *Tishkoff and Verrelli*, 2003; *Garrigan and Hammer*, 2006; *Fagundes et al.*, 2007; *Blum and Jakobsson*, 2011). Further, these distributions might also be useful for hypothesis testing; because many of the models

in this article are nested, likelihood ratio tests can be performed. For extending our work to perform maximum likelihood inference, it will be desirable to extend the computations to permit the sampling of multiple lineages in pairs of populations (*i.e.*,  $n_i$  lineages from population  $i$  and  $n_j$  lineages from population  $j$ ). Such an extension could potentially build upon the work of *Marth et al.* (2004), who derived the coalescence-time distribution for a sample of  $n$  lineages in a single population with multiple bottlenecks.

An additional feature of structured population models that would be desirable to incorporate is migration between populations after their initial founding. In the archaic serial founder model, some level of migration between neighboring populations might enable the model to make predictions that more closely match observations from human genetic data. For the modern serial founder model, simulations have shown that small to moderate levels of migration have relatively little impact on observed patterns of genetic diversity (*DeGiorgio et al.*, 2009). In any case, inclusion of migration would enable us to examine considerably more complex versions of the models that we have investigated.

Finally, one important quantity that we did not explore is linkage disequilibrium (LD). In simulations, we previously investigated whether the spatial distribution of LD observed in worldwide human populations is consistent with a serial founder model (*DeGiorgio et al.*, 2009). We found that the serial founder model can indeed predict the observed spatial distribution of LD. Moreover, we found that LD patterns can be useful in differentiating the patterns predicted by different evolutionary models. Therefore, incorporation of LD into our theoretical models would provide a distinct type of statistic that would further enable investigators to distinguish between models. For example, because excess long-range LD is a signature of ancient admixture (*e.g.*, *Plagnol and Wall*, 2006), incorporation of LD statistics would be useful for determining whether models that consider archaic admixture provide a

significantly better fit to observed human genetic variation than models that do not consider admixture. Because LD is such a valuable quantity, it would be informative to investigate patterns of LD produced by the various models by incorporating recombination into the theory.

## **5.9 Acknowledgments**

We thank Raquel Assis, Laurent Excoffier, Zach Szpiech, and two anonymous reviewers for their valuable comments. This work was supported by NIH grant R01 GM081441, NIH training grants T32 GM070449 and T32 HG000040, a grant from the Burroughs Wellcome Fund, a University of Michigan Rackham Merit Fellowship, and the New Zealand Marsden Fund.

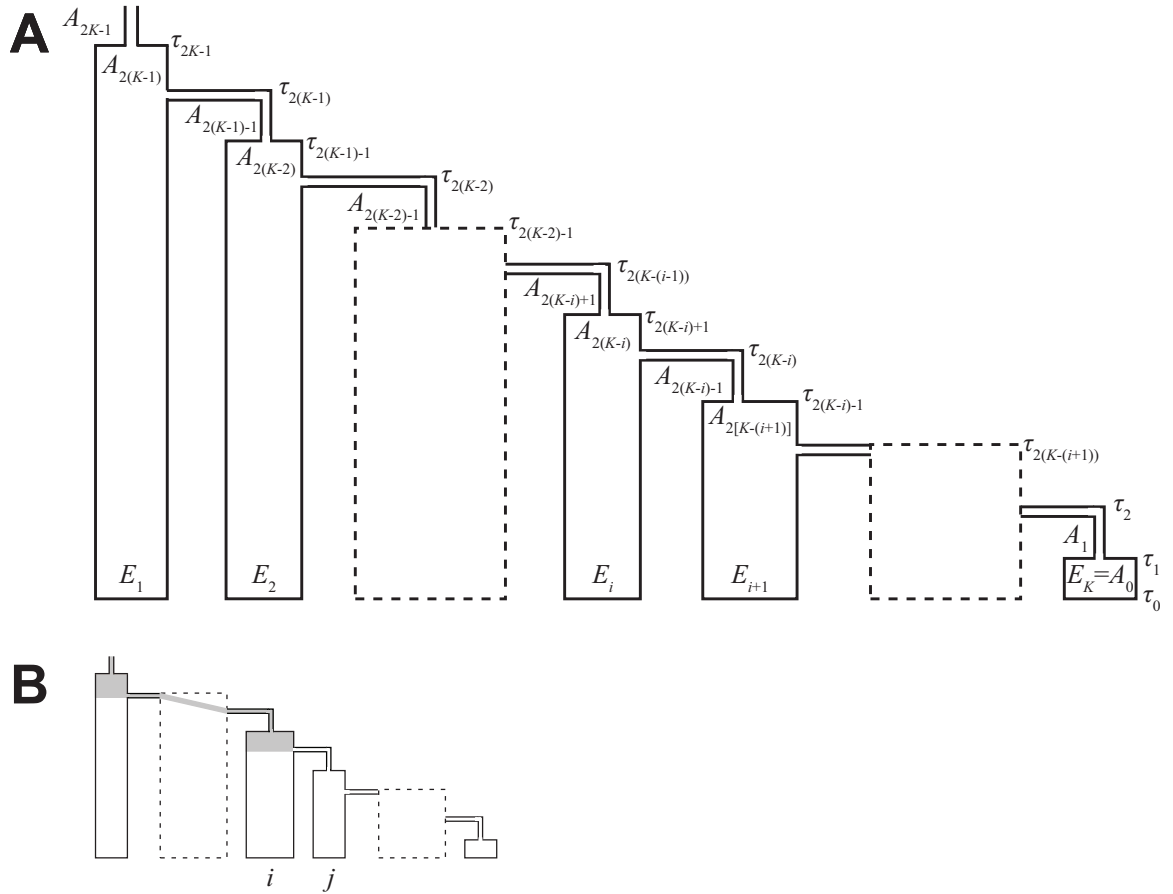


Figure 5.1: Serial founder model. (A) Serial founder model with  $K$  extant and  $2K$  ancestral populations. At time  $\tau_{2K-1}$ , ancestral population  $A_{2K-1}$  expands to a larger size to form ancestral population  $A_{2(K-1)}$ . Next, at time  $\tau_{2(K-1)}$ , ancestral population  $A_{2(K-1)}$  splits to form extant population  $E_1$  and newly founded ancestral population  $A_{2(K-1)-1}$ . At time  $\tau_{2(K-1)-1}$ , population  $A_{2(K-1)-1}$  expands to a larger size to form ancestral population  $A_{2(K-2)}$ . In general, at time  $\tau_{2(K-i)}$ , ancestral population  $A_{2(K-i)}$  splits into extant population  $E_i$  and newly founded ancestral population  $A_{2(K-i)-1}$ . At time  $\tau_{2(K-i)-1}$ , ancestral population  $A_{2(K-i)-1}$  expands to a larger size to form ancestral population  $A_{2(K-i-1)}$ . (B) Scenario in which lineages are sampled from populations  $E_i$  and  $E_j$ ,  $i \leq j$  ( $i < j$  is shown here). Regions in which coalescence can occur are represented in gray.



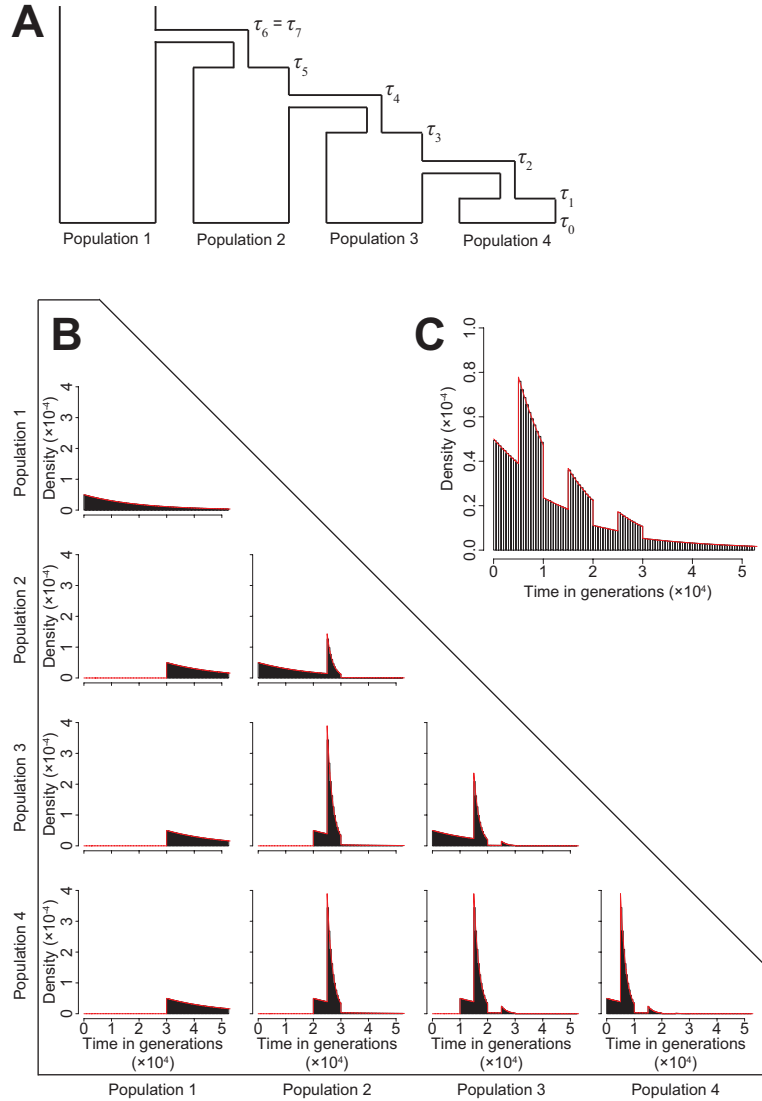


Figure 5.2: Distributions of coalescence times in the serial founder model. (A) Serial founder model with four extant populations. In the model, thick population sizes represent 10000 diploid individuals and thin population sizes represent 1000 diploid individuals. The times of founding events and population expansions are  $\tau_0 = 0$ ,  $\tau_h = \tau_{h-1} + 5000$  for  $h = 1, 2, \dots, 6$ , and  $\tau_6 = \tau_7 = 30000$  generations. (B) Probability density of coalescence times. Each sub-plot is the probability density of coalescence times for a pair of lineages sampled from the pair of populations listed in the row and column. (C) Probability density of coalescence times for a pair of lineages sampled from population 4, with identical parameter values to part A except that the bottlenecks (thin populations) have 5000 diploid individuals instead of 1000 diploid individuals. The figure can be compared with the plot for two lineages from population 4 in part B. Histograms are based on  $10^7$  coalescent simulations using MS (Hudson, 2002), and the red lines represent the analytical densities obtained from eq. 5.1.

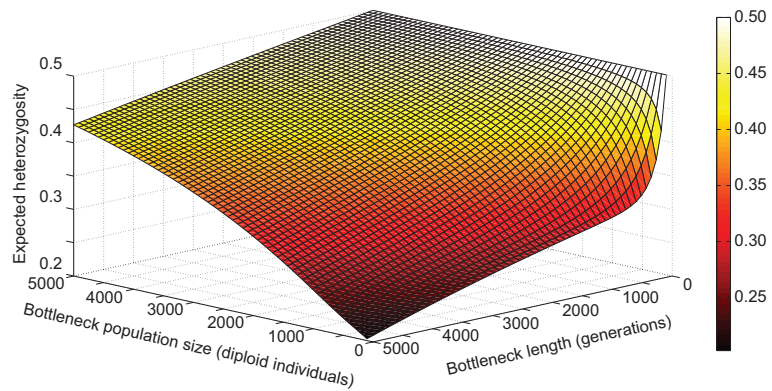


Figure 5.3: Expected heterozygosity for a pair of lineages sampled from population 4 of Figure 5.2A (eq. 5.8), as a function of population size for bottlenecks and bottleneck length measured in generations. A per-generation mutation rate of  $\mu = 2.5 \times 10^{-5}$  is assumed.

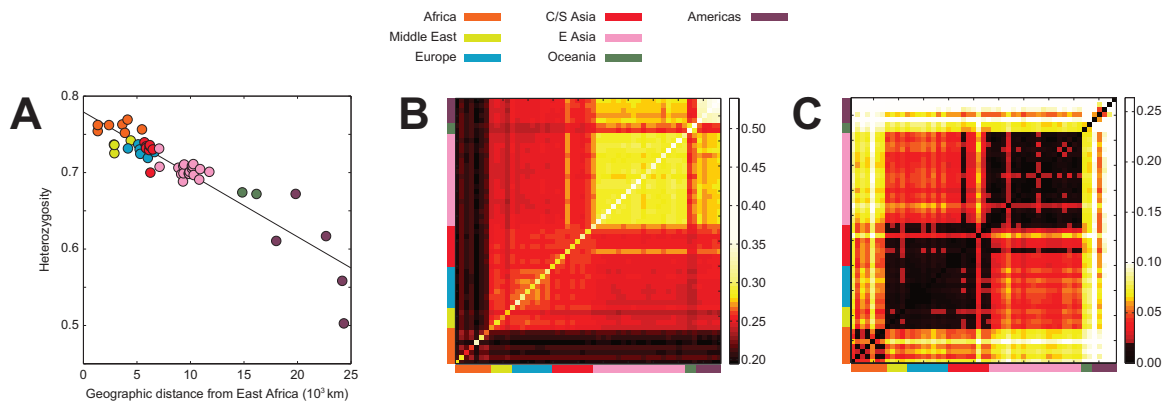


Figure 5.4: Patterns of within- and between-population summary statistics observed in human population-genetic data. Plots are based on 783 microsatellite loci from 53 worldwide populations in the HGDP-CEPH dataset (*Ramachandran et al., 2005; Rosenberg et al., 2005*). (A) Gene diversity as a function of distance from East Africa (redrawn from *DeGiorgio et al. (2009)*). Each point represents a particular population. (B) Between-population gene identity. Columns and rows each represent populations, and an entry in the matrix represents the gene identity for the population pair represented by the row and column. (C) Pairwise  $F_{ST}$  calculated from the same populations as in part B.

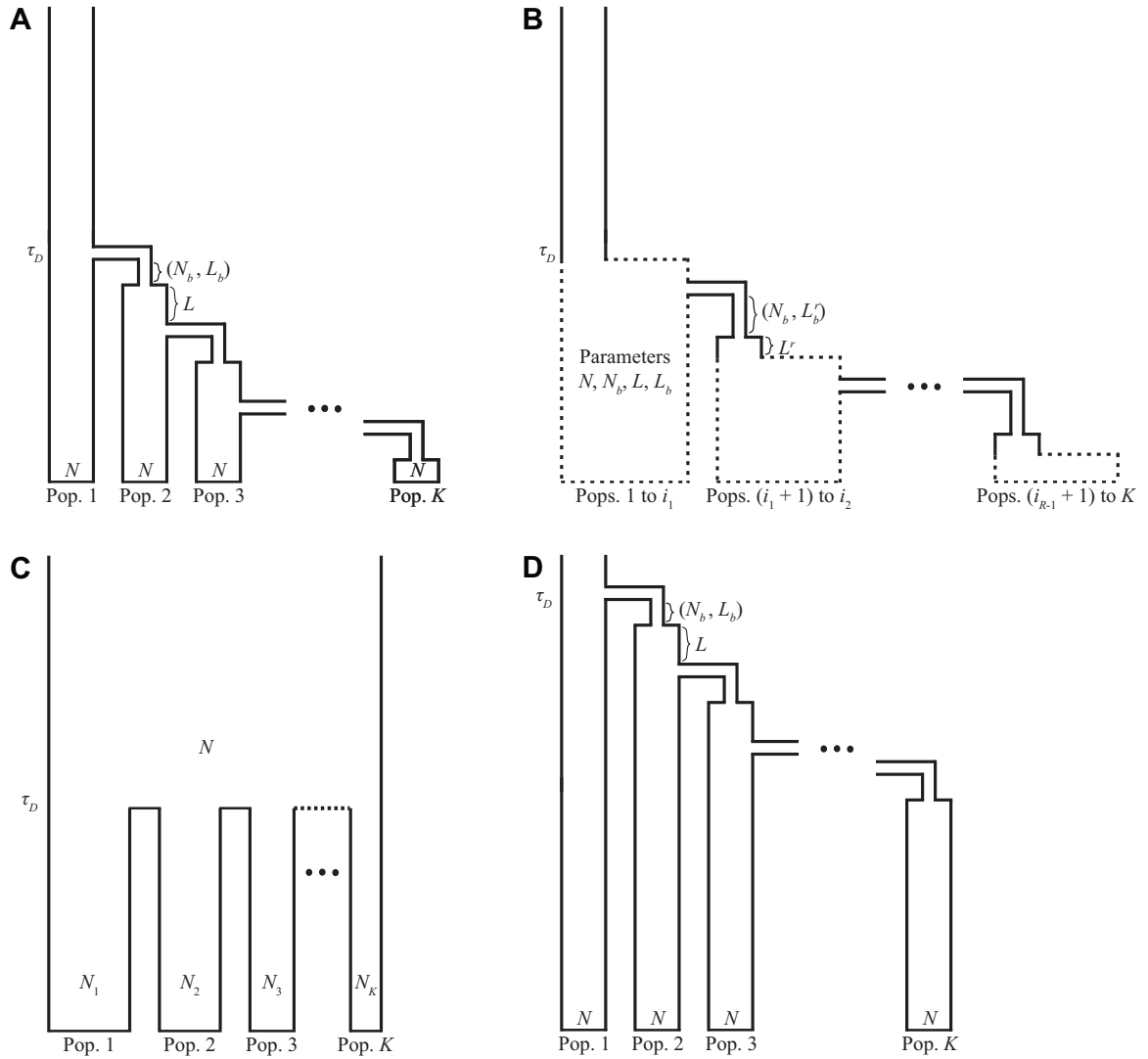


Figure 5.5: Models to which the general serial founder model reduces. (A) Modern serial founder model. (B) Nested regions model with  $R$  regions. (C) Instantaneous divergence model. (D) Archaic serial founder model. The models in parts A and C are exactly the models discussed by *DeGiorgio et al.* (2009).

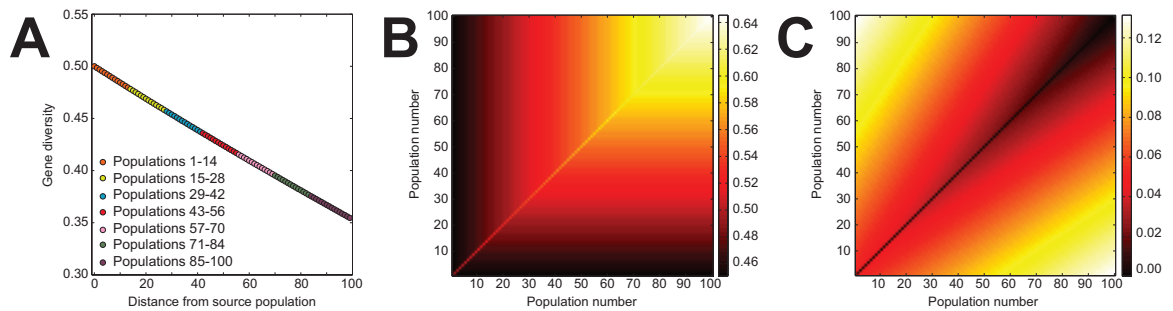


Figure 5.6: Patterns of genetic variation in a modern serial founder model. The values of the model parameters are indicated in the section “Modern serial founder model.” (A) Gene diversity of the populations as a function of distance from the source population, where distance is measured in units of populations. (B) Between-population gene identity for pairs of populations. (C) Pairwise  $F_{ST}$  for pairs of populations.

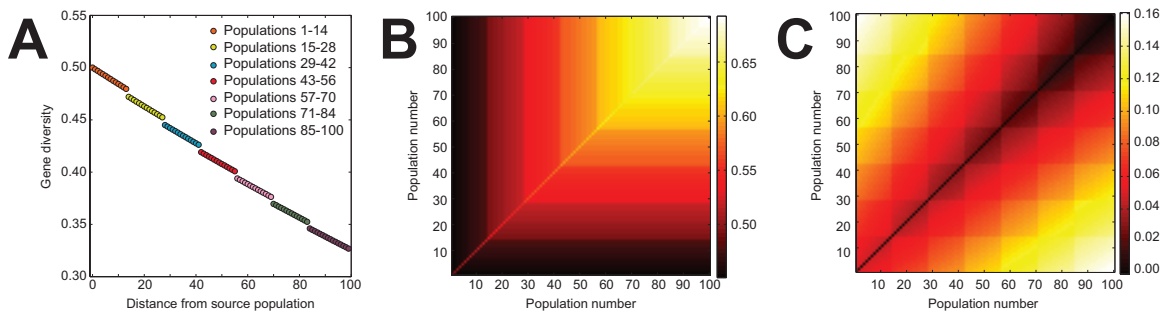


Figure 5.7: Patterns of genetic variation in a nested regions model. The values of the model parameters are the same as in Figure 5.6, with the exception that the bottleneck lasts 16 generations instead of 2 generations during the founding of modern populations 15, 29, 43, 57, 71, and 85. (A) Gene diversity of the populations as a function of distance from the source population, where distance is measured in units of populations. (B) Between-population gene identity for pairs of populations. (C) Pairwise  $F_{ST}$  for pairs of populations.

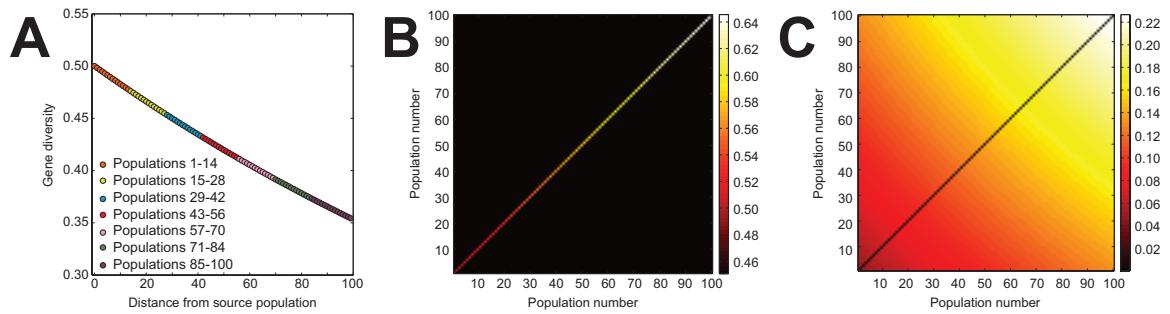


Figure 5.8: Patterns of genetic variation in the instantaneous divergence model. The values of the model parameters are indicated in the section “Instantaneous divergence model.” (A) Gene diversity of the populations as a function of distance from the source population, where distance is measured in units of populations. (B) Between-population gene identity for pairs of populations. (C) Pairwise  $F_{ST}$  for pairs of populations.

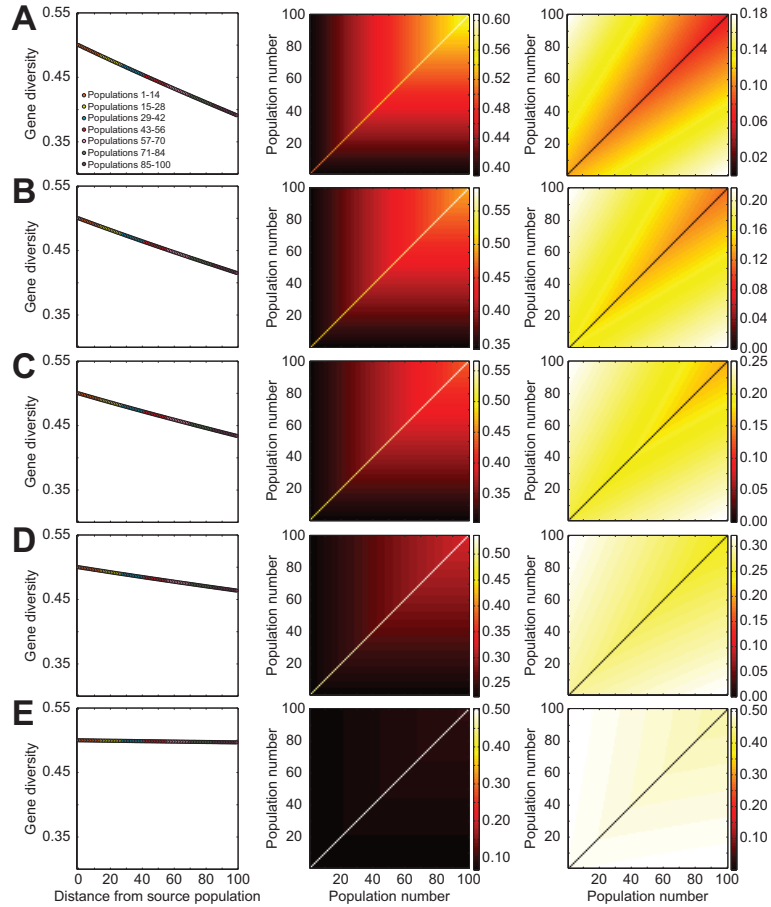


Figure 5.9: Patterns of genetic variation in an archaic serial founder model, as a function of varying divergence time  $\tau_D$ . The values of the model parameters for parts *A-E* are the same as in Figure 5.6*A*, with the exception that the divergence time  $\tau_D$ , measured in generations, varies across the plots. The first column is gene diversity of the populations as a function of distance from the source population, where distance is measured in units of populations. The second column is between-population gene identity for pairs of populations. The third column is pairwise  $F_{ST}$  for pairs of populations. (*A*)  $\tau_D = 5000$ . (*B*)  $\tau_D = 7500$ . (*C*)  $\tau_D = 10000$ . (*D*)  $\tau_D = 16000$ . (*E*)  $\tau_D = 40000$ .



## CHAPTER VI

# Fast and consistent estimation of species trees using supermatrix rooted triples

### 6.1 Introduction

A species tree is a branching pattern representing the divergence of multiple species, whereas a gene tree depicts the evolutionary history of a single gene. Though only a single species tree exists, trees for different genes often have conflicting topologies. This discordance of gene trees with the species tree is due to processes such as gene duplication, horizontal gene transfer, and incomplete lineage sorting (*Page and Charleston, 1997; Maddison, 1997; Than et al., 2007; Degnan and Rosenberg, 2009*).

When analyzing data from multiple loci, the most frequently occurring gene tree topology is sometimes used as an estimate of the species tree topology. For example, in a study of 30 loci, *Jennings and Edwards (2005)* used the gene tree that was inferred in 16 of 28 resolved topologies from three ingroup species of Australian grass finches as the species tree topology. However, even in the absence of complications such as hybridization (*Buckley et al., 2006; Holland et al., 2008; Meng and Kubatko, 2009*) and population structure (*Slatkin and Pollack, 2008*), this procedure is only justified for studies of three taxa. This is because the most likely three-taxon gene

tree is expected to match the species tree topology when incomplete lineage sorting is modeled by the multispecies coalescent (*Nei*, 1987; *Pamilo and Nei*, 1988). However, when a species tree has four taxa and is asymmetric, or has five or more taxa, the most likely gene tree does not necessarily match the species tree (*Degnan and Rosenberg*, 2006; *Rosenberg and Tao*, 2008). Such anomalous gene trees (AGTs; *Degnan and Rosenberg*, 2006) occur when the species tree falls into a particular space of branch lengths called the anomaly zone. Anomaly zones for four-taxon and five-taxon species trees are depicted in Figure 2 of *Degnan and Rosenberg* (2006) and Figures 3–5 of *Rosenberg and Tao* (2008), respectively.

The absence of AGTs for rooted three-taxon trees motivates the development of methods for inferring species trees using rooted triples, or three taxa at a time (*Degnan and Rosenberg*, 2006), as has been described for rooted triple consensus methods (*Ewing et al.*, 2008; *Degnan et al.*, 2009) and supertree methods (*Steel and Rodrigo*, 2008; *Willson*, 2009). Supertree methods generalize consensus methods to the setting in which input gene trees have overlapping subsets of taxa that need not be identical (*Bininda-Emonds*, 2004). Because a rooted tree is completely described by its set of rooted triples (*Steel*, 1992), we can utilize a supertree method to construct the species tree from correctly inferred rooted triples.

Supertree and other phylogenetic methods can be applied to sets of concatenated alignments, or supermatrices, to infer a species tree. A concatenated alignment contains sequences of multiple loci linked together to create a single “supergene” (*Rokas et al.*, 2003; *de Queiroz and Gatesy*, 2007), thus increasing the size of the dataset. Though statistical power generally increases with the size of a dataset, the accuracy of concatenation is currently under debate. *Rokas et al.* (2003) reported that the application of phylogenetic inference methods to concatenated sequence alignments can yield a strongly supported inferred species tree. However, several studies (*Kolaczkowski and Thornton*, 2004; *Mossel and Vigoda*, 2005; *Edwards*

*et al.*, 2007; *Kubatko and Degnan*, 2007) have also shown that inferring trees from concatenated data with maximum likelihood (ML) can perform poorly when sites are generated under different tree topologies and can produce bootstrap values that are misleadingly high (*Gadagkar et al.*, 2005; *Kubatko and Degnan*, 2007).

Here, we develop a divide-and-conquer approach (*Cormen et al.*, 2001) called SuperMatrix Rooted Triple (SMRT), which is a polynomial-time algorithm that circumvents some of the weaknesses of concatenation by linking it with rooted triple and supertree methods. SMRT assembles rooted triples inferred from concatenated alignments into a species tree using a supertree algorithm such as modified mincut (MMC) (*Page*, 2002). We compare SMRT in which rooted triples are inferred by maximum likelihood (SMRT-ML) to the method in which all taxa are analyzed simultaneously by applying ML to a supermatrix (SM-ML). In simulations that assume a molecular clock, SMRT-ML performs favorably on four- and five-taxon species trees both inside and outside the anomaly zone. Further, introducing two model violations—analysis under a molecular clock when gene trees are not clocklike and analysis under an incorrect substitution model—has little effect on the performance of SMRT-ML. We illustrate the SMRT-ML procedure using a yeast dataset frequently analyzed in phylogenetic studies (*Rokas et al.*, 2003; *Gatesy and Baker*, 2005; *Edwards et al.*, 2007) and find that SMRT-ML recovers the same species tree as that found using either SM-ML or the software BEST (*Liu*, 2008).

Assuming that incomplete lineage sorting is the source of discordance of gene trees with species trees and that there are no hybridization or horizontal gene transfer events, we prove that SM-ML is a statistically consistent estimator for three-taxon clocklike species trees when concatenated sequence alignments are generated from a coalescent distribution under a molecular clock and a binary substitution model (*Neyman*, 1971). Under the same set of assumptions, we then prove in Theorem VI.4 that SMRT-ML is a statistically consistent estimator of a species trees. Therefore,

our computationally efficient strategy is justified both theoretically and through simulations in the context of gene tree conflict due to incomplete lineage sorting. Although we assume here that rooted triples are inferred using a ML method, we stress that SMRT is a general approach that can utilize rooted triples that have been inferred from other methods such as parsimony and distance methods as well. We focus on triples inferred from ML because we compare our method to a method in which trees are inferred by ML from concatenated alignments.

## 6.2 Methods

### 6.2.1 Supermatrix rooted triple (SMRT)

The SMRT approach takes a concatenated alignment of  $n$  taxa and breaks it into  $\binom{n}{3}$  alignments, one for each set of three taxa. A rooted three-taxon tree is inferred for each alignment using any phylogenetic method by either assuming a molecular clock, or by including a known outgroup as a fourth taxon to root the tree. The species tree is then constructed by using the resulting rooted triples as input for a supertree algorithm. Here, we use MMC, which extends the mincut algorithm (*Semple and Steel, 2000*). The mincut algorithm satisfies five desirable properties: (1) the order of the input set of trees does not affect the method; (2) relabeling the set of taxa of the input trees produces the same output tree on the relabeled set of taxa; (3) if there exists a tree that has each input tree as a subtree, then the output tree will display all of these trees; (4) any taxon that is in the input set of trees is also in the output tree; and (5) the method is polynomial in the number of distinct taxa (*Semple and Steel, 2000*). *Page (2002)* created MMC by modifying the mincut method so that uncontradicted nestings are preserved in the output tree.

### 6.2.2 Simulation

We examined the performance of SMRT-ML using simulated sequence alignments. First we chose a species tree  $\sigma$  with topology  $((AB)C)D$ ,  $((AB)(CD))$ ,  $((((AB)C)D)E)$ ,  $((((AB)C)(DE)))$ , or  $((((AB)(CD))E))$ . Model species tree topologies are depicted in Figure 6.1. Branch lengths and probabilities for the matching gene tree topology and most probable nonmatching gene tree topologies are shown in Table 6.1. The branch lengths chosen for the species tree  $((AB)C)D$  are the same as those used in *Kubatko and Degnan (2007)*. One additional case was considered in which both internal branch lengths equal 0.1 coalescent units for the  $((AB)C)D$  species tree. For each species tree and each simulation replicate, using COAL (*Degnan and Salter, 2005*) conditional on  $\sigma$ , we simulated  $m = 100, 200, 300, 400, 500, 600, 700, 800, 900, 1000, 2000, 3000, 4000, 5000$ , and 6000 independent (within and across each set) gene trees with branch lengths. Branch lengths were simulated in coalescent units,  $t/(2N_e)$ , where  $t$  is the number of generations, and  $N_e$  is the effective population size. We converted the branch lengths for each gene tree to mutation units by multiplying each length by  $\theta/2$ , where  $\theta = 4N_e\mu$ , and  $\mu$  is the mutation rate per site per generation. As a consequence, all populations had equal values of  $\theta$ . For each gene tree, we converted branch lengths to the expected number of mutations by multiplying them by  $\theta/2$ , where  $\theta = 0.01$ . We generated sequence alignments of length  $L = 500$  nucleotides (nt) with SEQ-GEN (*Rambaut and Grassly, 1997*). These  $m$  independent alignments were concatenated to create single  $n$ -taxon alignments of length  $mL$ .

The concatenated alignments were then broken into all  $\binom{n}{3}$  three-taxon alignments of length  $mL$ . We inferred rooted ML trees for the  $n$ -taxon alignment, as well as for all three-taxon alignments, employing an exhaustive search over all tree topologies from PAUP\* (*Swofford, 2003*). All three-taxon rooted trees were entered as input to the program SUPERTREE (*Page, 2002*), which implements the MMC algorithm. Each

time PAUP\* was called, it returned  $k \geq 1$  tree topologies tied for the most likely species tree. The count for each of the tied topologies was increased by  $1/k$ . We repeated this procedure, beginning with the simulation of gene trees, 300 times for each combination of species tree topology and number of loci. The count for each tree topology was averaged over all replicate simulations. Unless otherwise stated, the results are for data simulated under a Jukes-Cantor (JC) model and analyzed under ML assuming JC and a molecular clock. A schematic of the simulation procedure is provided in Figure 6.2.

### 6.2.3 Empirical example

SMRT-ML was applied to analyze a yeast dataset consisting of 106 genes spanning over 127,000 nt (*Rokas et al.*, 2003). We used ML in PAUP\* under a GTR +  $\Gamma$  + I model without a molecular clock on each of the  $\binom{7}{3} = 35$  three-taxon subsets of the seven ingroup taxa, using the outgroup *C. albicans* to root the triples. In addition to the full concatenated alignment, we analyzed concatenated alignments of random subsets of  $m = 10, 20, 30, 40, 50, 60, 70, 80, 90$ , and 100 genes. For each value of  $m$ , SMRT-ML was applied to 300 random subsets of  $m$  genes, and we reported the proportion of times that SMRT-ML returned either the presumed species tree or a tree with at least one false clade. Bootstrapping for SMRT-ML was performed by reading the concatenated sequence data in R (*R Development Core Team*, 2008) and using the SAMPLE function to create 300 bootstrap eight-taxon alignments, SMRT-ML was applied to each bootstrap replicate in separate PAUP\* runs.

## 6.3 Results for simulations

### 6.3.1 Four taxa

A four-taxon asymmetric species tree is depicted in Figure 6.1A. Figures 6.3A–H and Figures 6.3I–P display simulation results for this species tree for SM-ML and SMRT-ML, respectively. As shown by *Kubatko and Degnan* (2007) and replicated here, SM-ML is misleading in that increasing the number of loci can make it more likely to return an incorrect species tree. In contrast, SMRT-ML outperformed SM-ML on the (((AB)C)D) species tree for all branch lengths tried except for  $(x, y) = (0.25, 0.01)$  (Figures 6.3H,P), where both methods performed poorly. For these branch lengths, using 6000 loci, SM-ML returned the species tree 52% of the time, SMRT-ML returned the species tree 54% of the time, and both methods returned each of the nonmatching trees (((AC)B)D) and (((BC)A)D) less than 25% of the time. For extremely small branch lengths of 0.01, the proportion of times that SMRT-ML recovers the species tree topology increases slowly (Figures 6.3I,M,P). However, the method does not appear to be misleading, suggesting that there is a tradeoff between consistency and speed of convergence as was seen for consensus methods by *Degnan et al.* (2009). For these sets of branch lengths, the proportion of times that SM-ML returned the species tree either increased just as slowly (Figure 6.3H) or was misleading (Figures 6.3A,E). Even though SMRT-ML did not always infer the matching species tree when  $x = 0.01$ , it often inferred the partially unresolved tree ((AB)CD) (e.g., Figures 6.3I,M), which is not misleading for the species tree topology. On the other hand, for  $(x, y) = (0.1, 1.0)$  (Figures 6.3C,K), both methods converged to the species tree with SMRT-ML converging more quickly than SM-ML; however, only SMRT-ML was increasingly likely to recover the species tree as loci were added for all branch lengths tried.

Simulation results on the four-taxon symmetric species tree (Figure 6.1B) are

shown in Figures 6.4A–H (SM-ML) and Figures 6.4I–P (SMRT-ML). In contrast to what was observed for the asymmetric tree, for the symmetric tree SM-ML is not misleading and converges to the true species tree faster than SMRT-ML for each set of branch lengths tested. This observation is not surprising, given that no anomaly zone exists for the four-taxon symmetric species tree and that SM-ML simultaneously analyzes all available sequence data for the four taxa. However, one must also be careful in assuming that SM-ML will perform well outside of the anomaly zone because the anomaly zone has no obvious relationship to the problems encountered with concatenation. As with the case for the asymmetric tree, SMRT-ML tends to have a slow rate of convergence at extremely small branch lengths (Figures 6.4I,M,P). However, it is still not misleading and frequently returns either the ((AB)CD) or ((CD)AB) partially unresolved tree. Thus, although SMRT-ML can be slower to converge to the species tree for symmetric four-taxon trees, simulations for both symmetric and asymmetric four-taxon species trees suggest that SMRT-ML has the desirable property of not being misleading regardless of the species tree topology or branch lengths.

### 6.3.2 Five taxa

Five-taxon trees are illustrated in Figures 6.1C–E. For these trees, SM-ML is misleading with certain branch lengths (Figures 6.5A–C and 6.6D). In contrast, SMRT-ML is not misleading under any parameters tested, attaining the correct tree 100% of the time with 6000 genes for all topologies and branch lengths tested.

Similarly to the results presented for four taxa, in cases where both SM-ML and SMRT-ML recover the species tree (given enough loci), the method that has faster convergence depends on the topology and branch lengths of the species tree. For the species tree (((((AB)C)D)E), the only set of branch lengths tested for which SM-ML was not misleading was  $(w, x, y) = (1.0, 0.1, 0.1)$ , in which case SMRT-ML



converged more quickly to the species tree than SM-ML. For these branch lengths, SMRT-ML recovered the species tree 94% of the time with 1000 loci, whereas SM-ML recovered the species tree 84% of the time. For the species tree  $((((AB)(CD))E))$ , SM-ML showed slightly faster convergence to the species tree for two branch length combinations (Figures 6.7A,D). For example, with 1000 loci and branch lengths  $(w, x, y) = (0.1, 0.1, 0.1)$ , SM-ML and SMRT-ML recovered the species 93% and 91% of the time, respectively. However, for the same species tree topology with  $(w, x, y) = (0.1, 0.1, 1.0)$ , SMRT-ML appears to converge more quickly, with the species tree being estimated  $\sim 89\%$  of the time with SMRT-ML using 1000 loci versus  $\sim 60\%$  of the time with SM-ML. Furthermore, whereas SM-ML was never found to be misleading for four-taxon symmetric species trees, SM-ML can fail to converge to the species tree for every five-taxon tree shape. SMRT-ML converged to the species tree for all branch lengths tested on every five-taxon tree shape.

### 6.3.3 Model violations

To assess how SM-ML and SMRT-ML perform with violations of assumptions, we made gene trees non-clocklike by independently multiplying each branch by a value sampled from an exponential distribution with mean 1. The concatenated alignment generated by these gene trees was then analyzed assuming JC and a molecular clock.

Figure 6.8 shows that, for the  $((((AB)C)D))$  tree, both methods were fairly robust to violation of the molecular clock (when compared with Figure 6.3). The molecular clock violation slowed down the convergence to the species tree that was inferred with clocklike gene trees. For example, the species tree was inferred 98% of the time with 1000 genes under a molecular clock (Figure 6.3K), whereas it was inferred 80% of the time with 1000 genes and 96% of the time with 3000 genes when the molecular clock was violated (Figure 6.8K). This trend also held for the symmetric four-taxon species tree (Figure 6.9) and the three five-taxon species trees (Figures 6.10–6.12). Also, the

violation of the molecular clock affected SMRT-ML more than SM-ML. For example, when the molecular clock is violated, it may require 2000 genes instead of 1000 genes to obtain the same fraction of correctly inferred trees (compare Figure 6.5E with Figure 6.10E, Figure 6.6E with Figure 6.11E, and Figure 6.7E with Figure 6.12E). From these results, we conclude that the performance of the two methods is only slightly influenced by the molecular clock violation.

We introduced a second model violation by generating sequence alignments under a complex substitution model (General Time-Reversible (GTR)) and then comparing SM-ML and SMRT-ML when trees were inferred assuming a simple substitution model (JC). As with the case of the molecular clock violation, the general patterns displayed by the two methods were not significantly altered (Figures 6.13–6.17). However, under this substitution model violation, SM-ML was more negatively affected than SMRT-ML. Based on simulations, SM-ML can converge more quickly to the wrong tree (compare Figures 6.3A,B with Figures 6.13A,B and Figure 6.7B with Figure 6.17B) and more slowly to the correct tree (compare Figure 6.3C with Figure 6.13C and Figure 6.7C with Figure 6.17C) compared to analysis under the correct model. Furthermore, this model violation can reverse the effect of adding more data. For example, when both branches of the four-taxon species tree (((AB)C)D) had lengths of 0.1568, SM-ML was increasingly likely to infer the correct tree when there was no model misspecification (Fig. 6.3D; 63% probability with 6000 genes), but decreasingly likely under model misspecification (Figure 6.13D; 26.7% chance of inferring the matching tree with 6000 genes). However, this model violation can also favorably influence SM-ML by causing a faster convergence to the correct tree (compare Figure 6.4A with Figure 6.14A and Figure 6.5B with Figure 6.15B).

In simulations, neither SM-ML nor SMRT-ML performed uniformly better than the other method for all possible species trees. Table 6.1 gives a summary of these results and notes whether each method recovered the species tree in more than 50% of

simulations with 6000 loci of 500 nt each. A “NO” in the table indicates that either the method was likely to pick one of several trees (including the species tree) or converged to the wrong tree. Convergence to an incorrect tree only occurred for SM-ML. In cases where less than 50% probability of recovering the species tree was observed for SMRT-ML, SMRT-ML typically returned the species tree topology  $> 40\%$  of the time and frequently returned some other tree, often a partially unresolved tree with no false positive clades. We note that a “NO” only occurred for SMRT-ML in the four-taxon cases where there was one extremely short branch length of 0.01 coalescent units, leading to a high probability of a partially unresolved tree. SM-ML had poorer performance as the number of taxa was increased even though branch lengths were less extreme than for most of the four-taxon simulations. SMRT-ML, however, had similar performance as the number of taxa increased.

## 6.4 Results for yeast data

Although the causes of gene tree conflict in the yeast dataset analyzed by *Rokas et al.* (2003) are unknown, the analysis of this dataset by several groups (e.g., *Gatesy and Baker*, 2005; *Edwards et al.*, 2007) makes it useful for comparing methods of inferring species trees. *Rokas et al.* (2003) reported that 20 concatenated genes were sufficient for maximum parsimony or ML to infer the same tree with high reliability. On the estimated species tree, the five taxa with the most difficult relationships to infer form the five-taxon subtree (((*S. cerevisiae*, *S. paradoxus*), *S. mikatae*), *S. kudriavzevii*), *S. bayanus*).

Using SMRT-ML on all 106 genes, we recovered the species tree found using SM-ML on the full data (*i.e.*, the same tree that was reported as the estimated species tree in *Rokas et al.* (2003)). When a clock was assumed, SMRT-ML returned the species tree with the five-taxon subtree replaced by (((*S. cerevisiae*, *S. paradoxus*), *S. mikatae*), (*S. kudriavzevii*, *S. bayanus*)). The same result was produced by the

program BEST (*Liu, 2008*) analyzing the full data under a molecular clock; however, the molecular clock assumption is unreasonable because the data are not clocklike at most loci (*Edwards et al., 2007*).

To compare the efficiency of species tree estimation methods when methods agree on the full data, it is useful to consider subsets of the genes. For example, although *Rokas et al. (2003)* found that 20 randomly chosen genes were sufficient for SM-ML to estimate the species tree with high probability, *Edwards et al. (2007)* found that eight genes were sufficient using BEST. Because of the tradeoff between consistency and speed of convergence, we expect SMRT-ML to perform less efficiently than SM-ML for many cases when both methods have a high probability of returning the same tree, and this expectation is indeed what we found with the yeast data. The proportion of times SMRT-ML returned the species tree, inferred from all 106 loci, using random subsets of 20 loci was approximately 33%, with another 8% of cases returning a tree that was unresolved with respect to the taxa *S. kudriavzevii* and *S. bayanus* and the  $\{S. cerevisiae, S. paradoxus, S. mikatae\}$  clade. With 60 genes, the proportion of times that SMRT-ML returned the species tree increased to 59% (Figure 6.18). The SMRT-ML method was therefore increasingly likely to return the tree reported by *Rokas et al. (2003)* as the number of genes from this dataset was increased.

Using SMRT-ML on the full dataset of 106 genes, the bootstrap support for clades  $\{S. cerevisiae, S. paradoxus\}$  and  $\{S. cerevisiae, S. paradoxus, S. mikatae\}$  was 99% and 91%, respectively (as opposed to the 100% bootstrap support observed for the total concatenated dataset in *Rokas et al. (2003)*), while the clade  $\{S. cerevisiae, S. paradoxus, S. mikatae, S. kudriavzevii\}$  had 61% bootstrap support (Figure 6.19) (see Methods under the “Empirical example” section for how the bootstrap with SMRT-ML was performed). The clade  $\{S. bayanus, S. kudriavzevii\}$  occurred in 29% of bootstrap replicates. Thus, although SMRT-ML and SM-ML produced the same estimated species tree for the 106-gene yeast dataset, SMRT-ML converged to this

estimated tree more slowly than SM-ML. The speed of approach to this tree could be either a product of the tradeoff between consistency and speed of convergence sometimes observed for SMRT-ML, or misleadingly high bootstrap support for SM-ML (*Gadagkar et al.*, 2005; *Kubatko and Degnan*, 2007). The slower convergence of SMRT-ML compared to SM-ML observed for this dataset is not expected to generalize to all species trees since simulations found that there are also species trees for which SMRT-ML converges more quickly than SM-ML.

In both simulations and analysis of the yeast data, SMRT-ML was not misleading, in the sense of becoming increasingly less likely to infer an incorrect tree with more data, even in cases where SM-ML converged to the wrong tree. To see whether the observation that SMRT-ML was not misleading is expected to be true in general, we next assess the properties of SMRT-ML theoretically. We derive the probability that a site has pattern  $\mathbf{x}$  for a three-taxon species tree by averaging over gene genealogies under a simple substitution model. This result is then used to prove that SMRT-ML is statistically consistent when estimating species trees from coalescent mixtures of site patterns, at least in a simplified setting.

## 6.5 Theory

In this section, we begin by developing the probability distribution of site patterns under a Cavender-Farris-Neyman (CFN) substitution model given a clocklike three-taxon species tree. This substitution model assumes binary characters with equal rates of mutation between the characters. Assuming that incomplete lineage sorting is the source of discordance between gene trees and species trees and that the species tree has no hybridization or horizontal gene transfer events, we then show that the frequency of a certain site pattern in a concatenated alignment converges in probability to the probability of the site pattern ((Lemma VI.2). From this result, we provide a proof that SM-ML is a consistent estimator of a clocklike three-taxon

species tree (Lemma VI.3). Utilizing Lemma VI.3, we show in Theorem VI.4 that SMRT-ML is consistent for estimating clocklike species trees under the CFN model.

Consider a species tree with three taxa. Denote the true species tree by  $\sigma$  with speciation times  $\rho_0$  and  $\rho_1$  (see Figure 6.20). Denote the topology of the species tree as ((AB)C). The species tree is therefore written as

$$\sigma = ((A:\rho_1, B:\rho_1):\rho_0 - \rho_1, C:\rho_0),$$

which has clocklike branch lengths. Further, denote the topology of the gene tree that matches the species tree  $\sigma$  as  $\tau_1$  and denote the other gene tree topologies as the star tree  $\tau_0 = (ABC)$  and the two discordant trees  $\tau_2 = ((AC)B)$  and  $\tau_3 = ((BC)A)$ .

Random gene trees evolving along the species tree  $\sigma$  can take on any of the topologies  $\tau_1$ ,  $\tau_2$ , or  $\tau_3$ . Define  $\theta$  as the population mutation rate for each branch of the tree. For a random gene tree topology, we define  $t$  as the total length of the gene tree and  $u$  as the time from the present to the most recent coalescent event in mutation units.

Our goal is to determine the probability of a site pattern  $\mathbf{x} = (x_1, x_2, x_3)$  under a CFN substitution model, where  $x_1$ ,  $x_2$ , and  $x_3$  are the characters at a site for species A, B, and C, respectively. If two species have the same character at a site, then they share the same letter. Therefore, the possible site patterns are  $xxx$ ,  $xyx$ ,  $xyx$ , and  $yxx$ . We note that only the  $xyx$  pattern supports the matching gene tree ((AB)C). We will show that when data are concatenated under a coalescent model and a ML tree is inferred from the concatenated data, the probability that the ML tree has topology  $\tau_1$  is higher than the probability of any other bifurcating tree topology. Further, we will show that this probability approaches 1 as the number of sites approaches infinity.

The probability of a site pattern given the species tree is obtained by conditioning on the gene genealogy and integrating over the joint density of the two coalescent

times. The form of the joint density depends on whether both coalescent events occur more anciently than the root of the species tree  $\sigma$  or whether one coalescent event occurs more recently than the root of the species tree. This latter case only occurs when the gene tree matches the species tree. In this case (Figure 6.20A), define  $g_\sigma(t, u, \tau_1)$  as the joint density for the coalescent times and gene tree topology  $\tau_1$ . Following *Rannala and Yang* (2003), this joint density is written as

$$g_\sigma(t, u, \tau_1) = \frac{4}{\theta^2} e^{\frac{2}{\theta}(\rho_0 + \rho_1 - t - u)}. \quad (6.1)$$

When all coalescent events occur more ancient than the root and the genealogy has topology  $\tau_i$  (Figures 6.20B–D), the joint density of coalescent times and topology is

$$f_\sigma(t, u, \tau_i) = \frac{4}{\theta^2} e^{\frac{2}{\theta}(2\rho_0 + \rho_1 - t - 2u)}. \quad (6.2)$$

Note that when all coalescent events occur above the root, the form of the joint density for the gene tree topology and two coalescent times is the same for each of the three topologies. The probability of site pattern  $\mathbf{x}$  given that the species tree is  $\sigma$  is

$$P_\sigma(\mathbf{x}) = \int_{\rho_0}^{\infty} \int_{\rho_1}^{\rho_0} P_\sigma(\mathbf{x} | t, u, \tau_1) g_\sigma(t, u, \tau_1) du dt + \sum_{i=1}^3 \int_{\rho_0}^{\infty} \int_{\rho_0}^t P_\sigma(\mathbf{x} | t, u, \tau_i) f_\sigma(t, u, \tau_i) du dt, \quad (6.3)$$

where  $P_\sigma(\mathbf{x} | t, u, \tau_i)$  is the probability of the site pattern given the gene genealogy with topology  $\tau_i$  and the branch lengths  $u$  and  $t - u$ . The first term in the probability is for the case that the gene tree matches the species tree and there is a coalescence between the A and B lineages more recent than the root of the species tree (Figure 6.20A). The second term is a summation corresponding to the three possible gene tree topologies

when all coalescent events are more ancient than the root (Figures 6.20B–D).

For a CFN substitution model, *Yang* (2000) provided the probabilities of the site pattern  $\mathbf{x}$  conditional on the gene tree topology with branch length  $t$  and  $u$  as

$$\begin{aligned} P(xxx | t, u, \tau_i) &= \frac{1}{4} + \frac{1}{4}e^{-4u} + \frac{1}{2}e^{-4t} \\ P(xxy | t, u, \tau_1) &= \frac{1}{4} + \frac{1}{4}e^{-4u} - \frac{1}{2}e^{-4t} \\ P(xyx | t, u, \tau_1) &= P(yxx | t, u, \tau_1) = \frac{1}{4} - \frac{1}{4}e^{-4u}, \end{aligned} \tag{6.4}$$

where the equality for  $xyx$  and  $yxx$  follows by symmetry of A and B with respect to C in tree  $\tau_1$ . We have dropped the subscript  $\sigma$  in  $P_\sigma(\cdot | \cdot)$  because the probability of a site pattern is independent of the species tree given the gene genealogy  $\tau_i$ . Similarly,

$$\begin{aligned} P(xxy | t, u, \tau_3) &= P(xxy | t, u, \tau_2) = P(xyx | t, u, \tau_1) \\ P(yxx | t, u, \tau_3) &= P(xyx | t, u, \tau_2) = P(xxy | t, u, \tau_1) \\ P(xyx | t, u, \tau_3) &= P(yxx | t, u, \tau_2) = P(yxx | t, u, \tau_1). \end{aligned} \tag{6.5}$$

We next derive the full distribution of site patterns for a given species tree  $\sigma$ .



Using the symmetries in equation (6.5),

$$\begin{aligned}
P_\sigma(xxy) &= \int_{\rho_0}^{\infty} \int_{\rho_1}^{\rho_0} P(xxy | t, u, \tau_1) g_\sigma(t, u, \tau_1) du dt \\
&\quad + \int_{\rho_0}^{\infty} \int_{\rho_0}^t \left\{ [P(xxy | t, u, \tau_1) + 2P(xyx | t, u, \tau_1)] \right. \\
&\quad \quad \quad \left. \times f_\sigma(t, u, \tau_i) \right\} du dt \\
&= \int_{\rho_0}^{\infty} \int_{\rho_1}^{\rho_0} \frac{1 + e^{-4u} - 2e^{-4t}}{4} \frac{4}{\theta^2} e^{\frac{2}{\theta}(\rho_0 + \rho_1 - t - u)} du dt \\
&\quad + \int_{\rho_0}^{\infty} \int_{\rho_0}^t \frac{3 - e^{-4u} - 2e^{-4t}}{4} \frac{4}{\theta^2} e^{\frac{2}{\theta}(2\rho_0 + \rho_1 - 2u - t)} du dt \\
&= \frac{1 + 2\theta + e^{-4\rho_1} - 2e^{-4\rho_0}}{4 + 8\theta}. \tag{6.6}
\end{aligned}$$

Analogously,

$$\begin{aligned}
P_\sigma(xyx) &= \int_{\rho_0}^{\infty} \int_{\rho_1}^{\rho_0} \frac{1 - e^{-4u}}{4} \frac{4}{\theta^2} e^{\frac{2}{\theta}(\rho_0 + \rho_1 - t - u)} du dt \\
&\quad + \int_{\rho_0}^{\infty} \int_{\rho_0}^t \frac{3 - e^{-4u} - 2e^{-4t}}{4} \frac{4}{\theta^2} e^{\frac{2}{\theta}(2\rho_0 + \rho_1 - 2u - t)} du dt \\
&= \frac{1 + 2\theta - e^{-4\rho_1}}{4 + 8\theta}. \tag{6.7}
\end{aligned}$$

By symmetry we have that  $P_\sigma(yxx) = P_\sigma(xyx)$  and by the law of total probability,

$$\begin{aligned}
P_\sigma(xxx) &= 1 - P_\sigma(xxy) - P_\sigma(xyx) - P_\sigma(yxx) \\
&= \frac{1 + 2\theta + e^{-4\rho_1} + 2e^{-4\rho_0}}{4 + 8\theta}. \tag{6.8}
\end{aligned}$$

The probability in equation (6.6) is greater than the probability in equation (6.7) if and only if  $\rho_0 > \rho_1$ , *i.e.*, the root of the species tree is more ancient than the divergence

of species A and B. Therefore, the probabilities of the segregating site patterns given the species tree  $\sigma$  are related by  $P_\sigma(xxy) > P_\sigma(yxx) = P_\sigma(xy x)$ . Hence, the most probable segregating site pattern is the pattern that supports the species tree.

It is possible to extend the above derivation to other substitution models by modifying the expression in equation (6.4) and including a term for  $P(xyz)$  if there are more than two possible character states. Extending to site pattern probabilities for four or more taxa is also accomplished using the same approach but is considerably more tedious. For example, with four taxa, there are 15 rooted gene tree topologies rather than three, and the form of the joint density of coalescent times and gene tree topology depends on the *coalescent history*, a list of ancestral populations from the species tree where each coalescence occurs (*Degnan and Salter, 2005; Rosenberg, 2007*). For four-taxon trees, there are up to five coalescent histories for a given gene tree in a species tree, in contrast to the two expressions for three taxa (eqs. (6.1) and (6.2)). Thus, the probability of a site pattern  $\mathbf{x}$  is found by summing over gene trees and computing triple integrals of  $P(\mathbf{x})$  with respect to each of the algebraic expressions taken by the joint densities of coalescent times and gene tree topologies. Because SMRT-ML only uses alignments of three taxa, we have only derived three-taxon site pattern probabilities. We next provide two lemmas which aid in the proof of the theorem that SMRT-ML is consistent.

Lemma VI.2 essentially says that the alignment lengths do not matter asymptotically (under reasonable conditions), because the proportion of sites with any given pattern  $\mathbf{x}$  will approach the probability of the site pattern. In practice the length of the alignments could affect the rate at which a method using concatenated data (e.g., SM-ML or SMRT-ML) converges to a particular species tree. Lemma VI.3 says that because the most likely segregating pattern supports the species tree, SM-ML is consistent on concatenated three-taxon alignments under some assumptions (e.g., a clocklike species tree, constant ancestral  $\theta$ s, and the CFN substitution

model). Theorem VI.4 puts these ideas together and states that, because SMRT-ML constructs the species tree from several SM-ML estimates restricted to rooted triples, SMRT-ML is a statistically consistent estimator of clocklike species trees under the CFN model.

We begin by stating assumptions used for proving the lemmas and theorem that follow:

1. Let the gene tree for the  $i$ th locus have topology  $\tau^{(i)} \in \{\tau_1, \tau_2, \tau_3\}$  and coalescent times  $u_i$  and  $t_i$  (Figure 6.20), where the joint distribution of topology and coalescent times is given by equations (6.1) and (6.2). Assume that each site  $j$  in locus  $i$  is independent given the gene tree and coalescent times and has site pattern probability  $P(\mathbf{x} | t_i, u_i, \tau^{(i)})$ , given by equations (6.6)–(6.8), where the mutation parameter  $\theta$  is constant for each ancestral population in the species tree. This derivation for site pattern probabilities depends on the following assumptions:

- Mutations occur under the CFN substitution model.
- The species tree is clocklike.
- Incomplete lineage sorting is the source of discordance between gene trees and species trees.
- There is no hybridization, horizontal gene transfer, or other gene flow between species.
- There is no population subdivision within species.

2. Consider a concatenated alignment of  $m$  non-recombining loci that are conditionally independent given the species tree, each with finite length  $L_i \geq 1$  for  $i = 1, 2, \dots, m$ . Define  $q_m = (\sum_{i=1}^m L_i^2) / (\sum_{i=1}^m L_i)^2$  and assume that, for any site pattern  $\mathbf{x}$ ,  $q_m \rightarrow 0$  as  $m \rightarrow \infty$ .

3. A supertree algorithm is used with the property that if the input trees are compatible, then the supertree is a rooted phylogenetic tree which displays all input trees.

The condition under assumption 2 that  $q_m \rightarrow 0$  as  $m \rightarrow \infty$  allows a version of the Law of Large Numbers to be applied to site pattern probabilities for concatenated alignments with different lengths and ensures that the length of the concatenated alignment does not grow too rapidly. For example, if we concatenate loci of constant length  $L$ , then  $q_m = mL^2/(mL)^2 \rightarrow 0$  as  $m \rightarrow \infty$ . Similarly, if the gene length is bounded, so that  $1 \leq L_i \leq B$ , for some upper bound  $B$ , then  $q_m \leq mB^2/m^2 \rightarrow 0$ . Since real genomes are finite, this assumption is reasonable for biological data. However, if every new locus were twice the length of the previous locus, say  $L_i = 2^i$  for  $i = 1, 2, \dots, m$ , then  $q_m \rightarrow 1/3$  as  $m \rightarrow \infty$ . Thus, if the concatenated alignment grows too quickly, Lemma VI.2 does not apply.

Assumption 3 states that the only characteristic of the supertree method that is necessary to prove Theorem VI.4 is that the method must return a tree which displays all input trees when they are compatible. Hence, if all rooted triples are inferred correctly, then the tree that displays those rooted triples is the species tree topology. A broad class of supertree algorithms can be used to prove this result including BUILD (*Aho et al.*, 1981), matrix representation using parsimony (*Baum*, 1992; *Ragan*, 1992), mincut (*Semple and Steel*, 2000), MMC (*Page*, 2002), matrix representation using flipping (*Chen et al.*, 2003), and normalized triplet supertrees (*Willson*, 2009).

Lemma VI.1 is a version of the Weak Law of Large Numbers that does not require identically distributed random variables. This lemma is used to prove Lemma VI.2.

**Lemma VI.1** (Modified Theorem 5.2.3 of *Chung* (1974)). *Consider the sequence  $X_1, X_2, \dots, X_n$ , where  $X_i > 0$ , of independent random variables each with their own distribution function. Define  $S_n = \sum_{i=1}^n X_i$ . Further, let  $\{b_n\}$  be a sequence*

that approaches infinity and assume that  $X_i \leq b_n$  for each  $i = 1, 2, \dots, n$ . If  $\lim_{n \rightarrow \infty} \sum_{i=1}^n E[X_i^2]/b_n^2 = 0$ , then  $S_n/b_n \xrightarrow{P} E[S_n/b_n]$ .

**Lemma VI.2.** *Under assumptions 1 and 2, the proportion of sites with pattern  $\mathbf{x}$  converges in probability to  $P(\mathbf{x})$ .*

*Proof.* First we show that the expected proportion of sites with a given site pattern  $\mathbf{x}$  is equal to the probability of that site pattern,  $P(\mathbf{x})$ . We consider the expected proportion of sites with each pattern and note that by Lemma VI.1, as the number of loci approaches infinity, the probability approaches 1 that the proportion of sites with a given pattern approaches the expected proportion. Let  $m$  denote the number of loci and let  $L_i$  denote the number of sites at locus  $i$ . The total number of sites is  $\sum_{i=1}^m L_i$ . For a site pattern  $\mathbf{x}$ , let  $\delta_{\mathbf{x},i,j} = 1$  if site  $j$  in locus  $i$  has site pattern  $\mathbf{x}$ ; otherwise  $\delta_{\mathbf{x},i,j} = 0$ . Let  $M_{\mathbf{x},i} = \sum_{j=1}^{L_i} \delta_{\mathbf{x},i,j}$  denote the number of sites in locus  $i$  that have site pattern  $\mathbf{x}$ . Let  $S_m = \sum_{i=1}^m M_{\mathbf{x},i}$  and let  $b_m = \sum_{i=1}^m L_i$ . Because the length of the concatenated alignment is increasing with each additional locus, we have that  $b_m \rightarrow \infty$  as  $m \rightarrow \infty$ . Note that  $E[\delta_{\mathbf{x},i,j}^2] = P(\mathbf{x})$ . Also note that  $E[\delta_{\mathbf{x},i,j} \delta_{\mathbf{x},i,k}] = P(\mathbf{x}, \mathbf{x})$  for  $j \neq k$  where  $P(\mathbf{x}, \mathbf{x})$  is the probability of getting pattern  $\mathbf{x}$  at two different sites. Note that we do not need to know the actual value of  $P(\mathbf{x}, \mathbf{x})$ —only that it is between 0 and 1. Then it follows that

$$\begin{aligned} \frac{1}{b_m^2} \sum_{i=1}^m E[M_{\mathbf{x},i}^2] &= \frac{1}{b_m^2} \sum_{i=1}^m \left( \sum_{j=1}^{L_i} E[\delta_{\mathbf{x},i,j}^2] + 2 \sum_{j=1}^{L_i-1} \sum_{k=j+1}^{L_i} E[\delta_{\mathbf{x},i,j} \delta_{\mathbf{x},i,k}] \right) \\ &= \frac{1}{b_m^2} \sum_{i=1}^m [P(\mathbf{x})L_i + P(\mathbf{x}, \mathbf{x})(L_i^2 - L_i)] \\ &= \frac{P(\mathbf{x}) - P(\mathbf{x}, \mathbf{x})}{b_m} + P(\mathbf{x}, \mathbf{x}) \frac{\sum_{i=1}^m L_i^2}{b_m^2}. \end{aligned} \tag{6.9}$$

The quantity in equation (6.9) approaches 0 as  $m \rightarrow \infty$  only if  $q_m = \sum_{i=1}^m L_i^2/b_m^2 \rightarrow 0$  as  $m \rightarrow \infty$ . We assumed that  $q_m \rightarrow 0$  as  $m \rightarrow \infty$ . Thus, Lemma VI.1 applies and

therefore,

$$\begin{aligned}
\frac{\sum_{i=1}^m M_{\mathbf{x},i}}{\sum_{i=1}^m L_i} &= \frac{S_m}{b_m} \xrightarrow{P} E \left[ \frac{\sum_{i=1}^m M_{\mathbf{x},i}}{b_m} \right] \\
&= \frac{\sum_{i=1}^m \sum_{j=1}^{L_i} P(\mathbf{x})}{\sum_{i=1}^m L_i} \\
&= \frac{\sum_{i=1}^m L_i}{\sum_{i=1}^m L_i} P(\mathbf{x}) = P(\mathbf{x}). \quad \square
\end{aligned}$$

**Lemma VI.3.** *Under assumptions 1 and 2, SM-ML is a statistically consistent estimator of a three-taxon clocklike species tree.*

*Proof.* Let  $m$  denote the number of loci. The total number of sites is  $\sum_{i=1}^m L_i$ . Let  $M_{\mathbf{x},i}$  denote the number of sites in locus  $i$  that have site pattern  $\mathbf{x}$ . Further, let  $M_{\mathbf{x}} = \sum_{i=1}^m M_{\mathbf{x},i}$  be the number of sites with pattern  $\mathbf{x}$  in the concatenated alignment. Suppose three species, A, B, and C, have the species tree  $\sigma = ((A:\rho_1, B:\rho_1):\rho_0 - \rho_1, C:\rho_0)$ , where  $\rho_0$  and  $\rho_1$  are measured in coalescent units, and the two ancestral populations each have the same  $\theta$ . Further, suppose there are  $M_{xxx}$ ,  $M_{xxy}$ ,  $M_{xyx}$ , and  $M_{yxx}$  sites with site patterns  $xxx$ ,  $xxy$ ,  $xyx$ , and  $yxx$ , respectively. By Lemma VI.2, we know that the relative frequency of pattern  $xxy$  ( $M_{xxy}/\sum_{i=1}^m L_i$ ) converges in probability to  $P(xxy)$ . Because  $xxy$  is the most likely segregating site pattern (eqs. (6.6)–(6.8)), it follows that the probability that  $xxy$  is the most frequently occurring segregating site pattern (*i.e.*,  $M_{xxy} > M_{xyx}, M_{yxx}$ ) approaches 1 as  $m \rightarrow \infty$ . Theorem 3 of *Chor et al.* (2007) states that if  $M_{xxy} > M_{xyx}, M_{yxx}$ , then ((AB)C) is the inferred ML under a molecular clock. Utilizing this theorem, the probability that the ML tree topology is ((AB)C) approaches 1 as  $m \rightarrow \infty$ .  $\square$

**Theorem VI.4.** *Under assumptions 1–3, SMRT-ML is a statistically consistent estimator of a clocklike species tree with three or more taxa.*

*Proof.* Suppose we have an  $n$ -taxon species tree. There are  $\binom{n}{3}$  subsets of three taxa. Let the rooted triples on the species tree be enumerated  $\sigma_1, \sigma_2, \dots, \sigma_J$ , where  $J = \binom{n}{3}$ .

Let  $\sigma_j^*$  denote a rooted triple defined on the same taxa as  $\sigma_j$  but which is not a rooted triple on the species tree. From equations (6.6) and (6.7) and from equation (6.5), if  $\mathbf{x}$  is the most probable segregating site pattern for  $\sigma_j$ , then  $P_{\sigma_j}(\mathbf{x}) > P_{\sigma_j^*}(\mathbf{x})$ . Let the most frequently occurring segregating site pattern for supermatrix rooted triple  $j$  be  $\mathbf{x}$ . Applying Lemma VI.2, for any  $\varepsilon > 0$ , we can choose the number of loci  $m$  such that the probability of  $P_{\sigma_j}(\mathbf{x}) > P_{\sigma_j^*}(\mathbf{x})$  is greater than  $1 - \varepsilon/\binom{n}{3}$ . By Lemma VI.3, the ML estimate for each of these  $J$  sets of three taxa is  $\sigma_j$ ,  $j = 1, 2, \dots, J$ . Therefore the probability that all  $J$  rooted triples in the species tree are inferred by SMRT-ML is greater than  $1 - \varepsilon$ . In this case, the rooted triples will be compatible and the tree with the same topology as the species tree is uniquely identified by these  $J$  triples by Proposition 4 of *Steel* (1992). Applying a supertree algorithm to these  $J$  rooted triples with the property that if the input trees are compatible, then the supertree method returns a tree that displays all of its input trees, the supertree algorithm is guaranteed to return the matching species tree with probability greater than  $1 - \varepsilon$ . Thus, the supertree method applied to the  $J$  supermatrix rooted triples returns the species tree with probability greater than  $1 - \varepsilon$ . Therefore, SMRT-ML is statistically consistent under the CFN substitution model when the species tree is clocklike.  $\square$

## 6.6 Discussion

### 6.6.1 Overview of results and implications

In this study, we have shown that combining concatenation and supertree methods on rooted triples can overcome the problems caused by incomplete lineage sorting for concatenation-based ML inference of species trees. From theory, we find that SMRT-ML is a consistent estimator of species trees when sequences are generated under a CFN substitution model assuming a molecular clock and equal values of  $\theta$  over the species tree.

Although neither SM-ML nor SMRT-ML performs uniformly better than the other, a scan of Table 6.1 shows that SMRT-ML often outperforms SM-ML when no single gene tree has high probability (typically  $\leq 25\%$ ), and when two gene trees have very similar probabilities. Because the yeast dataset has considerably less gene discordance than these cases, it is not surprising that SM-ML needs fewer loci than SMRT-ML to obtain the same species tree that was inferred from all 106 loci. The yeast data analysis also suggests that it may take a large number of genes for SMRT-ML to have a high probability of recovering the species tree, and therefore that SMRT-ML may have an advantage with sizeable genomic datasets. Simulations show that large amounts of data may also be necessary to resolve phylogenies when no single gene tree topology predominates.

Through simulations, we find that SMRT-ML is not misleading and often outperforms SM-ML given sufficiently severe gene tree discordance when sequences are generated under JC and GTR substitution models. This finding suggests that SMRT-ML is consistent when assuming models that are more complex than CFN. However, analytical results for three-taxon trees under more complex models are difficult to obtain. For example, *Chor et al.* (2006) found that the exact ML solution for a rooted three-taxon Jukes-Cantor problem required finding roots of an 11th degree polynomial.

An attractive property of the SMRT method is computational efficiency. For each rooted triple, the tree space contains only three trees and the number of branch lengths needed to optimize is small. Therefore, the total number of trees examined is  $3\binom{n}{3} = n(n-1)(n-2)/2$ . In contrast, the total number of rooted tree topologies in an  $n$ -taxon tree space is  $(2n-3)!!$ . Although there are methods, such as Branch and Bound (*Felsenstein*, 2004), that can ignore the irrelevant part of the tree space, finding globally optimal trees under criteria such as likelihood or parsimony is NP-hard (*Day et al.*, 1986; *Chor and Tuller*, 2005; *Roch*, 2006). Because MMC is a polynomial-time



algorithm (*Page, 2002*), and only a polynomial number of trees is evaluated using SMRT, both steps of inferring triples and constructing the tree are polynomial in the number of taxa. Thus, at least under a simple substitution model, SMRT-ML is a polynomial-time algorithm for inferring the species tree and is statistically consistent when gene tree discordance is described by the multispecies coalescent model.

### 6.6.2 Taxon sampling for species tree inference

An issue that has received a lot of attention in phylogenetics is whether increased taxon sampling can improve the accuracy of species tree inference. Some researchers argue that increased taxon sampling generally improves phylogenetic inference (*Zwickl and Hillis, 2002; Hedtke et al., 2006*), and others argue that it often does not (*Poe and Swofford, 1999; Rosenberg and Kumar, 2001; Rokas and Carroll, 2005*). These studies have all focused on the effect of taxon sampling on the estimation of gene trees, prompting the need for investigating its effects on species tree estimation (*Degnan and Rosenberg, 2009*).

Some of our results imply that the performance of SM-ML can either be improved or impaired when extra taxa are sampled, depending on the branch lengths and topology of the species tree. In general, SMRT-ML is less sensitive to taxon sampling than SM-ML for the range of species trees examined. As an example where SM-ML performs worse with more taxa, consider the species trees  $((AB)(CD))$  with branch lengths  $(x, y) = (0.1, 1.0)$  (Figure 6.4C) and  $((((AB)(CD))E))$  with branch lengths  $(w, x, y) = (0.1, 0.1, 1.0)$  (Figure 6.7C). For the four-taxon species tree, SM-ML recovered the species tree topology  $\sim 99\%$  of the time with 1000 loci. The addition of the E taxon with a short branch length separating the root of the tree from the most recent common ancestor of A, B, C, and D impaired the performance of SM-ML, making it incorrectly group E with (AB) 38% of the time with 1000 loci. In contrast, adding the E taxon to the same four-taxon tree had a much smaller influence on the

performance of SMRT-ML (compare Figure 6.4K with Figure 6.7G).

To investigate this effect further, we added a sixth taxon separated from the root of the tree  $((AB)(CD))E$  by 0.1 coalescent units to create the species tree  $(((((AB)(CD))E)F)$  (Figure 6.21A). Adding the sixth taxon caused the probability that SM-ML inferred the AGT  $(((((AB)(CD))(EF)))$  to approach 1 as more genes were added (Figure 6.21B). On the other hand, SMRT-ML had a similar performance with this six-taxon tree on taxa A–F as with the four- and five-taxon subtrees on taxa A–D and A–E, respectively.

For the five-taxon species tree  $((((AB)C)D)E)$  with branch lengths  $(w, x, y) = (1.0, 0.1, 0.1)$  (Figure 6.5D), SM-ML recovered the species tree 100% of the time given enough loci. However, when taxon E was removed from this species tree, SM-ML was misleading on the subtree  $((AB)C)D$  with branch lengths  $(x, y) = (0.1, 0.1)$  (Figure 6.22), with a probability approaching 1 of returning the AGT  $((AB)(CD))$ . SMRT-ML was not as influenced by the presence of taxon E for this example, though the extra taxon slightly hindered the speed of convergence to the species tree. This example shows not only that increased taxon sampling had a less dramatic influence on SMRT-ML than SM-ML, but also that the same parameters can produce opposite effects that aid one method while hurting the other.

### 6.6.3 Rooted triple consensus

A recent study used rooted triples estimated at each locus as input to the quartet puzzling algorithm (Ewing *et al.*, 2008) by treating a fourth taxon as a known outgroup. In quartet puzzling, maximum likelihood trees for all  $\binom{n}{4}$  quartets of a set of  $n$  species are estimated and a heuristic algorithm is used to construct the tree from the inferred quartets (Strimmer and von Haeseler, 1996). The  $R^*$  consensus method (Bryant, 2003; Degnan *et al.*, 2009) is similar in that it uses rooted triples at each locus, although these are generated by first inferring gene trees on the full

set of taxa.  $R^*$  consensus then applies a different non-heuristic algorithm from that of the quartet puzzling based rooted triple consensus to construct the tree from the estimated rooted triples. Like  $R^*$  consensus, rooted triple consensus construct the estimated species tree from  $m\binom{n}{3}$  rooted triples, where  $m$  is the number of loci. Neither method requires the estimation of coalescent or population parameters, and each avoids the problem of AGTs due to incomplete lineage sorting through the use of rooted triples.  $R^*$  consensus given known gene trees at each locus is proven to be statistically consistent when gene tree discordance is due to incomplete lineage sorting (Degnan *et al.*, 2009). A more general approach shows that supertree methods that have rooted triples as input can be statistically consistent in this setting given certain covering conditions and bounds on error in gene tree estimation (Steel and Rodrigo, 2008, Proposition 5). SMRT is different from rooted triple consensus methods in two respects: (1) only the  $\binom{n}{3}$  rooted triples from a supermatrix are inferred, and (2) the resulting triples are input into a supertree algorithm to construct the estimated species tree.

One advantage of rooted triple consensus and  $R^*$  over SMRT is that they use the information of all available taxa at a given locus to infer a gene tree whereas SMRT only uses information on three taxa. Because there may be a lack of phylogenetic signal among the three taxa analyzed by SMRT, the extra information about the relationship between taxa used by rooted triple consensus and  $R^*$  can aid in more accurate estimates of species tree when the total amount of sequence is small. However, SMRT has the advantage that it is both fast and tractable on a large number of taxa. Because ML inference of phylogenetic trees is NP-hard (Chor and Tuller, 2005; Roch, 2006), if gene trees are inferred using ML at each locus, then both rooted triple consensus and  $R^*$  are NP-hard whereas SMRT is polynomial in the number of taxa.

#### 6.6.4 Bayesian approaches

Recent methods, such as BEST (*Liu and Pearl, 2007; Liu, 2008*) and BUCKY (*Ané et al., 2007*), for inferring species trees from multilocus data take a Bayesian approach. The program BEST simultaneously estimates a joint posterior distribution of gene trees and species trees (*Rannala and Yang, 2008*) assuming that gene trees are distributed according to the coalescent process and that gene tree discordance is due solely to incomplete lineage sorting. In contrast to BEST, which models discordance among of gene trees using the coalescent process, BUCKY uses a prior to model the correlation between gene trees without assuming the source (e.g., incomplete lineage sorting) of discordance. These methods are attractive in that they are designed to handle gene-tree discordance. However, both are computationally intensive, relying on MCMC runs for separate loci and for estimates of the species tree and are therefore tractable only for small numbers of taxa and loci (*Edwards, 2009*). Because SMRT is polynomial in the number of taxa and not heavily affected by the number of loci, it is especially well-suited for genomic-level data and large numbers of taxa.

#### 6.6.5 Other sources of discordance

SMRT gains its strength from the fact that when gene trees are distributed according to the multispecies coalescent, there are no anomalous three-taxon trees when the source of gene-tree discordance is due only to incomplete lineage sorting. However, in the presence of other sources of discordance, such as hybridization, horizontal gene transfer, gene duplication, recombination, and population structure the most probable three-taxon gene tree might not match the species tree (*Slatkin and Pollack, 2008*). Hence, if there are forces acting strongly to create gene-tree discordance other than incomplete lineage sorting, then SMRT may not have enough information to obtain the correct tree. However, because SMRT has the ability to infer partially unresolved trees, then it may be the case that forces such as horizontal

gene transfer will cause SMRT to infer a partially unresolved tree. Future studies are needed to assess how SMRT and other methods perform under various types and degrees of gene tree discordance.

### 6.6.6 Summary

When genetic data from multiple loci are concatenated, the distribution of site patterns is a mixture that depends on the distribution of gene trees over the loci. Such mixture distributions on site patterns make it difficult to obtain analytical results for concatenated data and therefore to understand theoretical properties of phylogenetic methods that use concatenated data. We have obtained the distribution of site patterns for three-taxon concatenated sequences under a mixture distribution due to the multispecies coalescent using the CFN substitution model. Thus, despite the poor performance of SM-ML for some species trees, there is enough information in the concatenated alignment, and therefore in the distribution of site patterns, to recover the species tree topology. SMRT-ML uses this information in the concatenated alignment to consistently recover the species tree.

The consistency of SMRT-ML shows that the species tree topology is identifiable from concatenated data in the sense that two distinct species trees (with either different topologies or the same topology but different branch lengths) cannot have the same distribution of site patterns. The analytic framework in this paper could be extended to either more complex substitution models or to larger numbers of taxa to yield further insights into some of the properties of concatenated data.

As a tool for inferring species trees, SMRT-ML could be extended to cases where there are multiple individuals sampled per species. Here, there could be multiple inferred triples for each choice of three species, where one individual from within each of the three species is chosen randomly, or all possible combinations with one individual per species are used. If there are  $n$  species and  $i$  individuals sampled per

species, this procedure would result in  $i^3 \binom{n}{3}$  inferred rooted triples from which the species tree could be constructed using a supertree method such as MMC. With multiple rooted triples estimated on the same choice of three taxa, a supertree algorithm designed for high levels of conflict in the input triples might be useful, for example, Normalized Triplet Supertree (*Willson, 2009*).

We have not investigated the performance of SMRT when combined with methods of inferring gene trees other than ML, such as parsimony and distance methods. *Liu and Edwards* (2009) show that for concatenated data, under similar assumptions as in this paper, distance methods and in many cases parsimony methods recover the species tree when SM-ML is misleading. Although, because of long branch attraction (*Felsenstein, 1978*), maximum parsimony is not consistent for trees with five or more taxa, even when there is a molecular clock (*Hendy and Penny, 1989*). However, for cases in which rooted three-taxon gene trees can be inferred consistently from concatenated data—including distance and parsimony methods under a molecular clock—SMRT is also consistent for larger trees because of the fact that rooted triples identify a tree, independently of how those rooted triples were inferred. Future studies using simulation and real data will be needed to further assess the performance of SMRT methods and its extensions.

## 6.7 Acknowledgments

We thank Noah A. Rosenberg, Raquel Assis, Elizabeth S. Allman, Liang Liu, and two anonymous reviewers for their valuable comments. This work was supported by National Science Foundation grant DEB-0716904, National Institute of Health training grant T32 GM070449, and the new Zealand Marsden Fund.

Table 6.1: Probabilities of concordant and most probable discordant gene trees and performance of SM-ML and SMRT-ML with 6000 loci

Species tree $\sigma$	Branch lengths ( $x, y$ ) or ( $w, x, y$ ) (see Figure 6.1)	Highest-prob. nonmatching gene tree	Concordance probability	Highest-prob. nonmatching tree	SM-ML correct > 50%	SMRT-ML correct > 50%	Figures
(((AB)C)D)	(0.01, 2.0)	((AB)(CD))	0.30170	0.30039	NO	NO	6.3A,I
	(0.05, 1.0)	((AB)(CD))	0.25483	0.24116	NO	YES	6.3B,J
	(0.1, 1.0)	((AB)(CD))	0.27762	0.23099	YES	YES	6.3C,K
	(0.1568, 0.1568)	((AB)(CD))	0.13344	0.13349*	YES	YES	6.3D,L
	(0.01, 1.0)	((AB)(CD))	0.23595	0.24948*	NO	NO	6.3E,M
	(0.05, 0.05)	((AB)(CD))	0.07879	0.12079*	NO	YES	6.3F,N
	(0.1, 0.05)	((AB)(CD))	0.08867	0.11901*	NO	YES	6.3G,O
	(0.25, 0.01)	((AB)(CD))	0.10376	0.10511*	YES	YES	6.3H,P
	(0.1, 0.1)	((AB)(CD))	0.10370	0.12792*	NO	YES	6.22B
	(0.01, 2.0)	(((AB)C)D), (((AB)D)C)	0.30929	0.29280	YES	NO	6.4A,I
((AB)(CD))	(0.05, 1.0)	(((AB)C)D), (((AB)D)C)	0.27612	0.21987	YES	YES	6.4B,J
	(0.1, 1.0)	(((AB)C)D), (((AB)D)C)	0.29946	0.20915	YES	YES	6.4C,K
	(0.1568, 0.1568)	(((AB)C)D), (((AB)D)C)	0.18497	0.08196	YES	YES	6.4D,L
	(0.01, 1.0)	(((AB)D)C), ((CD)A)B), ((CD)B)A)	0.25659	0.22884	YES	NO	6.4E,M
	(0.05, 0.05)	(((AB)C)D), (((AB)D)C)	0.13384	0.10054	YES	YES	6.4F,N
		((AC)(BD)), ((BC)(AD))					

Species tree $\sigma$	Branch lengths ( $x, y$ ) or ( $w, x, y$ ) (see Figure 6.1)	Highest-prob. nonmatching gene tree	Concordance probability	Highest-prob. SM-ML nonmatching tree	SM-ML correct > 50%	SMRT-ML correct > 50%	Figures
	(0.1, 0.05)	((AC)(BD)), ((BC)(AD))	0.14516	0.09563	YES	YES	6.4G,O
	(0.25, 0.01)	((CD)A)B), ((CD)B)A)	0.16346	0.11584	YES	NO	6.4H,P
(((AB)C)D)E)	(0.1, 0.1, 0.1)	((AB)C)(DE))	0.02217	0.03321*	NO	YES	6.5A,E
	(0.1, 1.0, 0.1)	((AB)C)(DE))	0.09388	0.08158	NO	YES	6.5B,F
	(0.1, 0.1, 1.0)	((AB)C)(DE))	0.07055	0.08941*	NO	YES	6.5C,G
	(1.0, 0.1, 0.1)	((AB)(CD))E)	0.06547	0.07705*	YES	YES	6.5D,H
(((AB)C)(DE))	(0.1, 0.1, 0.1)	((DE)C)(AB))	0.04002	0.03034	YES	YES	6.6A,E
	(0.1, 1.0, 0.1)	((AC)B)(DE)), ((BC)A)(DE))	0.10506	0.07656	YES	YES	6.6B,F
	(0.1, 0.1, 1.0)	((AB)D)(CD)), ((AB)E)(CD))	0.10970	0.06465	YES	YES	6.6C,G
	(1.0, 0.1, 0.1)	((DE)C)(AB))	0.07781	0.08825*	NO	YES	6.6D,H
(((AB)(CD))E)	(0.1, 0.1, 0.1)	((AB)E)(CD)), ((CD)E)(AB))	0.02914	0.03626*	YES	YES	6.7A,E
	(0.1, 1.0, 0.1)	((CD)E)(AB))	0.07339	0.09065*	NO	YES	6.7B,F
	(0.1, 0.1, 1.0)	((AB)E)(CD))	0.07339	0.09065*	YES	YES	6.7C,G
	(1.0, 0.1, 0.1)	((AC)(BD))E), ((AD)(BC))E)	0.09591	0.05231	YES	YES	6.7D,H
(((AB)(CD))E)F)	see Figure 6.21A	((AB)(CD))(EF))	0.01525	0.02343*	NO	YES	6.18B

\*The most probable gene tree is an AGT.



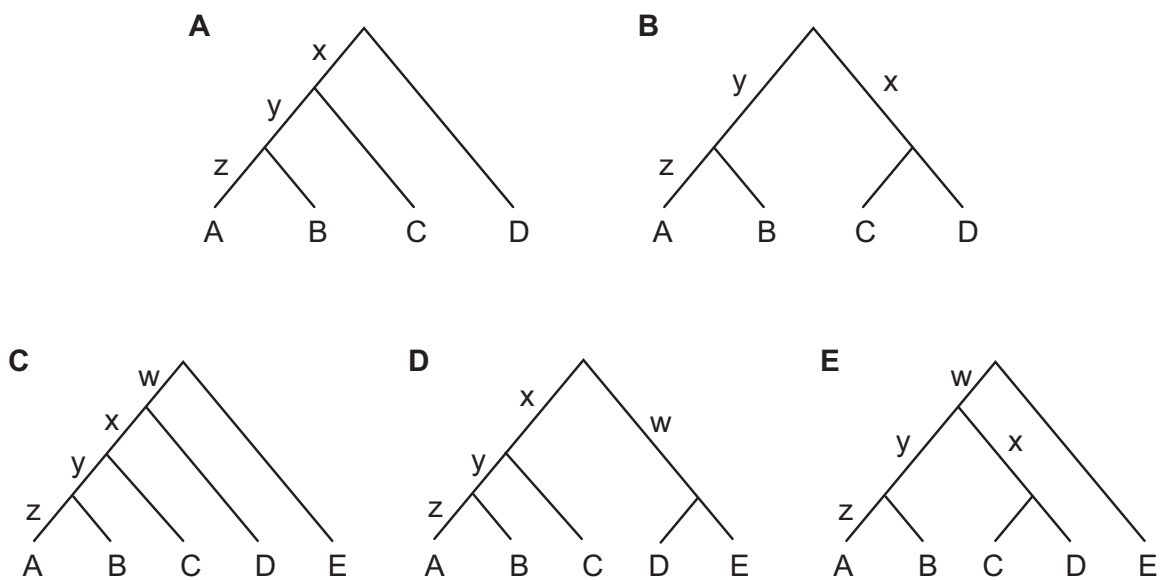


Figure 6.1: Four- and five-taxon clocklike species tree topologies. (A, B) Four-taxon species tree topologies with branch lengths  $x$ ,  $y$ , and  $z$ . (C-E) Five-taxon species tree topologies with branch lengths  $w$ ,  $x$ ,  $y$ , and  $z$ . Branch lengths are in coalescent time units  $t/(2N_e)$ , where  $t$  is the time in generations and  $N_e$  is the effective population size. For all simulations, we let  $z = 1$ .

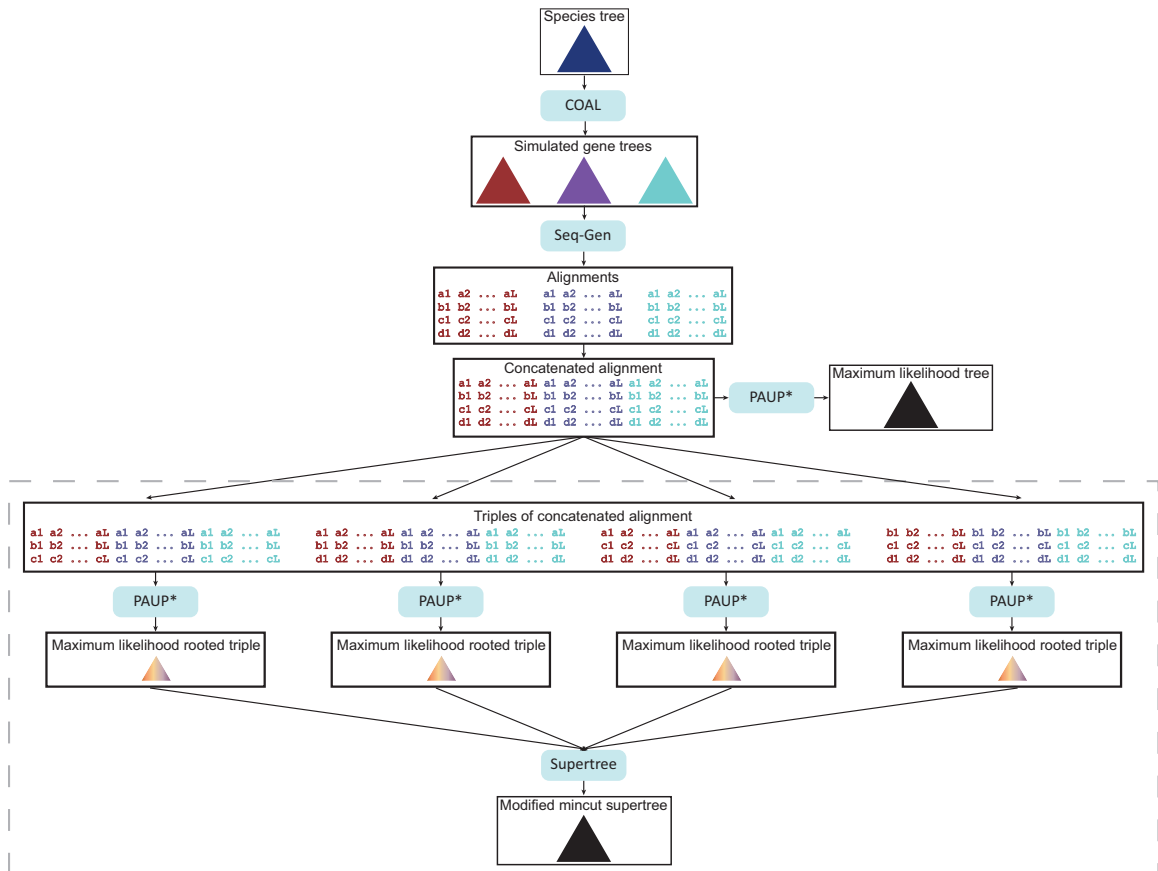


Figure 6.2: Schematic of our simulation procedure. First, an  $n$ -taxon species tree is chosen with branch lengths, which is fed through COAL (*Degnan and Salter, 2005*) to produce a set of  $n$ -taxon gene trees simulated under this species tree. SEQ-GEN (*Rambaut and Grassly, 1997*) is then used to create alignments of  $n$  species based on the gene trees, which are linked to create a single concatenated alignment. The concatenated alignment is analyzed under maximum likelihood (SM-ML) with PAUP\* (*Swofford, 2003*) to infer a species tree. The concatenated alignment is also broken into all  $\binom{n}{3}$  alignments of three species, which are then fed through PAUP\* to infer a total of  $\binom{n}{3}$  rooted triples. These rooted triples are used as input to SUPERTREE (*Page, 2002*) to infer a species tree (SMRT-ML). The dashed gray box represents the part of the procedure that is SMRT-ML.

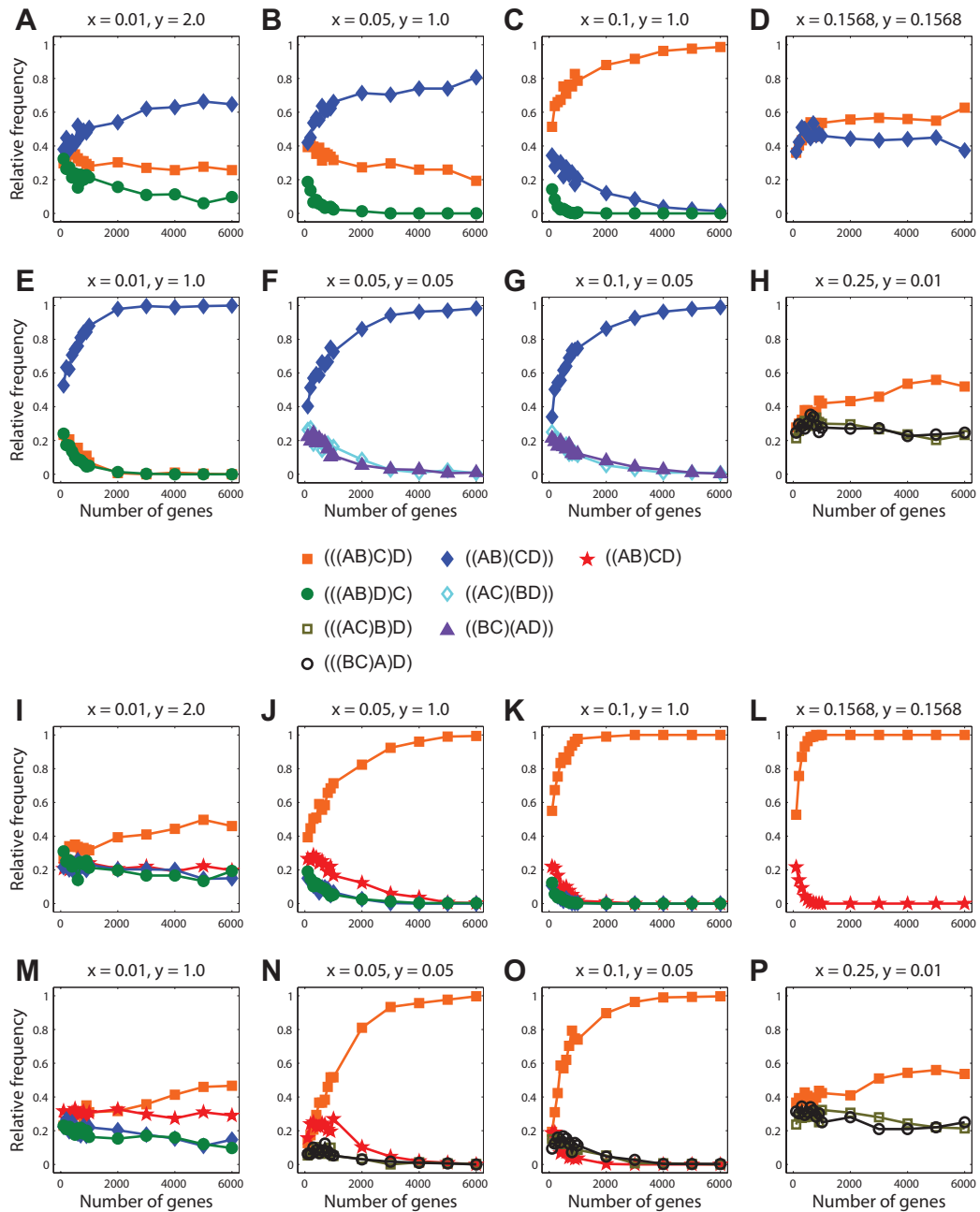


Figure 6.3: Results of simulations for the four-taxon tree  $((AB)C)D$  (Figure 6.1A) generated under a Jukes-Cantor model with  $\theta = 0.01$  and a molecular clock, and analyzed under maximum likelihood assuming a molecular clock and a Jukes-Cantor model. (A-H) SM-ML (resimulated from *Kubatko and Degnan (2007)*). (I-P) SMRT-ML. Data for each combination of branch lengths and number of loci were generated from 300 independent simulations.

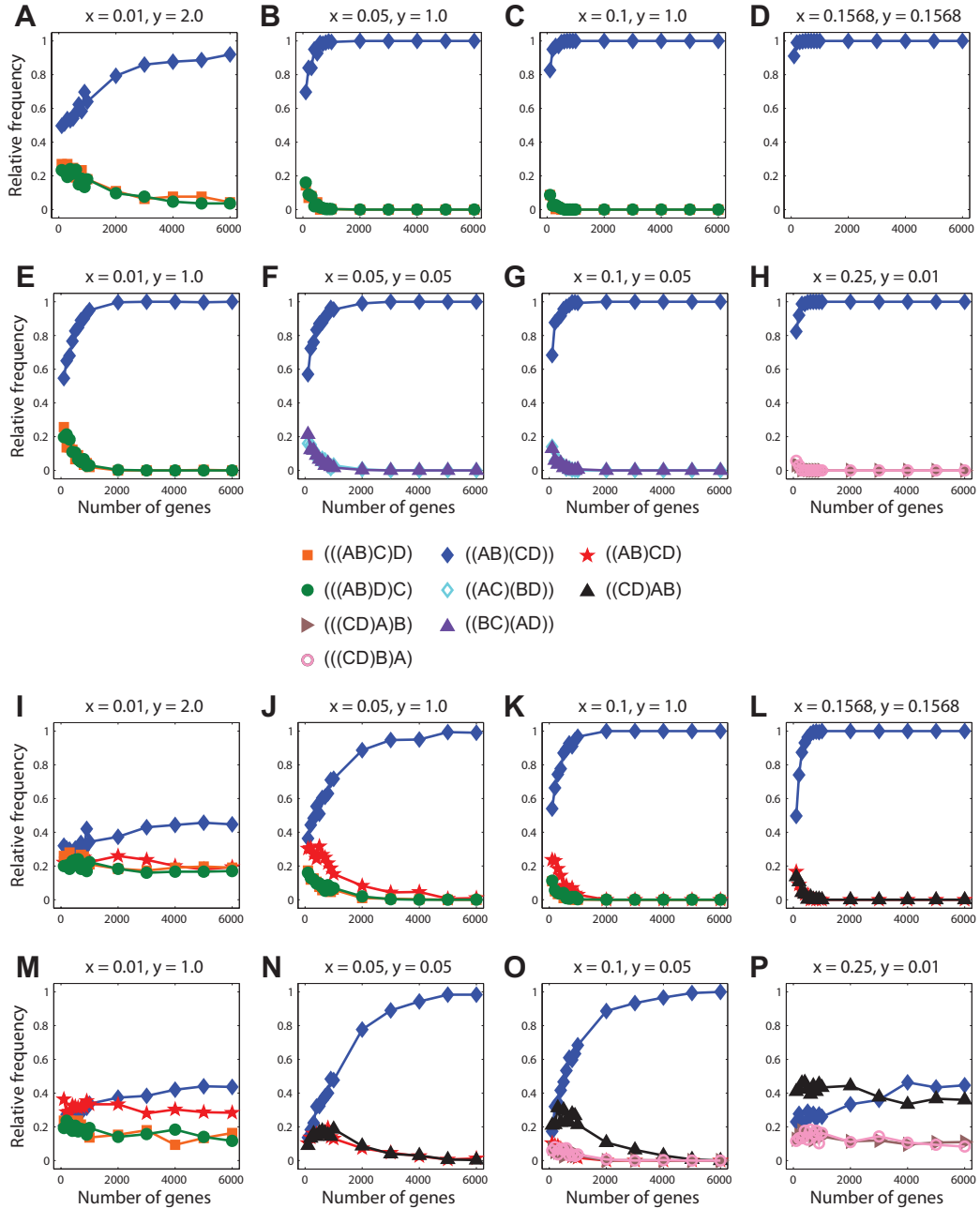


Figure 6.4: Results of simulations for the four-taxon tree  $((AB)(CD))$  (Figure 6.1B) generated under a Jukes-Cantor model with  $\theta = 0.01$  and a molecular clock, and analyzed under maximum likelihood assuming a molecular clock and a Jukes-Cantor model. (A-H) SM-ML. (I-P) SMRT-ML. Data for each combination of branch lengths and number of loci were generated from 300 independent simulations.

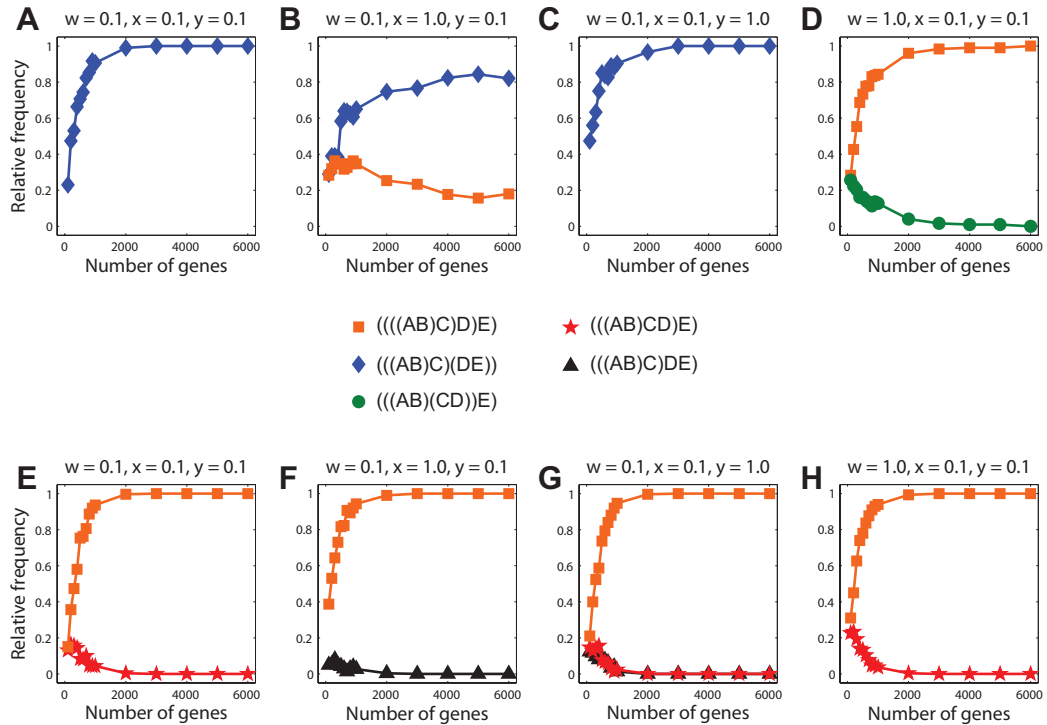


Figure 6.5: Results of simulations for the five-taxon tree  $((((AB)C)D)E)$  (Figure 6.1C) generated under a Jukes-Cantor model with  $\theta = 0.01$  and a molecular clock, and analyzed under maximum likelihood assuming a molecular clock and a Jukes-Cantor model. (A-E) SM-ML. (E-H) SMRT-ML. Data for each combination of branch lengths and number of loci were generated from 300 independent simulations..

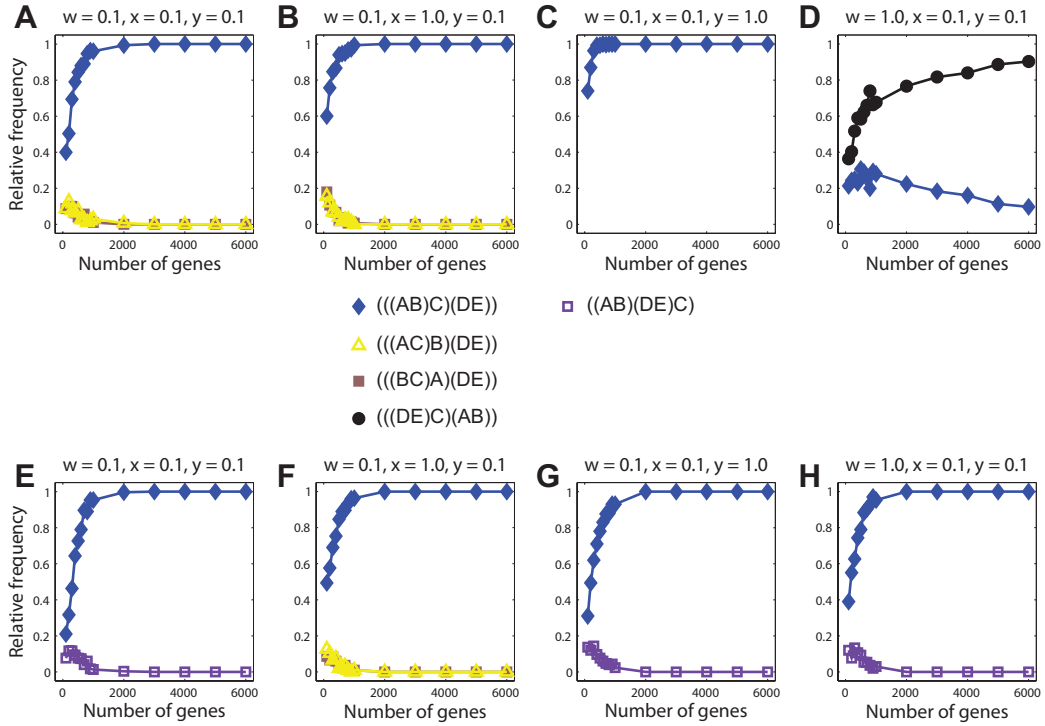


Figure 6.6: Results of simulations for the five-taxon tree  $(((AB)C)(DE))$  (Figure 6.1D) generated under a Jukes-Cantor model with  $\theta = 0.01$  and a molecular clock, and analyzed under maximum likelihood assuming a molecular clock and a Jukes-Cantor model. (A-E) SM-ML. (E-H) SMRT-ML. Data for each combination of branch lengths and number of loci were generated from 300 independent simulations.

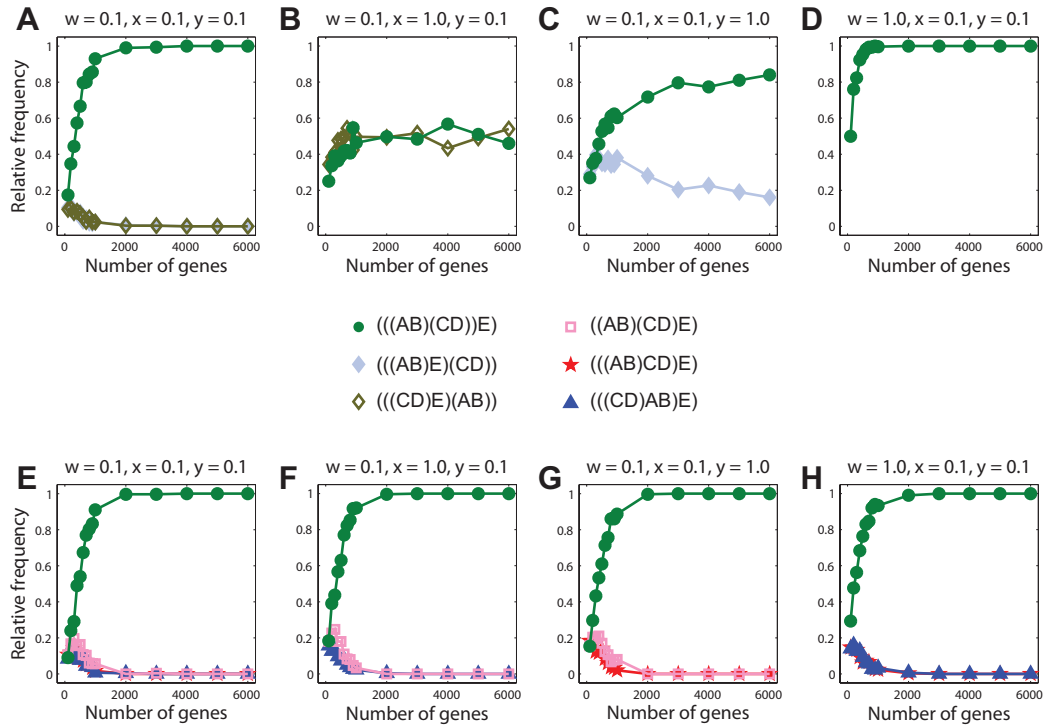


Figure 6.7: Results of simulations for the five-taxon tree  $((((AB)(CD))E)$  (Figure 6.1E) generated under a Jukes-Cantor model with  $\theta = 0.01$  and a molecular clock, and analyzed under maximum likelihood assuming a molecular clock and a Jukes-Cantor model. (A-E) SM-ML. (E-H) SMRT-ML. Data for each combination of branch lengths and number of loci were generated from 300 independent simulations.

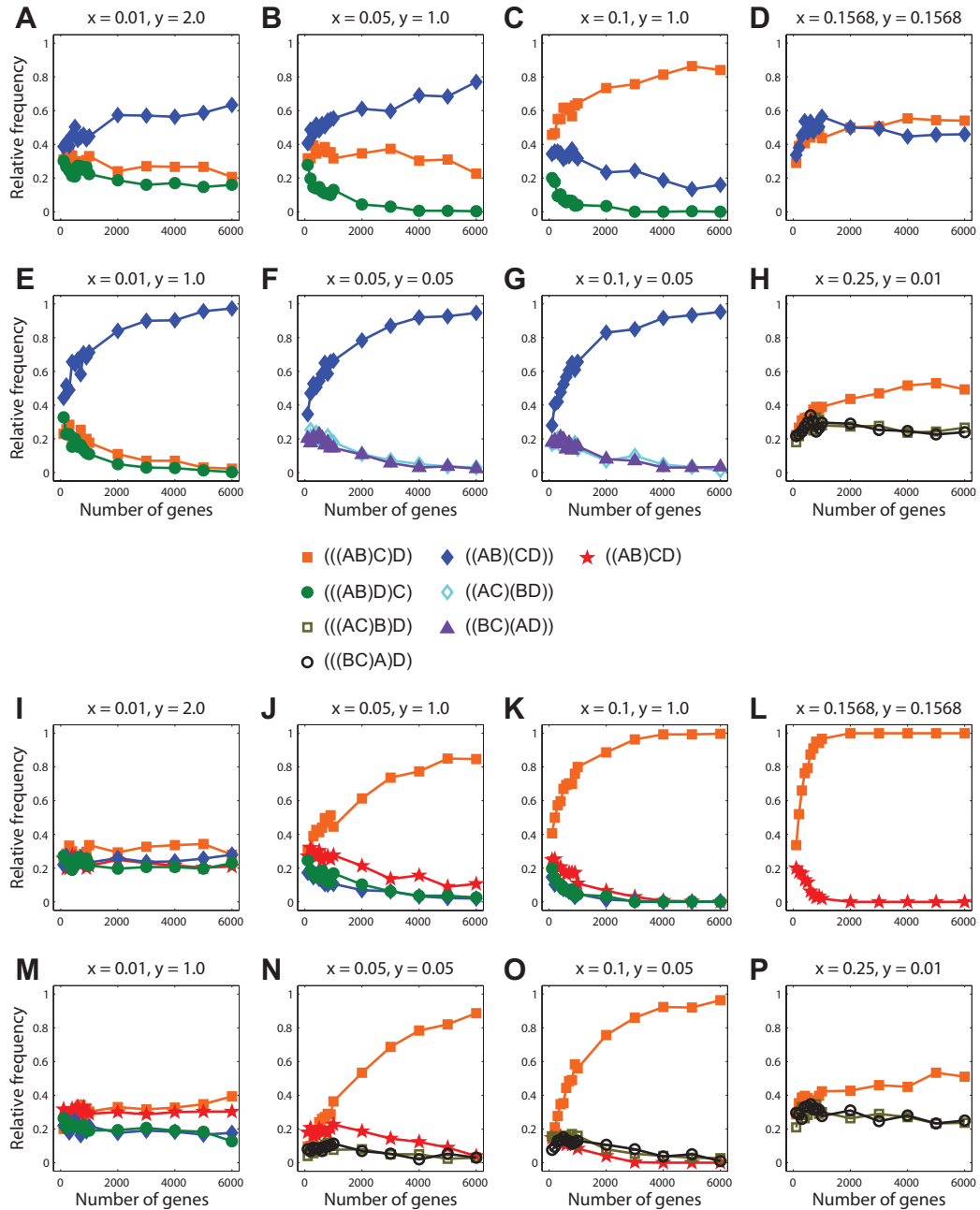


Figure 6.8: Results of simulations for the four-taxon tree  $(((AB)C)D)$  (Figure 6.1A) generated under a Jukes-Cantor model with  $\theta = 0.01$  and a violation of the molecular clock, and analyzed under maximum likelihood assuming a molecular clock and a Jukes-Cantor model. (A-H) SM-ML. (I-P) SMRT-ML. Data for each combination of branch lengths and number of loci were generated from 300 independent simulations.



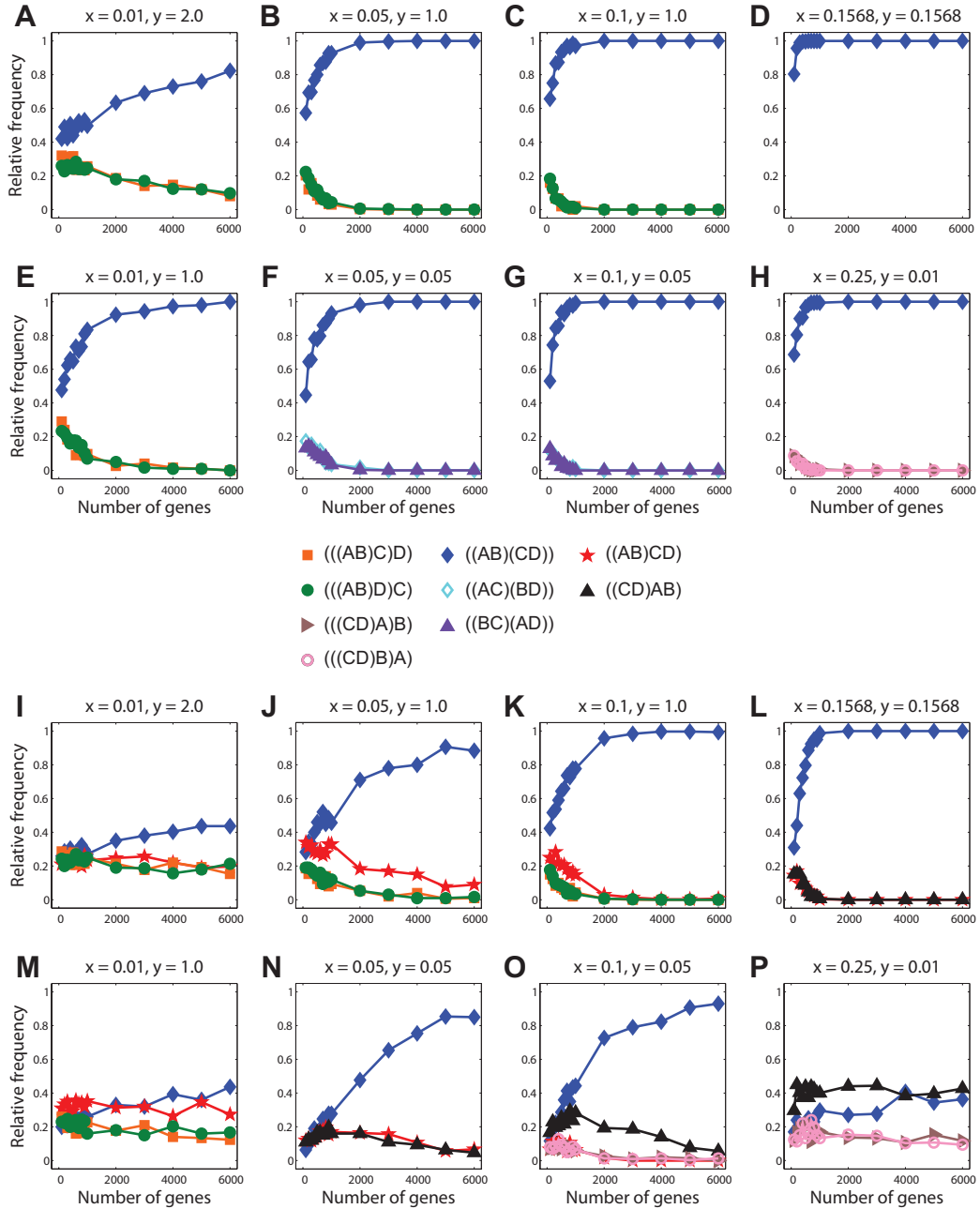


Figure 6.9: Results of simulations for the four-taxon tree  $((AB)(CD))$  (Figure 6.1B) generated under a Jukes-Cantor model with  $\theta = 0.01$  and a violation of the molecular clock, and analyzed under maximum likelihood assuming a molecular clock and a Jukes-Cantor model. (A-H) SM-ML. (I-P) SMRT-ML. Data for each combination of branch lengths and number of loci were generated from 300 independent simulations.

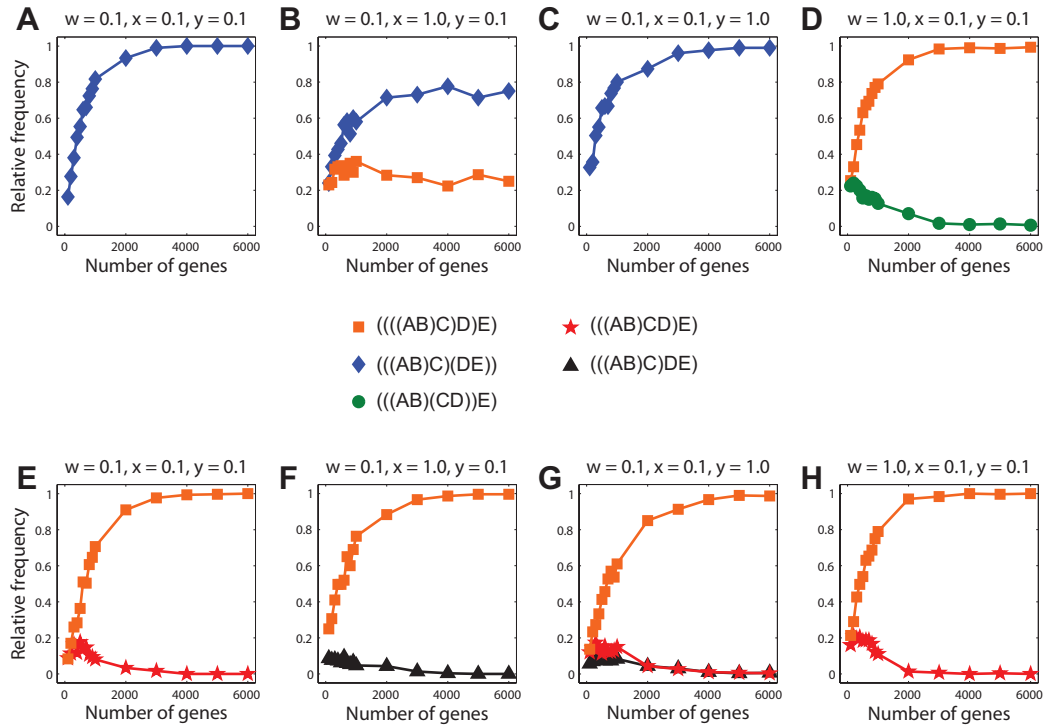


Figure 6.10: Results of simulations for the five-taxon tree  $((((AB)C)D)E)$  (Figure 6.1C) generated under a Jukes-Cantor model with  $\theta = 0.01$  and a violation of the molecular clock, and analyzed under maximum likelihood assuming a molecular clock and a Jukes-Cantor model. (A-D) SM-ML. (E-H) SMRT-ML. Data for each combination of branch lengths and number of loci were generated from 300 independent simulations.

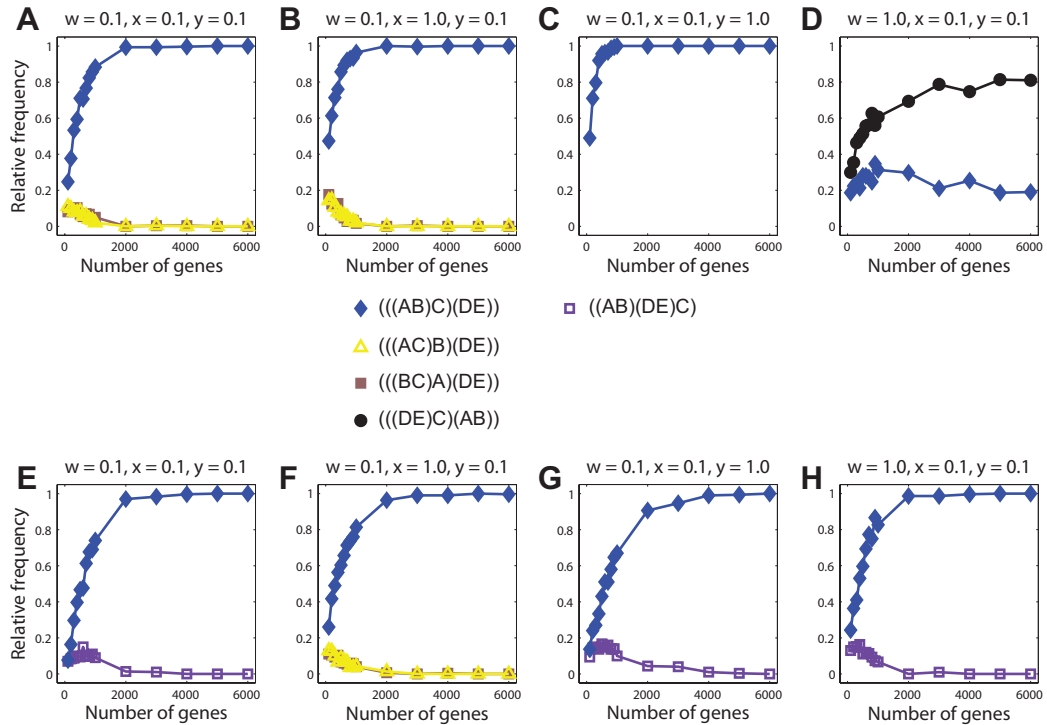


Figure 6.11: Results of simulations for the five-taxon tree  $(((AB)C)(DE))$  (Figure 6.1D) generated under a Jukes-Cantor model with  $\theta = 0.01$  and a violation of the molecular clock, and analyzed under maximum likelihood assuming a molecular clock and a Jukes-Cantor model. (A-D) SM-ML. (E-H) SMRT-ML. Data for each combination of branch lengths and number of loci were generated from 300 independent simulations.

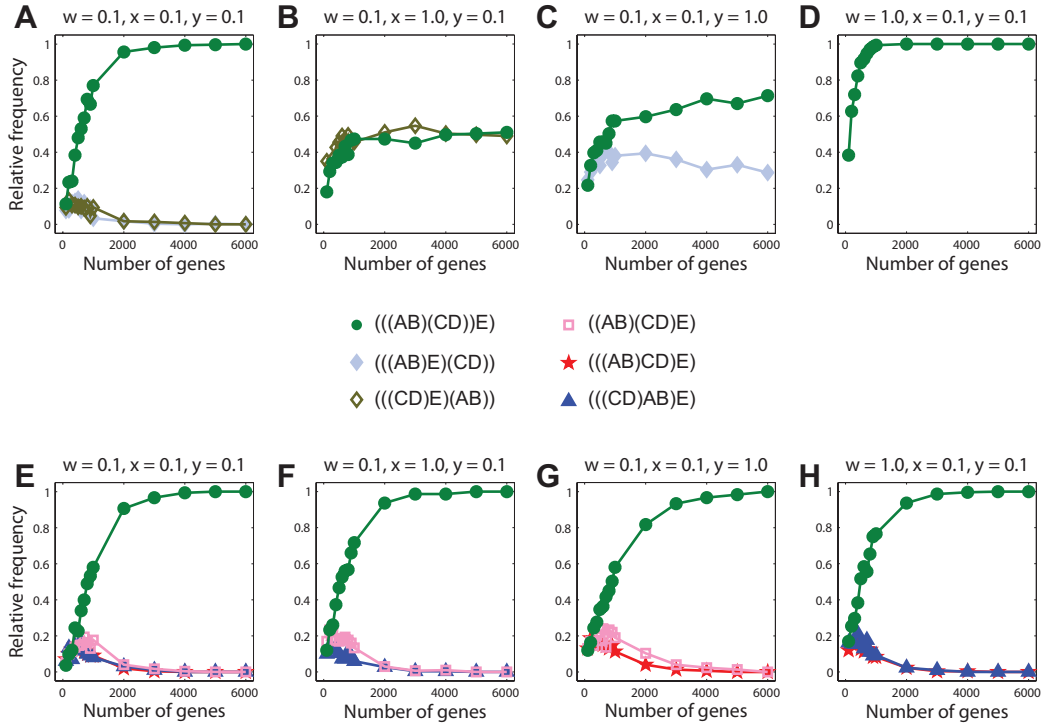


Figure 6.12: Results of simulations for the five-taxon  $((AB)(CD))E$  (Figure 6.1E) generated under a Jukes-Cantor model with  $\theta = 0.01$  and a violation of the molecular clock, and analyzed under maximum likelihood assuming a molecular clock and a Jukes-Cantor model. (A-D) SM-ML. (E-H) SMRT-ML. Data for each combination of branch lengths and number of loci were generated from 300 independent simulations.

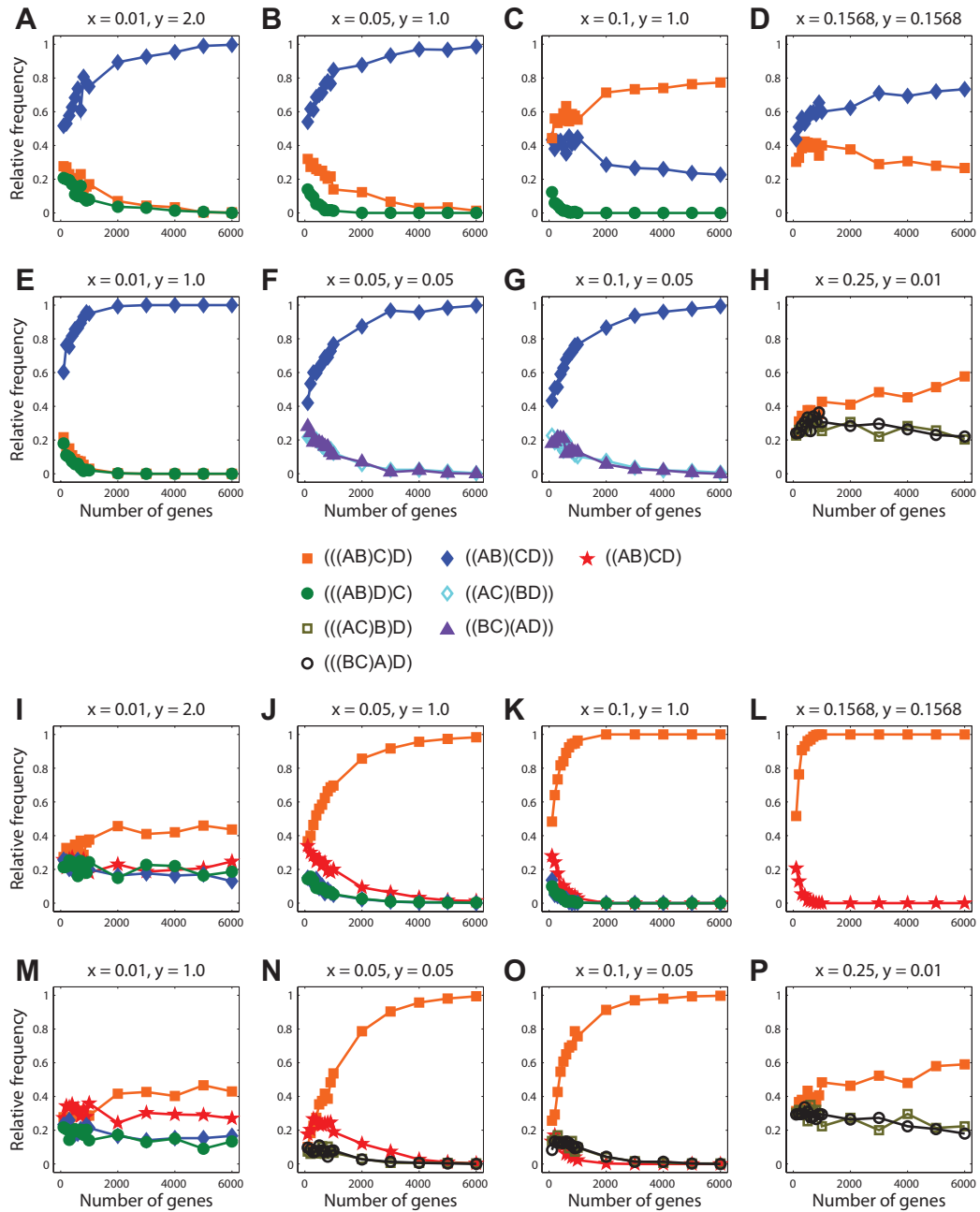


Figure 6.13: Results of simulations for the four-taxon tree  $((AB)C)D$  (Figure 6.1A) generated under a General Time-Reversible model with shape parameter  $\alpha = 1$ , relative frequencies for nucleotides  $(A, C, G, T) = (0.1, 0.2, 0.3, 0.4)$ , relative rates of substitutions  $(A \leftrightarrow C, A \leftrightarrow G, A \leftrightarrow T, C \leftrightarrow G, C \leftrightarrow T, G \leftrightarrow T) = (1.5, 1.5, 0.5, 10.5, 1.0, 6.0)$ ,  $\theta = 0.01$ , and a molecular clock, and analyzed under maximum likelihood assuming a molecular clock and a Jukes-Cantor model. (A-H) SM-ML. (I-P) SMRT-ML. Data for each combination of branch lengths and number of loci were generated from 300 independent simulations.

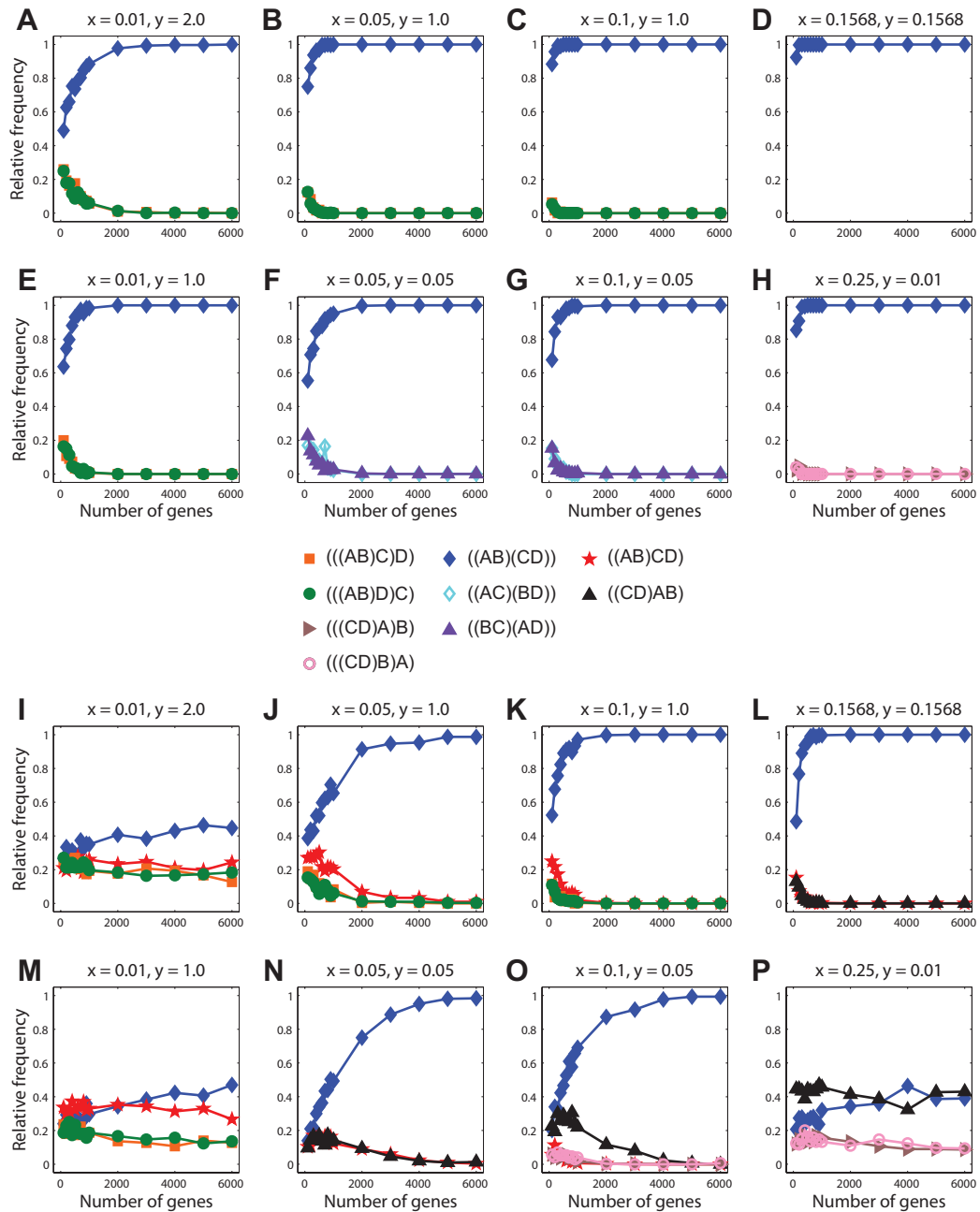


Figure 6.14: Results of simulations for the four-taxon tree  $((AB)(CD))$  (Figure 6.1B) generated under a General Time-Reversible model with shape parameter  $\alpha = 1$ , relative frequencies for nucleotides  $(A, C, G, T) = (0.1, 0.2, 0.3, 0.4)$ , relative rates of substitutions  $(A \leftrightarrow C, A \leftrightarrow G, A \leftrightarrow T, C \leftrightarrow G, C \leftrightarrow T, G \leftrightarrow T) = (1.5, 1.5, 0.5, 10.5, 1.0, 6.0)$ ,  $\theta = 0.01$ , and a molecular clock, and analyzed under maximum likelihood assuming a molecular clock and a Jukes-Cantor model. (A-H) SM-ML. (I-P) SMRT-ML. Data for each combination of branch lengths and number of loci were generated from 300 independent simulations.

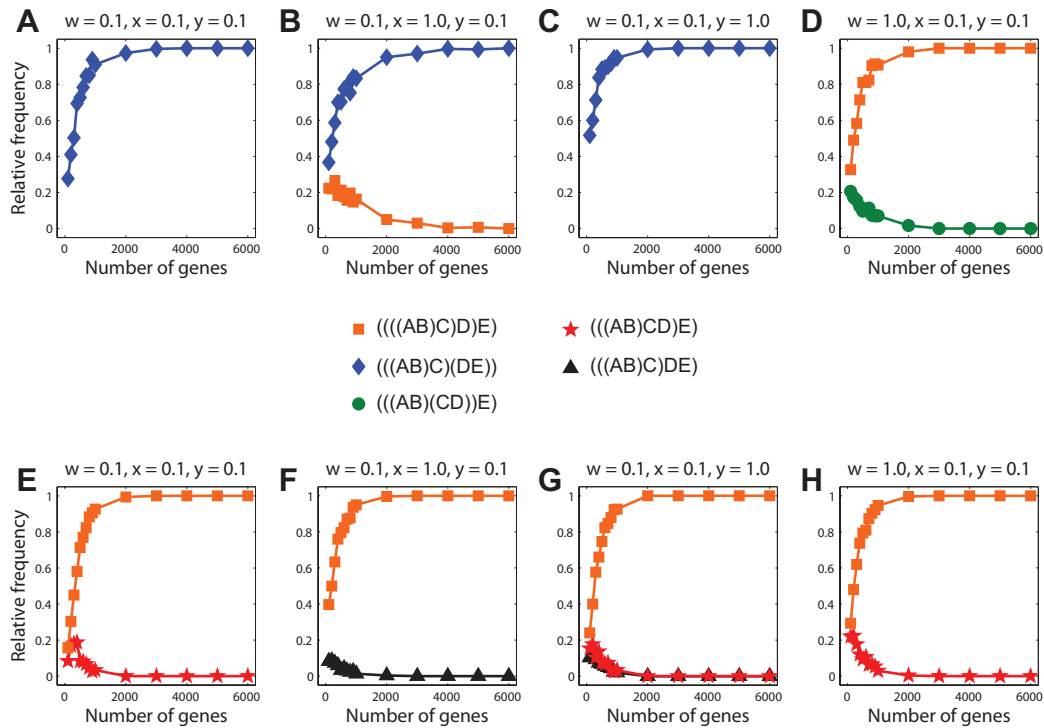


Figure 6.15: Results of simulations for the five-taxon tree  $((((AB)C)D)E)$  (Figure 6.1C) generated under a General Time-Reversible model with shape parameter  $\alpha = 1$ , relative frequencies for nucleotides (A, C, G, T) = (0.1, 0.2, 0.3, 0.4), relative rates of substitutions (A $\leftrightarrow$ C, A $\leftrightarrow$ G, A $\leftrightarrow$ T, C $\leftrightarrow$ G, C $\leftrightarrow$ T, G $\leftrightarrow$ T) = (1.5, 1.5, 0.5, 10.5, 1.0, 6.0),  $\theta = 0.01$ , and a molecular clock, and analyzed under maximum likelihood assuming a molecular clock and a Jukes-Cantor model. (A-D) SM-ML. (E-H) SMRT-ML. Data for each combination of branch lengths and number of loci were generated from 300 independent simulations.

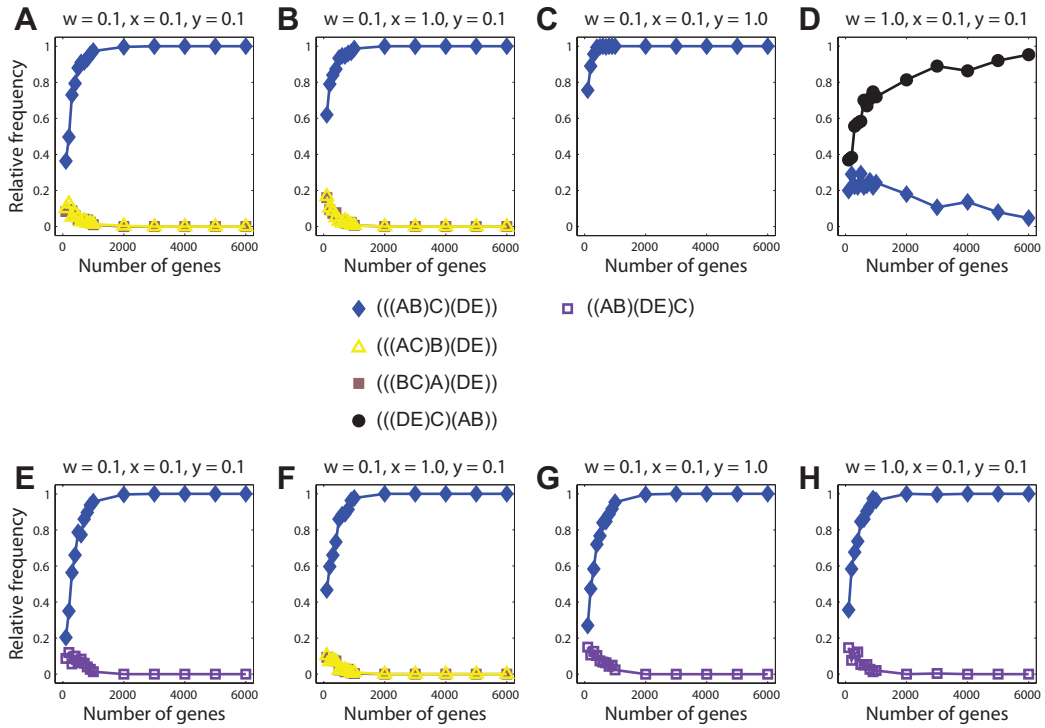


Figure 6.16: Results of simulations for the five-taxon tree  $(((AB)C)(DE))$  (Figure 6.1D) generated under a General Time-Reversible model with shape parameter  $\alpha = 1$ , relative frequencies for nucleotides (A, C, G, T) = (0.1, 0.2, 0.3, 0.4), relative rates of substitutions (A $\leftrightarrow$ C, A $\leftrightarrow$ G, A $\leftrightarrow$ T, C $\leftrightarrow$ G, C $\leftrightarrow$ T, G $\leftrightarrow$ T) = (1.5, 1.5, 0.5, 10.5, 1.0, 6.0),  $\theta = 0.01$ , and a molecular clock, and analyzed under maximum likelihood assuming a molecular clock and a Jukes-Cantor model. (A-D) SM-ML. (E-H) SMRT-ML. Data for each combination of branch lengths and number of loci were generated from 300 independent simulations.



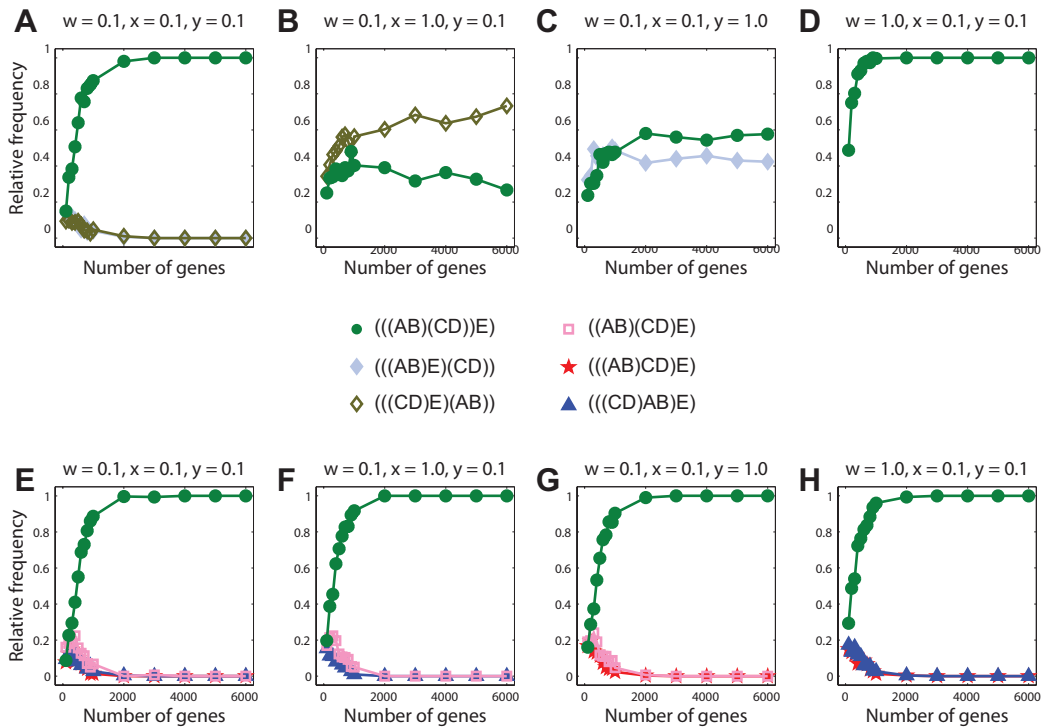


Figure 6.17: Results of simulations for the five-taxon tree  $((((AB)(CD))E)$  (Figure 6.1E) generated under a General Time-Reversible model with shape parameter  $\alpha = 1$ , relative frequencies for nucleotides (A, C, G, T) = (0.1, 0.2, 0.3, 0.4), relative rates of substitutions (A $\leftrightarrow$ C, A $\leftrightarrow$ G, A $\leftrightarrow$ T, C $\leftrightarrow$ G, C $\leftrightarrow$ T, G $\leftrightarrow$ T) = (1.5, 1.5, 0.5, 10.5, 1.0, 6.0),  $\theta = 0.01$ , and a molecular clock, and analyzed under maximum likelihood assuming a molecular clock and a Jukes-Cantor model. (A-D) SM-ML. (E-H) SMRT-ML. Data for each combination of branch lengths and number of loci were generated from 300 independent simulations.

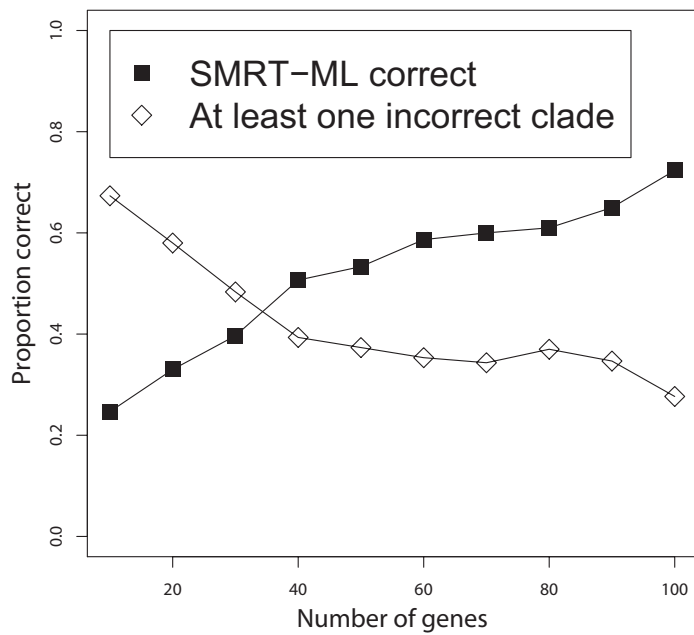


Figure 6.18: Proportion of times SMRT-ML recovers the estimated species tree or at least one false clade for random subsets of genes from the original data set. The two proportions do not add up to 100% because in some cases a partially unresolved tree, which does not have any false clades, is returned by SMRT-ML.

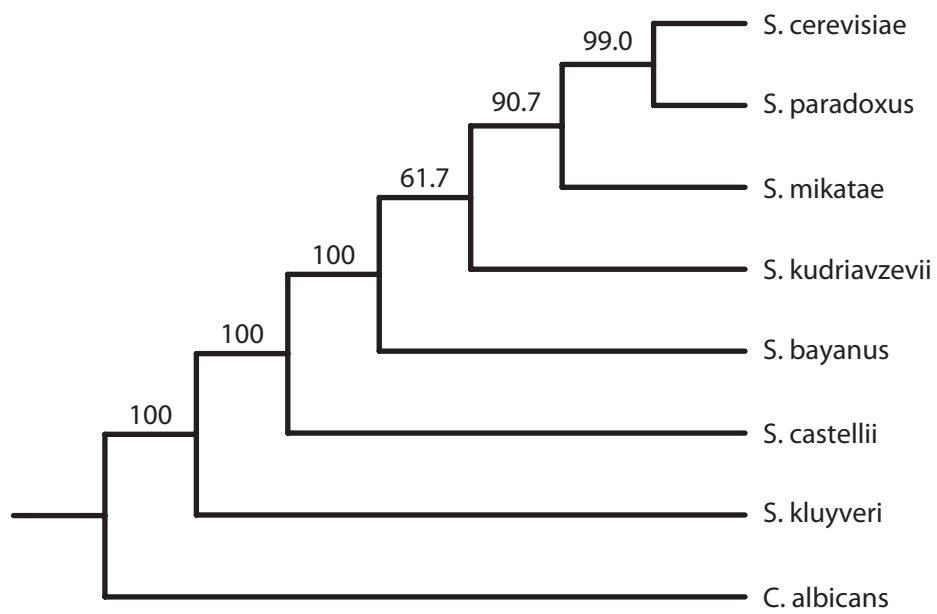


Figure 6.19: Bootstrap support percentages for nodes in the SMRT-ML yeast analysis using the 106-gene data set. Proportions are based on 300 bootstrap replicates.

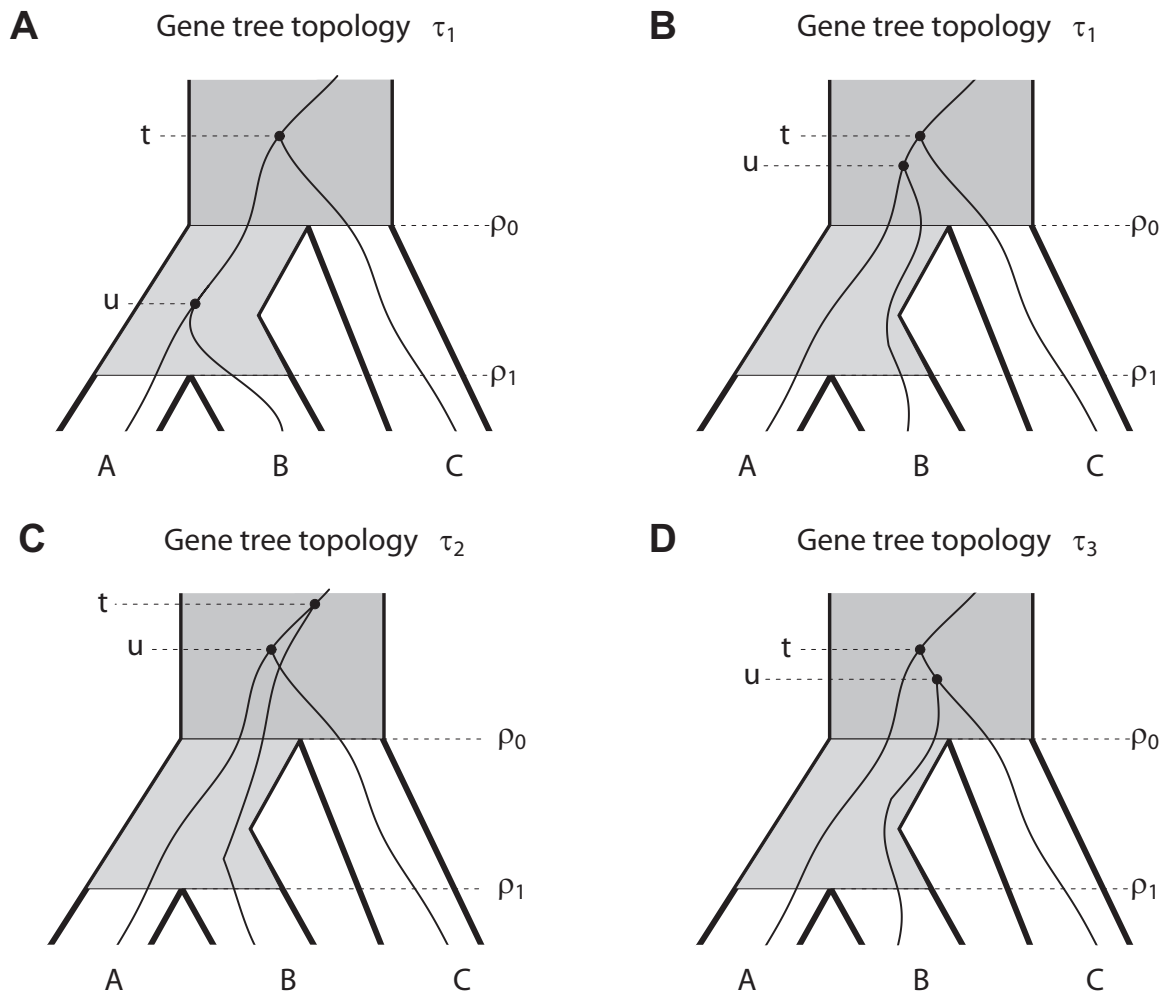


Figure 6.20: A three-taxon gene tree within a model species tree with notation used in the paper. In all cases, the species tree has the topology  $((AB)C)$ . Dots represent coalescent events. (A) and (B) depict the same gene tree topology with different coalescent histories. The gene tree in (C) has the  $((AC)B)$  topology; the gene tree in (D) has the  $((BC)A)$  topology.

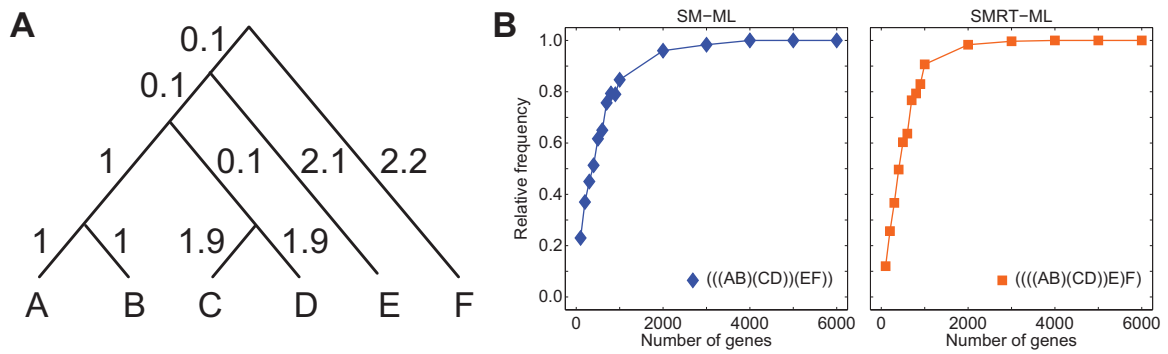


Figure 6.21: Results of simulations for a six-taxon tree with topology  $((((AB)(CD))E)F)$ . (A) Species tree. (B) SM-ML and SMRT-ML applied to simulated data under a Jukes-Cantor model with  $\theta = 0.01$  satisfying a molecular clock analyzed assuming a molecular clock. Data for each combination of branch lengths and number of loci were generated from 300 independent simulations.

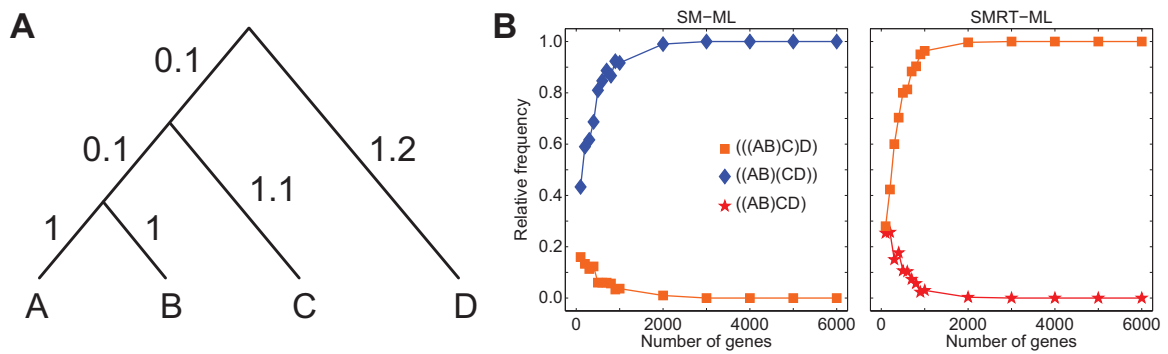


Figure 6.22: Results of simulations for a four-taxon tree with topology  $((((AB)C)D)$ . (A) Species tree. (B) SM-ML and SMRT-ML applied to simulated data under a Jukes-Cantor model with  $\theta = 0.01$  satisfying a molecular clock analyzed assuming a molecular clock. Data for each combination of branch lengths and number of loci were generated from 300 independent simulations.

## CHAPTER VII

# Consistency of phylogenetic consensus methods in the presence of ancestral population structure

### 7.1 Introduction

Recently, much attention has been given to the development of methods that consistently infer the correct species tree from discordant gene trees (*Rannala and Yang, 2008*). This work has largely focused on incomplete lineage sorting—which occurs when lineages from two different species fail to coalesce in the population immediately ancestral to the split of the two species—as a source of gene tree discordance (*Degnan and Rosenberg, 2009*).

Consensus methods, each of which takes a set of gene trees as input and returns a species tree estimate according to a specific rule (*Bryant, 2003*), have provided one important source of methods for species tree inference. A consensus method  $\hat{C}$  is a statistically consistent estimator of a species tree topology under some model if for each species tree  $\sigma$ ,  $\hat{C}$  applied to a set of gene trees randomly generated under the model, assuming that the species tree is  $\sigma$ , converges in probability to the topology of  $\sigma$  as the number of gene trees approaches infinity. Statistical consistency is a desirable property because it is reasonable to expect that as more data are gathered, evidence should accumulate in support of the true value of the parameter being

estimated. *Degnan and Rosenberg* (2006) showed that when gene trees are distributed according to the multispecies coalescent model of the evolution of gene lineages conditional on a species tree, an extreme case of incomplete lineage sorting arises in which the most likely gene tree topology does not match the species tree topology. This inconsistency implies that species tree estimation must use information other than the most frequently occurring gene tree topology to accurately infer the species tree topology. Indeed, many consensus methods relying on other principles provide statistically consistent estimators of the species tree topology under the multispecies coalescent model. This collection of methods includes STEAC (*Liu et al.*, 2009), STAR (*Liu et al.*, 2009), R\* Consensus (*Degnan et al.*, 2009), GLASS (*Mossel and Roch*, 2010), and Maximum Tree (*Liu et al.*, 2010).

In its simplest form, the multispecies coalescent model assumes that each modern species and each ancestral species has a constant population size, each pair of lineages within a given ancestral species has an equal chance of coalescing, and each species is unstructured. Because the multispecies coalescent model assumes that random mating occurs within species, when ancestral species are structured, as has been hypothesized for various species (e.g., *Garrigan et al.*, 2005; *Thalmann et al.*, 2007; *White et al.*, 2009), it is unclear whether methods that are consistent under the multispecies coalescent continue to be consistent.

The difficulty of species tree estimation in the presence of ancestral population structure lies in the way that population structure alters the probability distribution of gene trees given a species tree compared to the unstructured case (*Slatkin and Pollack*, 2008). Using a three-taxon example, *Slatkin and Pollack* (2008) showed that with ancestral population structure, the probability distribution of gene tree topologies can have a certain asymmetry, and the most likely three-taxon gene tree topology need not match the species tree topology. These consequences of the multispecies coalescent with ancestral population structure do not occur in the



standard multispecies coalescent.

Here, we describe an extension of the structured ancestral population model considered by *Slatkin and Pollack (2008)*. Using our extended structured model, we evaluate the consistency of several consensus methods, employing a single counterexample to show that many methods are inconsistent. For each inconsistent method, we show that it is in fact misleading in the sense that for a certain fixed species tree  $\sigma$  and a particular set of parameters, the probability that the consensus tree contains a clade not present on  $\sigma$  approaches 1 as the number of loci approaches infinity. To evaluate the behavior of the various consensus methods in practice (*i.e.*, the speed at which they converge to or diverge from the correct bifurcating species tree topology), we perform simulations, assuming an island migration model as our structured ancestral population model. As is predicted by our theoretical results, the only method that does not provide strong support for an incorrect species tree topology is GLASS/Maximum Tree. However, from simulations using model species trees both with and without ancestral population structure, we show that GLASS/Maximum Tree performs poorly when little information exists in sequence alignments (e.g., an absence of substitutions between species, causing inferred gene trees to have branches of length zero). From these results, we conclude that increased attention is needed to the development of consensus methods that accurately infer species trees in the presence of ancestral population structure.

### 7.1.1 Model

We use the notation in Table 7.1. Suppose time is measured in generations. Consider an ultrametric  $n$ -taxon bifurcating species tree  $\sigma$  with  $n \geq 3$  taxa (*i.e.*, each leaf has an identical sum of branch lengths to the root, in units of generations). Then we can always find a set of species A, B, and C on  $\sigma$  with the relationship  $((A:\tau_3, B:\tau_3):\tau_2 - \tau_3, C:\tau_2)$ , where  $\tau_2 > \tau_3 > 0$  and  $\tau_2$  and  $\tau_3$  are measured in

generations.

Each internal branch along the species tree specifies an ancestral population. An  $n$ -taxon species tree contains  $n - 1$  ancestral populations, including the branch above the root. Label the ancestral populations of  $\sigma$  by recursively visiting the root, then the left subtree, and finally the right subtree (a pre-order traversal of  $\sigma$ ). Each ancestral population might be a structured population and suppose that the structured population model is invariant across  $L$  independent loci (gene trees), so that each of the  $L$  gene trees can be viewed as a random variate conditional on the same assumed species tree. Let  $D^{(i)}$  be the number of demes in ancestral population  $i$  (the number of demes is a finite positive integer), let  $\mathbf{N}^{(i)}$  be the vector of sizes for the  $D^{(i)}$  demes in ancestral population  $i$  (each population size is a finite positive integer), and let  $\mathbf{M}^{(i)}$  be the backward migration matrix between demes in ancestral population  $i$  (Figure 7.1).

Denote a structured ancestral population model by  $\mathcal{S} = \mathcal{S}(\sigma, \mathbf{D}, \mathbf{N}, \mathbf{M}, \Psi)$ , where  $\mathbf{D} = [D^{(1)}, D^{(2)}, \dots, D^{(n-1)}]$ ,  $\mathbf{N} = [\mathbf{N}^{(1)}, \mathbf{N}^{(2)}, \dots, \mathbf{N}^{(n-1)}]$ ,  $\mathbf{M} = [\mathbf{M}^{(1)}, \mathbf{M}^{(2)}, \dots, \mathbf{M}^{(n-1)}]$ , and  $\Psi$  is a  $(n + \sum_{i=1}^{n-1} D^{(i)}) \times (n + \sum_{i=1}^{n-1} D^{(i)})$  matrix that describes how demes connect across species divergences. Each row (column) of  $\Psi$  corresponds to a distinct deme among the extant and ancestral populations. The first  $n$  rows (columns) correspond to the  $n$  extant populations, the next  $D^{(1)}$  rows (columns) correspond to the  $D^{(1)}$  demes in ancestral population 1, the next  $D^{(2)}$  rows (columns) correspond to the  $D^{(2)}$  demes in ancestral population 2, and so on, until the last  $D^{(n-1)}$  rows (columns) correspond to the  $D^{(n-1)}$  demes in ancestral population  $n - 1$ . Extant populations each contain only a single deme because they are unstructured. The element  $\Psi_{jk}$  provides the probability that a lineage merges into deme  $k$  from deme  $j$  at the moment in which, going back in time, deme  $j$  ends and deme  $k$  begins. If, going back in time, deme  $k$  does not directly receive lineages from deme  $j$  at the moment that deme  $j$  ends and deme  $k$  begins, then  $\Psi_{jk} = 0$ . By

construction, each row of  $\Psi$  sums to 1. Therefore,  $\Psi$  provides probability distributions on the locations in the ancestral populations of  $\sigma$  into which lineages sampled from the taxa of  $\sigma$  can merge.

For each ancestral population  $i$ ,  $\mathbf{M}_{xy}^{(i)}$  is the per-generation probability of backward migration from deme  $x^{(i)}$  to deme  $y^{(i)}$  for a lineage in deme  $x^{(i)}$ . Assume that in any ancestral population  $i$ , demes  $x^{(i)}$  and  $y^{(i)}$  communicate. In other words, the migration rate from deme  $x^{(i)}$  to deme  $y^{(i)}$  is nonzero, or for any pair of lineages there otherwise exists an indirect migration path through other demes from deme  $x^{(i)}$  to deme  $y^{(i)}$ . This assumption encodes the idea of what we mean by a structured population and, by ensuring that demes communicate in the ancestral population above the root, it guarantees that with probability 1 the coalescence process will terminate. The relationship between species A, B, and C within the  $n$ -taxon species tree  $\sigma$ , and the structured ancestral population model, are illustrated in Figure 7.1.

We are interested in computing the probabilities  $\mathbb{P}[E \mid \mathcal{S}]$  of events  $E$ , conditional on model  $\mathcal{S}$ . Such probabilities are possible to compute by connecting models of individual populations along the branches of species tree  $\sigma$  with rules given by model  $\mathcal{S}$  about what happens to lineages at divergence times.

### 7.1.2 Counterexample

We use a single counterexample to prove that, at least in part of the parameter space of our structured population model, Democratic Vote (*Degnan and Rosenberg, 2006, 2009*), STAR (*Liu et al., 2009*), STEAC (*Liu et al., 2009*), R\* Consensus (*Bryant, 2003; Degnan et al., 2009*), Rooted Triple Consensus (*Ewing et al., 2008*), Minimize Deep Coalescences (MDC; *Maddison, 1997; Maddison and Knowles, 2006; Than and Nakhleh, 2009*), and Majority-Rule Consensus (*Degnan et al., 2009*) are misleading in that the probability that the consensus tree contains a clade not on the species tree goes to 1 as the number of loci goes to infinity. Consider a sample of

$n$  individuals, one from each species within an  $n$ -taxon species tree  $\sigma$ . Figure 7.2A displays a model of three species A, B, and C that have the topological relationship ((AB)C) within an  $n$ -taxon species tree  $\sigma$ . Certain internal branches are made long so that  $\sigma$  resembles a three-taxon species tree, in the sense that coalescences of lineages from the  $n - 3$  taxa other than A, B, and C with lineages from A, B, and C are very likely to occur on these long internal branches (Figure 7.2B).

Let  $\lambda_A$  be the subtree of  $\sigma$  that contains species A and that descends from the split of species A and B, let  $\lambda_B$  be the subtree of  $\sigma$  that contains species B and that descends from the split of species A and B, and let  $\lambda_C$  be the subtree of  $\sigma$  that contains species C and that descends from the split of species (AB) and C. Further, let  $\Gamma_A$ ,  $\Gamma_B$ , and  $\Gamma_C$  denote the sets of taxa at the leaves of subtrees  $\lambda_A$ ,  $\lambda_B$ , and  $\lambda_C$ , respectively. By definition,  $\Gamma_A \cap \Gamma_B = \emptyset$ ,  $\Gamma_A \cap \Gamma_C = \emptyset$ ,  $\Gamma_B \cap \Gamma_C = \emptyset$ , and  $\Gamma_A \cup \Gamma_B \cup \Gamma_C$  is the set of all taxa on species tree  $\sigma$ . Given a set of taxa  $X$ , we denote the tree displayed by phylogenetic tree  $\mathcal{T}$  restricted to  $X$  by  $\mathcal{T}|X$ . We denote the topology of phylogenetic tree  $\mathcal{T}$  as  $\text{top}(\mathcal{T})$ . To show that a consensus method is misleading, it suffices to find a set of branch lengths on a fixed species tree  $\sigma$  such that as the number of loci approaches infinity, the probability approaches 1 that the inferred species tree contains a clade not on  $\sigma$ . Therefore, we only need to find one counterexample to prove that a consensus method is misleading.

In our counterexample, we suppose that certain internal branches are long enough so that for a fixed set of taxa  $X \in \{\Gamma_A, \Gamma_B, \Gamma_C\}$ , fixed species tree  $\sigma$ , and random gene tree  $\mathcal{T}$ ,  $\mathbb{P}[\text{top}(\mathcal{T}|X) = \text{top}(\sigma|X) | \mathcal{S}]$  is arbitrarily close to 1. Formally, for fixed arbitrarily small  $\delta > 0$ , we make certain internal branches long enough so that  $1 - \delta < \mathbb{P}[\text{top}(\mathcal{T}|X) = \text{top}(\sigma|X) | \mathcal{S}] < 1$  for set  $X$ . To prove that a consensus method is a misleading estimator of  $\text{top}(\sigma)$ , it is sufficient to show that the species tree estimate on the basis of the consensus method does not display the relationship ((AB)C) in the limit as the number of gene trees goes to infinity.

There are two ancestral populations of interest, one directly ancestral to the split of species A and B at time  $\tau_3$  (denoted by  $\mathcal{A}_2$ ) and one directly ancestral to the split of species (AB) and C at time  $\tau_2$  (denoted by  $\mathcal{A}_1$ ). Both ancestral populations  $\mathcal{A}_1$  and  $\mathcal{A}_2$  contain  $D \geq 2$  demes, each of size  $N$  diploid individuals, that exchange migrants according to migration matrices  $\mathbf{M}^{(1)}$  and  $\mathbf{M}^{(2)}$ . For simplicity, both for  $i = 1$  and  $i = 2$ , we assume a symmetric island migration model in which  $\mathbf{M}_{xy}^{(i)} = m$  for each pair of distinct demes  $x^{(i)}$  and  $y^{(i)}$  in ancestral population  $\mathcal{A}_i$  (*Wakeley, 2009*). We also assume that all other ancestral populations in the  $n$ -taxon tree  $\sigma$  have only one deme (*i.e.*, they are unstructured). At time  $\tau_2$ , for each  $x = 1, 2, \dots, D$ , deme  $x^{(2)}$  in  $\mathcal{A}_2$  merges into deme  $x^{(1)}$  in  $\mathcal{A}_1$ . At time  $\tau_3$ , lineages from the  $\lambda_A$  subtree merge into deme  $j^{(2)}$  in  $\mathcal{A}_2$  and lineages from the  $\lambda_B$  subtree merge into deme  $k^{(2)} \neq j^{(2)}$  in  $\mathcal{A}_2$ . At time  $\tau_2$ , lineages from the  $\lambda_C$  subtree merge into deme  $k^{(1)}$ , the same deme into which lineages from the  $\lambda_B$  subtree, which had entered  $k^{(2)}$  in  $\mathcal{A}_2$ , merge if they have not coalesced or migrated in  $\mathcal{A}_2$ . The following list summarizes the assumptions made in this counterexample.

1. Assumptions about species tree  $\sigma$ 
  - (a) The species tree  $\sigma$  is fixed and has  $n \geq 3$  taxa.
  - (b) Certain internal branches on  $\sigma$  are sufficiently long that for random gene tree  $\mathcal{T}$ , fixed set of taxa  $X \in \{\Gamma_A, \Gamma_B, \Gamma_C\}$ , and fixed arbitrarily small  $\delta > 0$ ,  $\mathbb{P}[\text{top}(\mathcal{T}|X) = \text{top}(\sigma|X) \mid \mathcal{S}] > 1 - \delta$ .
2. Assumptions about the structure of the populations (*i.e.*,  $\mathbf{D}$ ,  $\mathbf{N}$ , and  $\mathbf{M}$ )
  - (a) All populations have one deme except for ancestral populations  $\mathcal{A}_1$  and  $\mathcal{A}_2$ , which each have a fixed equal number of demes,  $D \geq 2$ .
  - (b) Each deme has a fixed population size of  $N$  diploid individuals.
  - (c) The structured population model is an island migration model in which the per-generation backward migration rate between each pair of distinct

demes within ancestral population  $\mathcal{A}_1$  and within ancestral population  $\mathcal{A}_2$  is a fixed value  $m$ .

### 3. Assumptions about $\Psi$

- (a) At time  $\tau_2$ , for each  $x = 1, 2, \dots, D$ , deme  $x^{(2)}$  in  $\mathcal{A}_2$  merges into deme  $x^{(1)}$  in  $\mathcal{A}_1$ .
- (b) At time  $\tau_3$ , lineages from the  $\lambda_A$  subtree merge into deme  $j^{(2)}$  in  $\mathcal{A}_2$  and lineages from the  $\lambda_B$  subtree merge into deme  $k^{(2)} \neq j^{(2)}$  in  $\mathcal{A}_2$ .
- (c) At time  $\tau_2$ , lineages from the  $\lambda_C$  subtree merge into deme  $k^{(1)}$  in  $\mathcal{A}_1$ .

We have constructed the counterexample based on assumptions 1–3 so that for a specific set of taxa A, B, and C with topological relationship ((AB)C) on  $\sigma$ , we can fix  $\tau_2 - \tau_3$ ,  $D$ , and an ancestral population migration rate  $m$  such that for arbitrarily small  $\varepsilon > 0$ , the probability that a random gene tree will display the topology ((AB)C) is less than  $\varepsilon$ . For example, in Figure 7.2B, given fixed arbitrarily small  $\varepsilon > 0$ , fixed  $\tau_2 - \tau_3$  and  $D$  and fixed sufficiently small  $m$ , with probability greater than  $1 - \varepsilon$ , the lineage from A and the lineage from B will not migrate, and the lineages from B and C will coalesce before either lineage coalesces with the lineage from A. This high probability for coalescence of lineages from B and C causes a large proportion of random gene trees, greater than  $1 - \varepsilon$ , to display the nonmatching topological relationship ((BC)A).

Define an “event” as either a migration of a lineage from one deme to another deme within an ancestral population or a coalescence of two lineages. Let  $p_S(X, Y)$  be the probability under model  $\mathcal{S}(\sigma, \mathbf{D}, \mathbf{N}, \mathbf{M}, \Psi)$  that a lineage sampled from species X and a lineage sampled from species Y are in the same deme at the speciation time of X and Y. Consider three sampled lineages, one each from species A, B, and C. By construction of the counterexample,  $p_S(A, B) = 0$  because lineages from A merge into deme  $j^{(2)}$  and lineages from B merge into deme  $k^{(2)} \neq j^{(2)}$ . Within the time interval

$[\tau_3, \tau_2)$ , the time to a migration event in which the lineage from deme  $j^{(2)}$  exits the deme is exponentially distributed with rate  $(D - 1)m$  and the time to a migration event in which the lineage from deme  $k^{(2)}$  exits the deme is exponentially distributed with rate  $(D - 1)m$ . Therefore, the time to the first migration event (either from deme  $j^{(2)}$  or from deme  $k^{(2)}$ ) is exponentially distributed with rate  $(D - 1)m + (D - 1)m = 2(D - 1)m$  per generation (*Wakeley*, 2009, p. 150, eq. 5.23). Hence, the probability of zero migration events over the interval  $[\tau_3, \tau_2)$ —neither for the lineage from A nor for the lineage from B—is

$$\beta_1 = e^{-2(D-1)m(\tau_2-\tau_3)}.$$

Treating  $D$  and  $\tau_2 - \tau_3$  as fixed finite positive values, for sufficiently small migration rate  $m$ ,  $\beta_1$  is arbitrarily close to 1. Note, however, that  $m$  need not be small for  $\beta_1$  to be close to 1—for example, if  $m$  is instead a fixed finite positive value and  $\tau_2 - \tau_3$  is sufficiently small.

Many possible migration paths exist that can cause a lineage sampled from species B to be located in deme  $k^{(1)}$  of population  $\mathcal{A}_1$  (the same deme into which lineages from species C merge) at time  $\tau_2$ . For instance, there could be no migration events or there could be multiple migration events that eventually bring the lineage sampled from species B back into deme  $k^{(1)}$  at time  $\tau_2$ . Because  $\beta_1$  is the probability of only one of many possible ways of obtaining a lineage sampled from species B in deme  $k^{(1)}$  at time  $\tau_2$ , it is necessarily a lower bound for  $p_S(\text{B}, \text{C})$ . It follows that  $\beta_1 < p_S(\text{B}, \text{C}) < 1$ , and similarly, a bound can be placed on  $p_S(\text{A}, \text{C})$  such that  $0 < p_S(\text{A}, \text{C}) < 1 - \beta_1$ . Hence,  $p_S(\text{A}, \text{C})$  is arbitrarily close to 0 and  $p_S(\text{B}, \text{C})$  is arbitrarily close to 1 for sufficiently small migration rate  $m$  holding  $\tau_2 - \tau_3$  fixed, or for sufficiently small  $\tau_2 - \tau_3$  holding  $m$  fixed. Because the lineages from A and B are in different demes at time  $\tau_3$ ,  $p_S(\text{A}, \text{B}) = 0$ .

Let  $P_S[\mathcal{T}]$  denote the probability of gene tree topology  $\mathcal{T}$  under model  $\mathcal{S}(\sigma, \mathbf{D}, \mathbf{N}, \mathbf{M}, \Psi)$ . If no migration event occurs in the interval  $[\tau_3, \tau_2)$ , then the

lineage from species A is in deme  $j^{(1)}$  and the lineages from species B and C are in deme  $k^{(1)} \neq j^{(1)}$  at time  $\tau_2$ . Within the time interval  $[\tau_2, \infty)$ , the time to the first migration event that causes the lineage from deme  $j^{(1)}$  to migrate is exponentially distributed with rate  $(D - 1)m$ , the time to the first migration event that causes one of the two lineages from deme  $k^{(1)}$  to migrate is exponentially distributed with rate  $2(D - 1)m$ , and the time to the event in which the two lineages from deme  $k^{(2)}$  coalesce is exponentially distributed with rate  $1/(2N)$ . Therefore, the time to the first event (migration or coalescence) on the interval  $[\tau_2, \infty)$  is exponentially distributed with rate  $(D - 1)m + 2(D - 1)m + 1/(2N) = 3(D - 1)m + 1/(2N)$  per generation (*Wakeley, 2009, p. 150, eq. 5.23*). Hence, the probability that the first event in the interval  $[\tau_2, \infty)$  is a coalescence between the lineages from species B and C is  $[1/(2N)]/[3(D - 1)m + 1/(2N)]$ . Treating the parameters  $D$  and  $N$  as fixed finite positive values, the probability that the first event in the interval  $[\tau_2, \infty)$  is a coalescence between the lineages from species B and C is arbitrarily close to 1 for sufficiently small migration rate  $m$ . Multiplying by the probability  $\beta_1$  of observing zero migration events on the interval  $[\tau_3, \tau_2)$ , we obtain a lower bound on the probability,  $\beta_2$ , that the first event on the interval  $[\tau_3, \infty)$  is a coalescence event between the lineages from species B and C

$$\beta_2 = \frac{1/(2N)}{3(D - 1)m + 1/(2N)}\beta_1 = \frac{1}{6(D - 1)Nm + 1}\beta_1. \quad (7.1)$$

This probability  $\beta_2$  is arbitrarily close to 1 for sufficiently small migration rate  $m$ . Note, however, that  $m$  may not need to be too small for a coalescence between lineages from species B and C to be the most probable first event on the interval  $[\tau_2, \tau_3)$ . For example, if  $\tau_2 - \tau_3$  is sufficiently small, then  $\beta_1$  is arbitrarily close to 1. Because  $P_S[((BC)A)] \geq \beta_2$ , for some constant  $1/(c + 1)$  with  $c > 0$ , to get  $P_S[((BC)A)] > 1/(c + 1)$ , assuming  $\beta_1$  is sufficiently close to 1, we would only need



$$m \approx c/[6(D - 1)N].$$

Our counterexample (assumptions 1–3) together with certain parameter values chosen such that  $\beta_2$  is arbitrarily close to 1 provides a case in which gene trees strongly support clades that are not present on species tree  $\sigma$ . As we will see in the next section, the particular gene tree distribution provided by specific choices for  $\mathcal{S}(\sigma, \mathbf{D}, \mathbf{N}, \mathbf{M}, \Psi)$  causes a large class of consensus methods to infer species trees with clades not present on  $\sigma$ . This lack of concordance between the inferred species tree topology and  $\sigma$  occurs when the parameters  $\sigma$ ,  $\mathbf{D}$ ,  $\mathbf{N}$ ,  $\mathbf{M}$ , and  $\Psi$  are in the space defined by assumptions 1–3, when  $\tau_2 - \tau_3$  and  $D$  are fixed, and when fixed  $m$  is sufficiently small.

## 7.2 Consistency and inconsistency of methods

In this section, under the multispecies coalescent model with ancestral population structure, we investigate the statistical consistency of consensus methods that are based on seven different criteria for inferring species tree topologies. The seven methods involve using a uniquely favored topology (Democratic Vote), using average coalescence times (STEAC), using average ranks of coalescences (STAR), using uniquely favored rooted triples (R\* Consensus and Rooted Triple Consensus), minimizing the number of deep coalescences (MDC), taking the majority-rule (Majority-Rule Consensus), and using minimum coalescence times (GLASS/Maximum Tree). We show, through the use of the counterexample developed in the previous section, that consensus methods based on six of the seven criteria are misleading. We also provide a proof that consensus methods that use the minimum coalescence times criterion are statistically consistent. The proofs that Democratic Vote is misleading (Theorem VII.1) and that GLASS/Maximum Tree is consistent (Theorem VII.7) are provided in the main text. The proofs that the other consensus methods are misleading (Theorems VII.2-VII.6) are similar and for

completeness, they are provided in the Appendix.

### 7.2.1 Uniquely favored topology

An intuitive approach to the inference of species tree topologies is to use the Democratic Vote consensus method. Democratic Vote estimates a species tree topology using the most frequently occurring gene tree topology in a sample of gene trees (*Degnan and Rosenberg, 2009*). Discordant gene tree topologies with greater probability than the matching topology have been termed “anomalous gene trees” (AGTs), and the space of branch lengths in which AGTs arise has been termed the “anomaly zone” (*Degnan and Rosenberg, 2006*). Because of the existence of AGTs, and because of gene tree discordance more generally, it is difficult for consensus methods to achieve statistical consistency (*Degnan et al., 2009*). Under the multispecies coalescent model with no ancestral population structure, the space in which Democratic Vote is misleading corresponds exactly to the anomaly zone (*Degnan and Rosenberg, 2006*). A consequence of this direct correspondence between Democratic Vote and the anomaly zone is that Democratic Vote is a statistically consistent estimator of a species tree topology only for three-taxon species trees and for four-taxon species trees with a symmetric topology.

*Slatkin and Pollack* (2008) showed that the most likely gene tree topology does not necessarily match the species tree topology for three-taxon species trees under a specific multispecies coalescent model with ancestral population structure. This result implies that in structured ancestral population models, Democratic Vote can be misleading for three-taxon species tree topologies. Our general structured ancestral population model,  $\mathcal{S}(\sigma, \mathbf{D}, \mathbf{N}, \mathbf{M}, \Psi)$ , contains the model of *Slatkin and Pollack* (2008) as a special case. Under this general model, we use the counterexample (assumptions 1–3 in the “Counterexample” section) to show that Democratic Vote is a misleading estimator for the topology of fixed species tree  $\sigma$  with  $n \geq 3$  taxa under model

$\mathcal{S}(\sigma, \mathbf{D}, \mathbf{N}, \mathbf{M}, \Psi)$ .

To provide intuition as to why Democratic Vote is misleading, assuming the species tree has topology  $((\lambda_A \lambda_B) \lambda_C)$ , note that under the counterexample, if we fix  $\tau_2 - \tau_3$  and  $D$  and set the migration rate  $m$  sufficiently small, then gene trees generated under model  $\mathcal{S}(\sigma, \mathbf{D}, \mathbf{N}, \mathbf{M}, \Psi)$  display a nonmatching topology with probability arbitrarily close to 1. Because of this large probability, the most frequently occurring gene tree topology—the Democratic Vote topology—is  $((\lambda_B \lambda_C) \lambda_A)$  instead of  $((\lambda_A \lambda_B) \lambda_C)$ . Thus, Democratic Vote is a misleading estimator for the species tree topology under model  $\mathcal{S}(\sigma, \mathbf{D}, \mathbf{N}, \mathbf{M}, \Psi)$ .

Formally, Let  $\hat{P}[\mathcal{T}]$  denote the sample proportion of topology  $\mathcal{T}$  in a set of  $L$  gene trees.

**Theorem VII.1.** *Consider a species tree  $\sigma$  with  $n \geq 3$  taxa under model  $\mathcal{S}(\sigma, \mathbf{D}, \mathbf{N}, \mathbf{M}, \Psi)$ . Further, consider a consensus method  $\hat{C}_L$  that estimates  $\text{top}(\sigma)$  from a set of  $L$  gene trees using the most frequently occurring gene tree topology. Then  $\hat{C}_L$  is a misleading estimator of  $\text{top}(\sigma)$ .*

*Proof.* We use the counterexample (assumptions 1–3 in the “Counterexample” section) to show that  $\hat{C}_L$  is misleading. For  $\hat{C}_L$  to not be misleading, we must have that  $\hat{C}_L \xrightarrow{P} \text{top}(\sigma)$  as  $L \rightarrow \infty$ . Note that  $\text{top}(\sigma) = ((\lambda_A \lambda_B) \lambda_C)$ . Consider an alternative species tree  $\sigma^*$  with topology  $\text{top}(\sigma^*) = ((\lambda_B \lambda_C) \lambda_A)$ . Set the migration rate  $m$  sufficiently small such that  $P_S[\text{top}(\sigma^*)] = P_S[((\lambda_B \lambda_C) \lambda_A)] > (1 - \delta)^3 \beta_2$ , which is arbitrarily close to 1. Using the Law of Large Numbers,  $\hat{P}[\text{top}(\sigma^*)] \xrightarrow{P} P_S[\text{top}(\sigma^*)]$  as  $L \rightarrow \infty$ . Because  $P_S[\text{top}(\sigma^*)]$  is arbitrarily close to 1,  $\text{top}(\sigma^*) \neq \text{top}(\sigma)$  is the uniquely favored topology and so  $\mathbb{P}[\hat{C}_L = \text{top}(\sigma^*) | \mathcal{S}] \rightarrow 1$  as  $L \rightarrow \infty$ . Therefore,  $\hat{C}_L \xrightarrow{P} \text{top}(\sigma^*)$ , and  $\hat{C}_L$  is a misleading estimator of  $\text{top}(\sigma)$  under  $\mathcal{S}(\sigma, \mathbf{D}, \mathbf{N}, \mathbf{M}, \Psi)$ .  $\square$

### 7.2.2 Average coalescence times

Consider a sample of  $L$  independent loci and let  $T_{XY}^\ell$  denote the random coalescence time in the gene tree of locus  $\ell$  for a lineage sampled from species X and a lineage sampled from species Y. Define the average random coalescence time across  $L$  loci between one lineage sampled from species X and one lineage sampled from species Y as  $\bar{t}_{XY} = (1/L) \sum_{\ell=1}^L T_{XY}^\ell$ . *Liu et al.* (2009) developed the consensus method STEAC, which utilizes the average coalescence times  $\bar{t}_{XY}$ , considering each distinct pair of species X and Y, to infer a species tree. The average time  $\bar{t}_{XY}$  provides a distance between species X and Y. STEAC creates a distance matrix for all pairs of species, including an outgroup species, and infers a species tree using neighbor-joining. The outgroup is then used to root the tree. STEAC is a statistically consistent estimator of a species tree topology when gene trees are distributed according to the multispecies coalescent (*Liu et al.*, 2009). This consistency stems from the statistical consistency of neighbor-joining (*Atteson*, 1999) and the observation that under the multispecies coalescent model, for species X, Y, and Z, if the divergence time of species X and Y is smaller than that for X and Z and for Y and Z, then the expected coalescence time is smaller for lineages from X and Y than for lineages from X and Z and for lineages from Y and Z. We show that in the presence of ancestral population structure, STEAC is a misleading estimator for the topology of fixed species tree  $\sigma$  with  $n \geq 3$  taxa under model  $\mathcal{S}(\sigma, \mathbf{D}, \mathbf{N}, \mathbf{M}, \Psi)$ .

To provide intuition as to why STEAC is misleading, assuming the species tree has topology  $((\lambda_A \lambda_B) \lambda_C)$ , we can fix  $\tau_2 - \tau_3$  and  $D$  and set the migration rate  $m$  sufficiently small such that the lineages from B and C very likely coalesce more recently than either coalesces with the lineage from A. Thus, the average coalescence time for a pair of lineages, one sampled from B and one sampled from C, will be smaller than the average coalescence time for a pair of lineages, one sampled from A and one sampled from B. Consequently, because of how STEAC uses expected coalescence times to

estimate a species tree topology, STEAC is misleading.

**Theorem VII.2.** *Consider a species tree  $\sigma$  with  $n \geq 3$  taxa under model  $\mathcal{S}(\sigma, \mathbf{D}, \mathbf{N}, \mathbf{M}, \Psi)$ . Define the rule  $\mathcal{R}$  such that for species  $X$ ,  $Y$ , and  $Z$ ,  $X$  and  $Y$  join more recently in the past than do species  $X$  and  $Z$  and species  $Y$  and  $Z$  when  $\bar{t}_{XY} < \bar{t}_{XZ}$  and  $\bar{t}_{XY} < \bar{t}_{YZ}$ . Consider a consensus method  $\hat{C}_L$  that estimates  $\text{top}(\sigma)$  from a set of  $L$  gene trees using average coalescence times  $\bar{t}_{XY}$  for all distinct pairs of species  $X$  and  $Y$  according to rule  $\mathcal{R}$ . Then  $\hat{C}_L$  is a misleading estimator of  $\text{top}(\sigma)$ .*

### 7.2.3 Average ranks of coalescences

Coalescence ranks describe the relative order of internal nodes in a rooted tree topology. A ranking is an integer assignment with  $n$  for the root, and monotonically decreasing with ancestor/descendent relationships. The minimum rank in a tree is 2. There are several ways of assigning the ranks. For example, for the topology  $((((AB)C)D)$ , the root node has rank 4, the node connecting clade  $\{AB\}$  to  $C$  has rank 3, and the node connecting  $A$  and  $B$  has rank 2 (Figure 7.3A). For the topology  $((AB)(CD))$ , the root node has rank 4. Three possible configurations of ranks exist for the other internal nodes:

- The node connecting  $A$  and  $B$  has rank 2, and the node connecting  $C$  and  $D$  has rank 2 (Figure 7.3B) In this ranking, the rank of an internal node is the number of leaves descending from it.
- The node connecting  $A$  and  $B$  has rank 3, and the node connecting  $C$  and  $D$  has rank 3 (Figure 7.3C) In this ranking, the rank of an internal node is the rank of the node directly ancestral to it minus 1. This ranking is the ranking used by STAR (Liu et al., 2009).
- The node connecting  $A$  and  $B$  has rank 2, and the node connecting  $C$  and  $D$  has rank 3 (Figure 7.3D) In this ranking, the rank of an internal node is the rank

relative to all other internal nodes in the tree. Each possible value for a rank (*i.e.*, ranks  $2, 3, \dots, n$ ) is used once. This is the common method for assigning ranks to nodes in a tree.

The examples of coalescence ranks in Figure 7.3 all share the property that the rank of an internal node  $y$  is larger than the rank of a different internal node  $x$ , if node  $y$  lies along the path to the root from node  $x$ . This property defines what we mean by a rank.

STAR, a consensus method developed by *Liu et al.* (2009), assumes that the rank of the root node in an  $n$ -taxon tree is  $n$ . Then, descending toward the leaf nodes, internal nodes are assigned the rank of the node directly ancestral to it minus 1. Consider a sample of  $L$  independent loci and let  $R_{XY}^\ell$  denote the random coalescence rank in the gene tree of locus  $\ell$  for a lineage sampled from species X and a lineage sampled from species Y. Denote the random average coalescence rank across  $L$  loci between a lineage sampled from species X and a lineage sampled from species Y as  $\bar{r}_{XY} = (1/L) \sum_{\ell=1}^L R_{XY}^\ell$ . The STAR consensus method utilizes the average ranks of coalescences  $\bar{r}_{XY}$ , for each distinct pair of species X and Y, to infer a species tree. The average rank  $\bar{r}_{XY}$  provides a distance between species X and Y. Analogous to the procedure for STEAC, STAR creates a distance matrix for all pairs of species (including an outgroup) and infers a species tree using neighbor-joining. The outgroup is then used to root the tree. STAR is a statistically consistent estimator of a species tree topology when gene trees are distributed according to the multispecies coalescent (*Liu et al.*, 2009). This consistency stems from the statistical consistency of neighbor-joining and the observation that under the multispecies coalescent model, for species X, Y, and Z, if the divergence time of species X and Y is smaller than that for X and Z and for Y and Z, then the expected rank in the gene tree is smaller for the coalescence of lineages from X and Y than for the coalescence of lineages from X and Z and for the coalescence of lineages from Y and Z. We show that in the presence

of ancestral population structure, STAR is a misleading estimator for the topology of fixed species tree  $\sigma$  with  $n \geq 3$  taxa under model  $\mathcal{S}(\sigma, \mathbf{D}, \mathbf{N}, \mathbf{M}, \Psi)$ .

Similar to STEAC, assuming the species tree has topology  $((\lambda_A \lambda_B) \lambda_C)$ , we can fix  $\tau_2 - \tau_3$  and  $D$  and set the migration rate  $m$  sufficiently small such that the average coalescence rank for a pair of lineages, one sampled from B and one sampled from C, will be smaller than the average coalescence rank for a pair of lineages, one sampled from A and one sampled from B. Because STAR uses average coalescence ranks in the same way that STEAC uses average coalescence times to infer species tree topologies, STAR is misleading.

**Theorem VII.3.** *Consider a species tree  $\sigma$  with  $n \geq 3$  taxa under model  $\mathcal{S}(\sigma, \mathbf{D}, \mathbf{N}, \mathbf{M}, \Psi)$ . Define the rule  $\mathcal{R}$  such that for species  $X$ ,  $Y$ , and  $Z$ ,  $X$  and  $Y$  join more recently in the past than do species  $X$  and  $Z$  and species  $Y$  and  $Z$  when  $\bar{r}_{XY} < \bar{r}_{XZ}$  and  $\bar{r}_{XY} < \bar{r}_{YZ}$ . Consider a consensus method  $\hat{C}_L$  that estimates  $\text{top}(\sigma)$  from a set of  $L$  gene trees using average ranks of coalescences  $\bar{r}_{XY}$  for all distinct pairs of species  $X$  and  $Y$  according to rule  $\mathcal{R}$ . Then  $\hat{C}_L$  is a misleading estimator of  $\text{top}(\sigma)$ .*

#### 7.2.4 Uniquely favored rooted triples

Define a uniquely favored rooted triple among a set of three taxa  $X$ ,  $Y$ , and  $Z$  as the rooted topological relationship among  $X$ ,  $Y$ , and  $Z$  with the largest frequency in a sample of rooted gene trees. Because AGTs do not exist for three-taxon species trees under the multispecies coalescent model, consensus methods have been developed that infer species trees based on the topologies of uniquely favored rooted triples. These consensus methods are  $R^*$  Consensus (Bryant, 2003; Degnan et al., 2009) and Rooted Triple Consensus (Ewing et al., 2008).  $R^*$  Consensus constructs a species tree from uniquely favored rooted triples through an exact algorithm. Following Degnan et al. (2009), the set  $\mathcal{K}$  is a clade in the  $R^*$  Consensus tree if for each distinct pair of taxa

$X', X'' \in \mathcal{K}$  and every taxon  $Z \notin \mathcal{K}$ ,  $((X'X'')Z)$  is a uniquely favored rooted triple. The Rooted Triple Consensus tree is constructed with a heuristic algorithm that combines the  $\binom{n}{3}$  uniquely favored rooted triples using the tree puzzle heuristic (*Ewing et al.*, 2008). *Degnan et al.* (2009) proved that  $R^*$  Consensus is a statistically consistent estimator of a species tree topology when gene trees are distributed according to the multispecies coalescent model. We show that in the presence of ancestral population structure,  $R^*$  Consensus and Rooted Triple Consensus are misleading estimators for the topology of fixed species tree  $\sigma$  with  $n \geq 3$  taxa under model  $\mathcal{S}(\sigma, \mathbf{D}, \mathbf{N}, \mathbf{M}, \Psi)$ .

Analogous to the case of Democratic Vote, assuming the species tree has topology  $((\lambda_A \lambda_B) \lambda_C)$ , we can fix  $\tau_2 - \tau_3$  and  $D$  and set the migration rate  $m$  sufficiently small such that the probability that a gene tree displays topology  $((\lambda_B \lambda_C) \lambda_A)$  is arbitrarily close to 1. From this high probability, each rooted triple displayed by  $((\lambda_B \lambda_C) \lambda_A)$  is a uniquely favored rooted triple as the number of loci tends to infinity. Because a rooted binary tree topology is defined by its set of rooted triples (*Steel*, 1992, Proposition 4), each of the rooted triples are in the  $R^*$  and Rooted Triple Consensus trees. In particular, the  $R^*$  Consensus and Rooted Triple Consensus trees are  $((\lambda_B \lambda_C) \lambda_A)$ . Consequently,  $R^*$  Consensus and Rooted Triple Consensus are misleading.

**Theorem VII.4.** *Consider a species tree  $\sigma$  with  $n \geq 3$  taxa under model  $\mathcal{S}(\sigma, \mathbf{D}, \mathbf{N}, \mathbf{M}, \Psi)$ . Define the rule  $\mathcal{R}$  such that for species  $X, Y$ , and  $Z$ ,  $X$  and  $Y$  join more recently in the past than do species  $X$  and  $Z$  and species  $Y$  and  $Z$  when  $\hat{P}[(XY)Z] > \hat{P}[(XZ)Y]$  and  $\hat{P}[(XY)Z] > \hat{P}[(YZ)X]$ . Consider a consensus method  $\hat{C}_L$  that estimates  $\text{top}(\sigma)$  from a set of  $L$  gene trees using uniquely favored rooted triples according to rule  $\mathcal{R}$ . Then  $\hat{C}_L$  is a misleading estimator of  $\text{top}(\sigma)$ .*

### 7.2.5 Minimizing deep coalescences

Another sensible approach to inferring species trees from gene trees in the presence of incomplete lineage sorting is to minimize the number of deep coalescences



(Maddison, 1997). A coalescence event for species X and Y is called “deep” if the event does not occur in the population directly ancestral to the split of species X and Y in the species tree. The MDC criterion seeks to find a species tree that minimizes the number of lineages that do not coalesce in the first population in which they have the opportunity to find a common ancestor. Recently, *Than and Nakhleh* (2009) presented an exact method to infer a species tree from gene trees using the MDC criterion. A subsequent study showed that when gene trees are distributed according to the multispecies coalescent model, MDC is a misleading estimator of a species tree topology for four-taxon asymmetric species trees and for species trees with five or more taxa (*Than and Rosenberg*, 2011). In this section, we provide a theorem (proven in the Appendix) which states that in the presence of ancestral population structure, MDC is a misleading estimator for the topology of fixed species tree  $\sigma$  with  $n \geq 3$  taxa under model  $\mathcal{S}(\sigma, \mathbf{D}, \mathbf{N}, \mathbf{M}, \Psi)$ .

Assuming the species tree has topology  $((\lambda_A \lambda_B) \lambda_C)$ , fix  $\tau_2 - \tau_3$  and  $D$  and set the migration rate  $m$  sufficiently small such that gene trees display topology  $((\lambda_B \lambda_C) \lambda_A)$  with probability arbitrarily close to 1. The number of extra lineages that are needed to reconcile a gene tree with topology  $((\lambda_B \lambda_C) \lambda_A)$  and species trees with topologies  $((\lambda_A \lambda_B) \lambda_C)$  and  $((\lambda_B \lambda_C) \lambda_A)$  is one and zero, respectively. Because the probability of observing a gene tree with topology  $((\lambda_B \lambda_C) \lambda_A)$  is high, then the species tree with topology  $((\lambda_B \lambda_C) \lambda_A)$  will minimize the number of deep coalescences (*i.e.*, the number of extra lineages needed to reconcile the set of gene tree topologies with the species tree topology). Because  $((\lambda_B \lambda_C) \lambda_A)$  does not match the species tree topology, MDC is misleading.

**Theorem VII.5.** *Consider a species tree  $\sigma$  with  $n \geq 3$  taxa under model  $\mathcal{S}(\sigma, \mathbf{D}, \mathbf{N}, \mathbf{M}, \Psi)$ . Further, consider a consensus method  $\widehat{C}_L$  that estimates  $\text{top}(\sigma)$  from a set of  $L$  gene trees by minimizing the number of deep coalescences. Then  $\widehat{C}_L$  is a misleading estimator of  $\text{top}(\sigma)$ .*

### 7.2.6 Majority-rule

One widely used consensus method, termed Majority-Rule Consensus, constructs a species tree using only clades that appear with a frequency greater than some fixed  $\alpha$ ,  $\alpha \in [0.5, 1)$  (Bryant, 2003). The Majority-Rule Consensus tree is either resolved (bifurcating) or unresolved (multifurcating), partially unresolved, or fully unresolved. For the case of  $\alpha = 0.5$ , Majority-Rule Consensus has been shown to be a statistically inconsistent, but not misleading, estimator of a species tree topology when gene trees are distributed according to the multispecies coalescent (Degnan *et al.*, 2009). In this section, we provide a theorem (proven in the Appendix) which states that in the presence of ancestral population structure, Majority-Rule Consensus is a misleading estimator for the topology of fixed species tree  $\sigma$  with  $n \geq 3$  taxa under model  $\mathcal{S}(\sigma, \mathbf{D}, \mathbf{N}, \mathbf{M}, \Psi)$ .

Assuming the species tree has topology  $((\lambda_A \lambda_B) \lambda_C)$ , by fixing  $\tau_2 - \tau_3$  and  $D$  and setting the migration rate  $m$  sufficiently small such that gene trees display topology  $((\lambda_B \lambda_C) \lambda_A)$  with probability arbitrarily close to 1, all clades on  $((\lambda_B \lambda_C) \lambda_A)$  appear with frequency greater than fixed  $\alpha$ . All clades with frequency greater than  $\alpha$  appear on the Majority-Rule Consensus tree. Consequently, Majority-Rule Consensus is misleading.

**Theorem VII.6.** *Consider a species tree  $\sigma$  with  $n \geq 3$  taxa under model  $\mathcal{S}(\sigma, \mathbf{D}, \mathbf{N}, \mathbf{M}, \Psi)$ . Further, consider a consensus method  $\widehat{C}_L$  that estimates  $\text{top}(\sigma)$  from a set of  $L$  gene trees by only using clades present with a frequency greater than fixed  $\alpha$ ,  $\alpha \in [0.5, 1)$ . Then  $\widehat{C}_L$  is a misleading estimator  $\text{top}(\sigma)$ .*

### 7.2.7 Minimum coalescence time

Consider a sample of  $L$  independent loci. Define the minimum coalescence time across  $L$  loci between one lineage sampled from species X and one lineage sampled from species Y as  $t_{XY}^{\min} = \min_{\ell=1, \dots, L} T_{XY}^{\ell}$ . The final method we examine is one that uses the

minimum coalescence time  $t_{XY}^{\min}$ , considering each distinct pair of species  $X$  and  $Y$ , to infer a species tree. GLASS (*Mossel and Roch, 2010*) and Maximum Tree (*Liu et al., 2010*) are two names for the same method that constructs species trees using these minimum coalescence times. The minimum time  $t_{XY}^{\min}$  provides a distance between species  $X$  and  $Y$ . GLASS/Maximum Tree creates a distance matrix for all pairs of species and infers a species tree using single-linkage clustering. GLASS/Maximum Tree is a statistically consistent estimator of a species tree topology when gene trees are distributed according to the multispecies coalescent (*Mossel and Roch, 2010; Liu et al., 2010*). In this section, we show that in the presence of ancestral population structure according to our model  $\mathcal{S}(\sigma, \mathbf{D}, \mathbf{N}, \mathbf{M}, \Psi)$ , GLASS/Maximum Tree is a statistically consistent estimator for the topology of fixed species tree  $\sigma$  with  $n \geq 3$  taxa under model  $\mathcal{S}(\sigma, \mathbf{D}, \mathbf{N}, \mathbf{M}, \Psi)$ .

To provide intuition as to why GLASS/Maximum Tree is consistent, note that an assumption of model  $\mathcal{S}(\sigma, \mathbf{D}, \mathbf{N}, \mathbf{M}, \Psi)$  is that demes within an ancestral population communicate through a path of nonzero migration. Because of this communication between demes, as the number of gene trees grows large, for some gene tree a pair of lineages sampled from distinct species will likely coalesce in the population directly ancestral to the divergence of those species. Because GLASS/Maximum Tree uses minimum coalescence times to estimate a species tree topology, as the number of gene trees grows large, the single-linkage clustering algorithm applied to these minimum coalescence times will yield a tree topology that matches  $\text{top}(\sigma)$ . Therefore, GLASS/Maximum Tree is a statistically consistent estimator for the species tree topology under model  $\mathcal{S}(\sigma, \mathbf{D}, \mathbf{N}, \mathbf{M}, \Psi)$ .

**Theorem VII.7.** *Consider a species tree  $\sigma$  with  $n \geq 3$  taxa under model  $\mathcal{S}(\sigma, \mathbf{D}, \mathbf{N}, \mathbf{M}, \Psi)$ . Further, consider a consensus method  $\widehat{C}_L$  that estimates  $\text{top}(\sigma)$  from a set of  $L$  gene trees using single-linkage clustering applied to the set of minimum coalescence times  $t_{XY}^{\min}$  for each distinct pair of species  $X$  and  $Y$ . Then  $\widehat{C}_L$  is a*

statistically consistent estimator of  $\text{top}(\sigma)$ .

*Proof.* This proof is similar to that of *Mossel and Roch (2010)* for GLASS. Suppose we have  $L$  independent loci. For  $\widehat{C}_L$  to be consistent, we must have that  $\widehat{C}_L \xrightarrow{P} \text{top}(\sigma)$  as  $L \rightarrow \infty$ . Let  $b$  be the length of the shortest internal branch in species tree  $\sigma$ . Fix the species tree  $\sigma$  and fix the structured ancestral population model  $\mathcal{S}(\sigma, \mathbf{D}, \mathbf{N}, \mathbf{M}, \Psi)$ . Define  $f_k$ ,  $k = 1, 2, \dots, n - 1$ , as the probability under model  $\mathcal{S}(\sigma, \mathbf{D}, \mathbf{N}, \mathbf{M}, \Psi)$  that, going back in time, no coalescence occurs between any pair of lineages within  $b$  generations from entering ancestral population  $A_k$ . Then over the set of  $L$  gene trees, the probability that, going back in time, no coalescence occurs between any pair of lineages within  $b$  generations from entering  $A_k$  is  $(f_k)^L$ . It follows that over the set of  $L$  gene trees, the probability that, going back in time, at least one coalescence occurs between each pair of lineages within  $b$  generations from entering  $A_k$  is  $1 - (f_k)^L$ . Therefore, over the set of  $L$  gene trees and the set of ancestral populations, the probability that, going back in time, at least one coalescence occurs between each pair of lineages within  $b$  generations from entering each of the  $n - 1$  ancestral populations is

$$f_{\min} = \prod_{k=1}^{n-1} [1 - (f_k)^L].$$

Because the demes in  $\mathcal{S}(\sigma, \mathbf{D}, \mathbf{N}, \mathbf{M}, \Psi)$  communicate, we have  $0 < f_k < 1$  for  $k = 1, 2, \dots, n - 1$ . It follows that  $f_{\min} \rightarrow 1$  as  $L \rightarrow \infty$ . Consequently, as  $L \rightarrow \infty$ , for each pair of lineages sampled from a pair of species, the minimum coalescence time for those lineages lies within the population directly ancestral to the split of the two species. Hence, applying single-linkage clustering to the set of  $t_{XY}^{\min}$  for all distinct pairs of species  $X$  and  $Y$  yields  $\text{top}(\sigma)$ , and so  $\mathbb{P}[\widehat{C}_L = \text{top}(\sigma) | \mathcal{S}] \rightarrow 1$  as  $L \rightarrow \infty$ . Therefore,  $\widehat{C}_L \xrightarrow{P} \text{top}(\sigma)$ , and  $\widehat{C}_L$  is a statistically consistent estimator of  $\text{top}(\sigma)$  under  $\mathcal{S}(\sigma, \mathbf{D}, \mathbf{N}, \mathbf{M}, \Psi)$ .  $\square$

## 7.3 Simulations

### 7.3.1 Performance of methods on true gene trees

#### 7.3.1.1 Simulation procedure

To examine how robust the eight consensus methods are to ancestral population structure, we evaluated the performance of the methods using simulations. These simulations enable us to investigate the performance of the consensus methods on a finite number of loci, rather than solely their asymptotic behavior. The consensus methods we investigated are Democratic Vote, STEAC, STAR, R\* Consensus, Rooted Triple Consensus, MDC, Majority-Rule Consensus (with  $\alpha = 0.5$ ), and GLASS/Maximum Tree. We used the three-taxon species tree  $\sigma = ((A:1.0, B:1.0):0.1, C:1.1)$  illustrated in Figure 7.4A. The ancestral populations each follow an island migration model with  $D = 10$  demes and a scaled migration rate between demes of  $M = 4N_e m$ , where  $N_e$  is a reference effective number of diploid individuals in a population. Note that because both time and migration rate are scaled by the same effective population size  $N_e$ , the specific value of  $N_e$  does not matter. Because we are assuming an island migration model, within each ancestral population, for all  $i \neq j \in \{1, 2, \dots, 10\}$ , the migration rate from deme  $i$  to deme  $j$  is  $M$ . Time in our model is measured in coalescent units  $T = t/(2N_e)$ , where  $t$  is measured in generations. Going back in time, the lineage from species A merges into deme 1 in ancestral population  $\mathcal{A}_2$ , the lineage from species B merges into deme 10 of ancestral population  $\mathcal{A}_2$ , and the lineage from species C merges into deme 10 of ancestral population  $\mathcal{A}_1$ . At time  $\tau_2$ , lineages in deme  $x^{(2)}$  of  $\mathcal{A}_2$  merge into deme  $x^{(1)}$  of  $\mathcal{A}_1$  for each  $x = 1, 2, \dots, 10$ . This model is precisely the model used for the counterexample within the Theory section with the number of demes set to 10.

We generated gene trees for  $L = 100, 200, \dots, 2000$  independent loci (with each set of  $L$  gene trees generated independent of each other set of gene trees) using the

coalescent simulator MS (*Hudson, 2002*), which enables the simulation of gene trees given a species tree model. These  $L$  gene trees were then used as input to each of the consensus methods (each of which was applied to the same set of  $L$  gene trees), and a species tree estimate was obtained as output. We repeated this process for a total of 1000 independent replicate simulations of  $L$  loci.

### 7.3.1.2 Results

The results for these simulations are displayed in Figure 7.4B. For scaled migration rate  $M = 10.0$ , the tree topology with greatest support for every consensus method except for Majority-Rule Consensus is ((AB)C), which matches the species tree. Majority-Rule Consensus instead provides greatest support for the star phylogeny (ABC), reaching a frequency of 1.0 by 200 gene trees. This result for Majority-Rule applied to three-taxon gene trees is not surprising because Majority-Rule Consensus will return an unresolved three-taxon topology if there does not exist an input gene tree topology with frequency greater than 0.5 (*Degnan et al., 2009*). Because the internal branch length is small (0.1 coalescent units) and because the migration rate between demes is large ( $M = 10.0$ ), it is unlikely that any three-taxon gene tree will have frequency greater than 0.5 as the number of input gene trees gets large. The method that performs best is GLASS/Maximum Tree, reaching probability 1 of estimating species tree topology ((AB)C) by 800 gene trees. Although the other six consensus methods provide the strongest support to ((AB)C) at 2000 gene trees, the methods still have low support for ((AB)C), with frequencies of  $\sim 0.55$  for STAR,  $\sim 0.54$  for Democratic Vote, Rooted Triple Consensus and MDC,  $\sim 0.53$  for R\* Consensus, and  $\sim 0.49$  for STEAC.

Decreasing the migration rate to  $M = 1.0$ , we find, as with the case for  $M = 10.0$ , that GLASS/Maximum Tree has highest support for topology ((AB)C). GLASS/Maximum Tree also takes longer compared to the case for  $M = 10.0$  to

converge to the correct topology, reaching a frequency of 1.0 for the ((AB)C) topology with 1900 genes instead of 800 genes with the case of  $M = 10.0$ . As with the case of  $M = 10.0$ , Majority-Rule Consensus provides greatest support for (ABC), reaching a frequency of 1.0 by 200 genes. In contrast to the results for  $M = 10.0$ , we find that the other six consensus methods no longer have their highest support for the correct tree topology. Instead, the most favored topology is ((BC)A), reaching a frequency at 2000 gene trees of  $\sim 0.99$  for Democratic Vote, STAR, R\* Consensus, Rooted Triple Consensus, and MDC and a frequency of  $\sim 0.96$  for STEAC. By construction of the simulation, with sufficiently small migration, we would expect that each of the consensus methods (except for GLASS/Maximum Tree) would infer the topology ((BC)A) with highest frequency.

Reducing the migration rate to  $M = 0.1$ , we find that GLASS/Maximum Tree continues to support the correct species tree topology ((AB)C). Unlike for the two higher migration rates, GLASS/Maximum Tree does not infer the correct topology with a frequency of 1.0 by 2000 gene trees, obtaining ((AB)C) with frequency  $\sim 0.64$  at 2000 gene trees. However, the frequency of ((AB)C) when inferred by GLASS/Maximum Tree increases as a function of the number of gene trees. Consequently, we expect that the frequency would approach 1.0 with enough gene trees, as Theorem VII.7 predicts. Indeed, as expected, the other seven consensus methods provide highest support to the topology ((BC)A) with a frequency of 1.0 for all sets of  $L = 100, 200, \dots, 2000$  gene trees tested. Majority-Rule gives greatest support for the ((BC)A) topology instead of the (ABC) topology as in the cases for  $M = 1.0$  and  $M = 10.0$  because, when the migration rate is sufficiently small ( $M = 0.1$ ), the probability is far greater than 0.5 of a gene tree displaying topology ((BC)A).

### 7.3.2 GLASS/Maximum Tree from inferred gene trees

#### 7.3.2.1 Simulation procedure

The theoretical and simulation results for GLASS/Maximum Tree presented in the previous sections has only incorporated genealogical discordance due to the stochasticity of the coalescent process. However, an additional form of stochasticity that can cause genealogical discordance is mutation. To examine the behavior of GLASS/Maximum Tree when gene trees are estimated instead of known with certainty, we applied GLASS/Maximum Tree to gene trees that were inferred from sequence alignments. We examined the influence of mutation on GLASS/Maximum Tree under two scenarios: a scenario with and a scenario without ancestral population structure. The species tree used in this analysis is identical to the species tree used in the previous section and in Figure 7.4. The only exception is that for the unstructured ancestral population analysis, we let the number of demes in each ancestral population equal one. To create very structured ancestral populations, we will use the scaled migration rate  $M = 0.1$  for the structured ancestral population model. We generated gene trees for  $L = 10, 20, 30, 40, 50, 60, 70, 80, 90, 100, 200, 300, 400, 500, 600, 700, 800, 900,$  and 1000 independent loci (with each set of  $L$  gene trees generated independent of each other set of gene trees) using MS (*Hudson, 2002*). To convert branch lengths from coalescent units to mutation units (average number of mutations along the branch), we multiplied each length by  $\theta/2$ , where  $\theta = 4N_e\mu = 0.01$ , and  $\mu$  is the mutation rate per site per generation. Each gene tree was input into SEQ-GEN (*Rambaut and Grassly, 1997*), which generated sequence alignments of length 500 nucleotides under a Jukes-Cantor substitution model. For each sequence alignment, we used PAUP\* (*Swofford, 2003*) to infer rooted gene trees with maximum likelihood assuming the Jukes-Cantor substitution model and a molecular clock. GLASS/Maximum Tree was then applied to the  $L$  inferred gene



trees. We repeated this process for a total of 1000 independent replicate simulations of  $L$  loci. Unlike our other simulation, this simulation incorporated mutation.

### 7.3.2.2 Results

Under the model with no ancestral population structure, when the number of loci is small, GLASS/Maximum Tree is increasingly likely to infer the correct species tree topology  $((AB)C)$  as the number of loci increases (Figure 7.5A). However, the frequency of the phylogeny with zero branch lengths  $(ABC)_0 = (A:0,B:0,C:0)$  also increases as the number of loci increases. This increase in frequency of  $(ABC)_0$  is caused by maximum likelihood estimating gene trees with branches of length zero due to no mutations. Once a single input tree has branches of length zero between a pair of species, the GLASS/Maximum Tree estimate must also have branches of length zero between the pair of species. As the number of loci increases, the probability increases that the inferred GLASS/Maximum Tree will contain branches of length zero. This increased probability is reflected in the simulations in which the frequency of  $(ABC)_0$  increases and the frequency of  $((AB)C)$  decreases as the number of loci gets increasing large.

Similar to the unstructured ancestral population case, when ancestral populations are structured, the inferred frequency of  $(ABC)_0$  increases as the number of loci increases (Figure 7.5B). Because the structured ancestral model is the same as that used in Figure 7.4 with  $M = 0.1$ , the probability is small that a gene tree will display the  $((AB)C)$  topology. This low probability for the relationship  $((AB)C)$  is reflected in the small fraction of species trees that have topology  $((AB)C)$  in Figure 7.5B. By incorporating the mutation process in addition to the genealogical process, we find GLASS/Maximum Tree is increasingly likely to infer an unresolved tree as the number of loci increases. This result indicates that GLASS/Maximum Tree may perform poorly in practice, as gene trees can only be estimated and are, therefore,

not known with certainty.

## 7.4 Discussion

In this article, we have described a general structured ancestral population model that extends the basic multispecies coalescent. Using this model, we have proven that in the presence of ancestral population structure, many commonly used consensus methods for inferring species trees from gene trees are no longer statistically consistent (Table 7.2). The only method we found to be consistent is GLASS/Maximum Tree, which relies on minimum coalescence times across gene trees between pairs of species. This result, however, is discomfoting because this method has the limitation that if little information exists in only a single locus in a sample collection of loci, it is possible to obtain an estimated divergence time of 0 between species (Figure 7.5). Although using the minimum coalescence times between pairs of species is statistically consistent when gene trees are known exactly, the utility of GLASS/Maximum Tree in practice is uncertain.

The observation that most consensus methods evaluated are misleading in the presence of ancestral population structure prompts the need to develop consensus methods that are robust to ancestral structure. Although our model is more general than the multispecies coalescent model, it still provides a simplification of true ancestral population structure. For example, our model assumes that the migration matrix, the number of demes, and the sizes of the demes are constant along an internal branch of the species tree. Real ancestral population structure will probably involve changes in population sizes (e.g., bottlenecks), changes in the number of demes (e.g., fission and fusion of demes), and changes in the migration rates between demes over time. However, because our model is a simplification of more complex structured population models, we expect that if consensus methods are inconsistent under our model, then the situation will only get worse under more complicated

models. Consequently, more studies need to be performed to empirically characterize the properties of ancestral population structure. These studies can be accomplished by looking at the frequencies of gene trees because certain types of discordance, such as asymmetries in the frequencies of gene trees, are signatures of ancestral population subdivision (*Slatkin and Pollack, 2008*). Results from such studies may be useful in developing consensus methods that are robust to gene tree discordance caused by specific types of subdivided ancestral populations.

## **7.5 Acknowledgments**

We thank Ethan Jewett and Cuong Than for their valuable comments. This work was supported by NSF grant DEB-0716904, NIH training grants T32 GM070449 and T32 HG00040, and a University of Michigan Rackham Merit Fellowship.

Table 7.1: Notation

Notation	Definition
$\mathcal{S}(\sigma, \mathbf{D}, \mathbf{N}, \mathbf{M}, \Psi)$	Ancestral structured population model with parameters $\sigma$ , $\mathbf{D}$ , $\mathbf{N}$ , $\mathbf{M}$ , and $\Psi$
$\mathbb{P}[E   \mathcal{S}]$	Probability of event $E$ under model $\mathcal{S}(\sigma, \mathbf{D}, \mathbf{N}, \mathbf{M}, \Psi)$
$\lambda_A$	Subtree of species tree $\sigma$ that contains species $A$ and that descends from the split of species $A$ and $B$
$\lambda_B$	Subtree of species tree $\sigma$ that contains species $B$ and that descends from the split of species $A$ and $B$
$\lambda_C$	Subtree of species tree $\sigma$ that contains species $C$ and that descends from the split of species $(AB)$ and $C$
$\Gamma_A, \Gamma_B, \Gamma_C$	Sets of taxa at the leaves of subtrees $\lambda_A$ , $\lambda_B$ , and $\lambda_C$ , respectively
$\mathcal{T} X$	Tree displayed by phylogenetic tree $\mathcal{T}$ restricted to the set of taxa $X$
$\text{top}(\mathcal{T})$	Topology of phylogenetic tree $\mathcal{T}$
$p_S(X, Y)$	Probability that a lineage sampled from species $X$ and a lineage sampled from species $Y$ are in the same deme at the speciation time of $X$ and $Y$ under model $\mathcal{S}(\sigma, \mathbf{D}, \mathbf{N}, \mathbf{M}, \Psi)$
$P_S[\mathcal{T}]$	Probability of gene tree topology $\mathcal{T}$ under model $\mathcal{S}(\sigma, \mathbf{D}, \mathbf{N}, \mathbf{M}, \Psi)$
$\hat{P}[\mathcal{T}]$	Sample proportion of topology $\mathcal{T}$ in a set of gene trees
$T_{XY}^\ell$	Random coalescence time at locus $\ell$ for a lineage sampled from species $X$ and a lineage sampled from species $Y$
$\mathbb{E}_S[T_{XY}^\ell]$	Expected coalescence time under model $\mathcal{S}(\sigma, \mathbf{D}, \mathbf{N}, \mathbf{M}, \Psi)$ for a lineage sampled from species $X$ and a lineage sampled from species $Y$ at locus $\ell$
$\bar{t}_{XY}$	Mean coalescence time across all sampled gene trees between one lineage sampled from species $X$ and one lineage sampled from species $Y$
$R_{XY}^\ell$	Rank of the coalescent event at locus $\ell$ for a lineage sampled from species $X$ and a lineage sampled from species $Y$
$\mathbb{E}_S[R_{XY}^\ell]$	Expected rank under $\mathcal{S}(\sigma, \mathbf{D}, \mathbf{N}, \mathbf{M}, \Psi)$ of the coalescent event for a lineage sampled from species $X$ and a lineage sampled from species $Y$ at locus $\ell$

Notation	Definition
$\bar{r}_{XY}$	Mean rank of coalescent events across all sampled gene trees between one lineage sampled from species $X$ and one lineage sampled from species $Y$
$t_{XY}^{\min}$	Minimum coalescence time across all sampled gene trees between one lineage sampled from species $X$ and one lineage sampled from species $Y$
$xl(\text{top}(\sigma), \mathcal{T})$	Number of extra lineages contributed by topology of fixed species tree $\sigma$ for a fixed gene tree topology $\mathcal{T}$
$xl(\text{top}(\sigma))$	Number of extra lineages contributed by topology of fixed species tree $\sigma$ for a fixed set of gene trees

Table 7.2: Summary of the behavior of consensus methods

Criterion	Asymptotic behavior	Asymptotic behavior (structure)	Theorem	Method	Reference
Uniquely favored topology	Misleading	Misleading	VII.1	Democratic vote	<i>Degnan and Rosenberg (2006)</i>
Average coalescence time	Consistent	Misleading	VII.2	STEAC	<i>Liu et al. (2009)</i>
Average ranks of coalescences	Consistent	Misleading	VII.3	STAR	<i>Liu et al. (2009)</i>
Uniquely favored triples	Consistent	Misleading	VII.4	R* Consensus	<i>Degnan et al. (2009)</i>
Minimizing deep coalescences	Misleading	Misleading	VII.5	Rooted Triple Consensus MDC	<i>Ewing et al. (2008)</i> <i>Maddison (1997); Than and Nakhleh (2009)</i>
Majority-rule	Inconsistent	Misleading	VII.6	Majority-Rule Consensus	<i>Degnan et al. (2009)</i>
Minimum coalescence time	Consistent	Consistent	VII.7	GLASS	<i>Mossel and Roch (2010)</i>
				Maximum Tree	<i>Liu et al. (2010)</i>

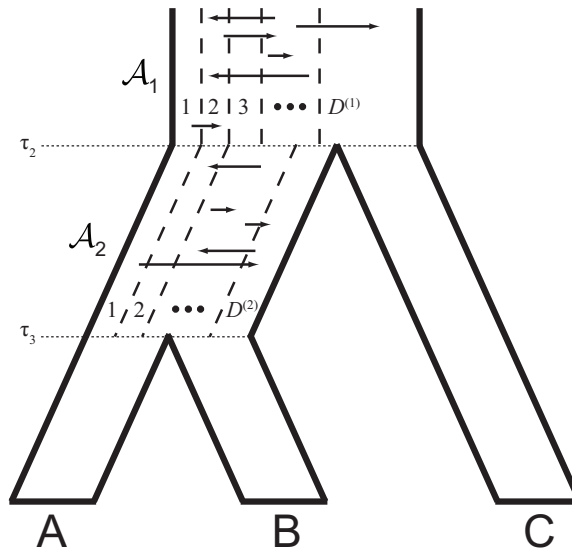


Figure 7.1: Model for the relationship among species A, B and C in a fixed species tree  $\sigma$ . Ancestral population  $\mathcal{A}_1$  has  $D^{(1)}$  demes and ancestral population  $\mathcal{A}_2$  has  $D^{(2)}$  demes. Migration occurs between the  $D^{(1)}$  demes in ancestral population  $\mathcal{A}_1$  and between the  $D^{(2)}$  demes in ancestral population  $\mathcal{A}_2$ . At  $\tau_2$  and  $\tau_3$ , going back in time, lineages merge into specific demes in ancestral populations  $\mathcal{A}_1$  and  $\mathcal{A}_2$  according to the matrix  $\Psi$  (see “Model” section).

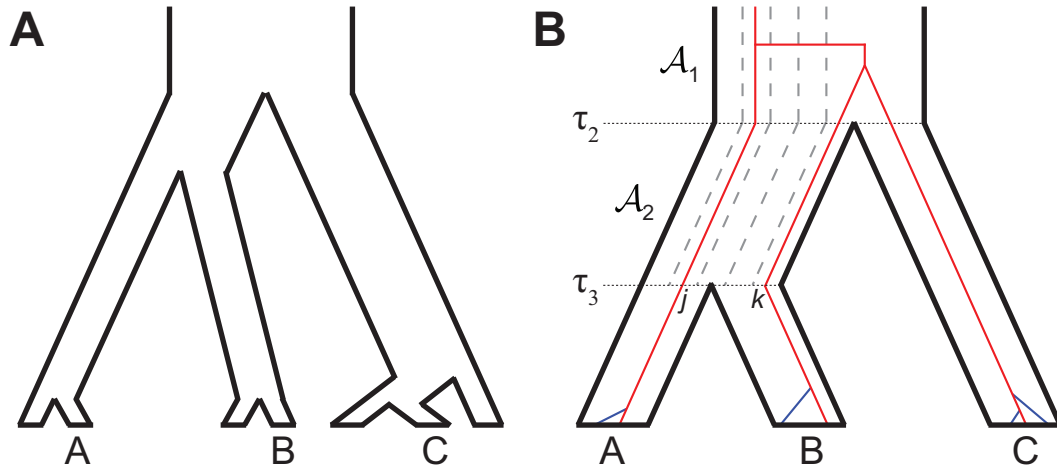


Figure 7.2: Counterexample used to prove that consensus methods are misleading. (A) Certain internal branches are made long so that the  $n$ -taxon species tree  $\sigma$  resembles a three-taxon species tree. (B) Lineages from species A, B, and C are in red and lineages from other taxa that have coalesced along the branches leading to species A, B, and C are in blue. The lineage from species A merges into deme  $j^{(2)}$  of ancestral population  $\mathcal{A}_2$ , the lineage from species B merges into deme  $k^{(2)} \neq j^{(2)}$  of ancestral population  $\mathcal{A}_2$ , and the lineage from species C merges into deme  $k^{(1)}$  of ancestral population  $\mathcal{A}_1$ . The migration rates out of demes are small, so that the lineages from A, B, and C each have low probabilities of leaving the deme in which they started. As a consequence, the probability is high that the lineage from B coalesces with the lineage from C before it coalesces with the lineage from A.



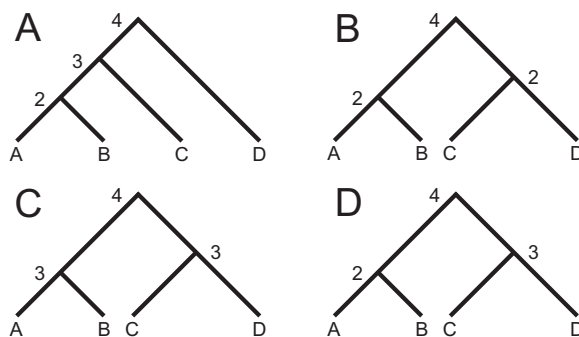


Figure 7.3: Four possible coalescence ranks of four-taxon trees. (A) Coalescence ranks for an asymmetric four-taxon tree. (B) Coalescence ranks for a symmetric four-taxon tree in which the rank of an internal node is the number of leaves descending from it. (C) Coalescence ranks for a symmetric four-taxon tree in which the rank of an internal node is the rank of the node directly ancestral to it minus 1. (D) Coalescence ranks for a symmetric four-taxon tree in which the rank of an internal node is assigned relative to all other internal nodes in the tree. In this ranking, each possible value for a rank (*i.e.*, ranks  $2, 3, \dots, n$ ) is used once.

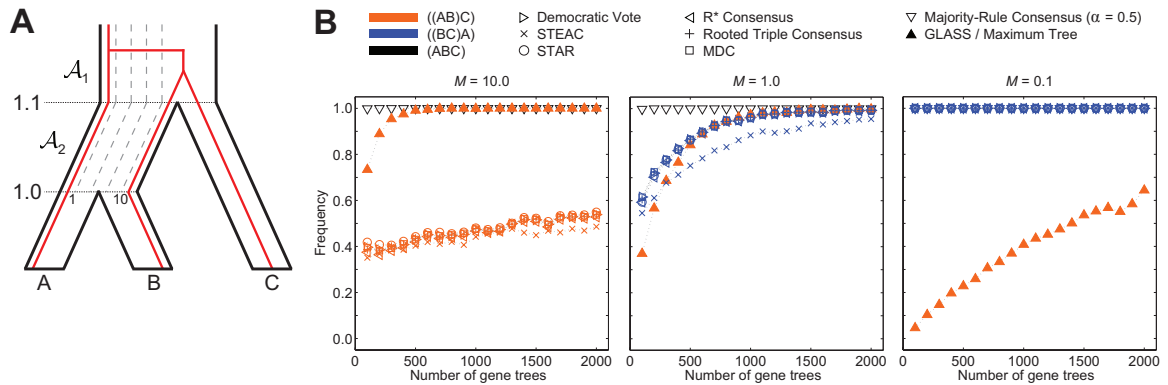


Figure 7.4: Simulation results for the three-taxon species tree ((AB)C). (A) Species tree with a structured ancestral population model. Time is measured in coalescent units  $t/(2N_e)$ , where  $t$  is time in generations and  $N_e$  is a reference diploid effective population size. The structured population model is an island migration model with  $D = 10$  demes in each ancestral population. The scaled migration rate between deme  $x$  and deme  $y \neq x$  is  $M = 4N_e m$ , which corresponds to  $M/4$  individuals per generation in each direction. Species A merges into deme 1 and species B and C each merge into deme 10. (B) Simulation results for scaled migration rates  $M = 10.0$ ,  $M = 1.0$ , and  $M = 0.1$ . Each tree topology is represented by a distinct color. Each consensus method is represented by a distinct symbol. For each consensus method, the tree topology traced is the topology for that method that has the highest frequency at 2000 gene trees. The frequency of a topology is calculated as the fraction among 1000 replicate simulations for which that topology was inferred.

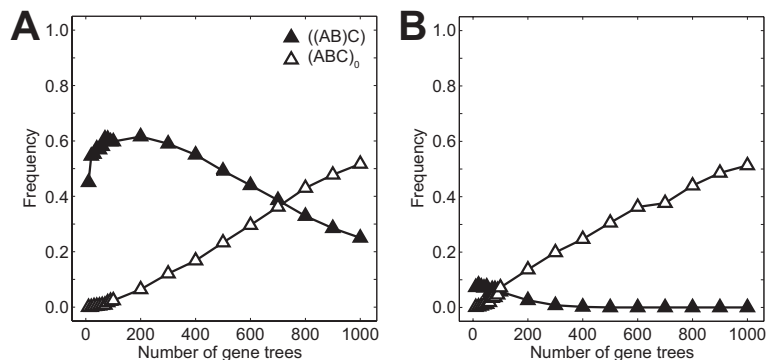


Figure 7.5: Inference of species trees using GLASS/Maximum Tree under a Jukes-Cantor substitution model (per-site mutation rate  $\theta = 0.01$ ) when gene trees are generated under the three-taxon species tree  $\sigma = ((A:1.0, B:1.0):0.1, C:1.1)$ . Time is measured in coalescent units  $t/(2N_e)$ , where  $t$  is time in generations and  $N_e$  is a reference diploid effective population size. (A) Simulation results with no ancestral population structure. (B) Simulation results with ancestral population structure. The structured population model is an island migration model with  $D = 10$  demes in each ancestral population (the same model as in Figure 7.4A). The scaled migration rate between deme  $x$  and deme  $y \neq x$  is  $M = 4N_e m = 0.1$ , which corresponds to one individual per 40 generations in each direction. Species A merges into deme 1 and species B and C each merge into deme 10. Each tree topology is represented by a symbol: an open triangle for a tree with zero branch lengths  $(ABC)_0$  and a closed triangle for  $((AB)C)$ . The frequency of a topology is calculated as the fraction among 1000 replicate simulations for which that topology was inferred.

## 7.6 Appendix

In this section, we provide proofs that STEAC, STAR, R\* Consensus, Rooted Triple Consensus, Minimize Deep Coalescences, and Majority-Rule Consensus are misleading estimators of a species tree topology under model  $\mathcal{S}(\sigma, \mathbf{D}, \mathbf{N}, \mathbf{M}, \Psi)$ .

### 7.6.1 Average coalescence times

Under model  $\mathcal{S}(\sigma, \mathbf{D}, \mathbf{N}, \mathbf{M}, \Psi)$ , define the expected time to coalescence at a random locus for a random lineage sampled from species X and a random lineage sampled from species Y as  $\mathbb{E}_{\mathcal{S}}[T_{XY}]$ .

*Proof of Theorem VII.2.* We use the counterexample (assumptions 1–3 in the “Counterexample” section) to show that  $\widehat{C}_L$  is misleading. For  $\widehat{C}_L$  to not be misleading, we must have that  $\widehat{C}_L \xrightarrow{P} \text{top}(\sigma)$  as  $L \rightarrow \infty$ . Note that  $\text{top}(\sigma) = ((\lambda_A \lambda_B) \lambda_C)$ . Because the migration model is the same across loci, we will drop the superscript in  $T_{AB}^\ell$  and  $T_{BC}^\ell$  for convenience.

Denote the expected coalescence time for two random lineages residing in the same deme by  $\mathbb{E}[T \mid \text{same}]$  and the expected coalescence time for two random lineages residing in different demes by  $\mathbb{E}[T \mid \text{diff}]$ . We can write the expected coalescence time for a random lineage sampled from species A and a random lineage sampled from species B as

$$\mathbb{E}_{\mathcal{S}}[T_{AB}] = \tau_3 + \mathbb{E}[T \mid \text{same}]p_{\mathcal{S}}(A, B) + \mathbb{E}[T \mid \text{diff}][1 - p_{\mathcal{S}}(A, B)],$$

where  $\tau_3$  is the divergence time of species A and B. Under the assumptions of the counterexample, because  $p_{\mathcal{S}}(A, B) = 0$ ,

$$\mathbb{E}_{\mathcal{S}}[T_{AB}] = \tau_3 + \mathbb{E}[T \mid \text{diff}].$$

The expected coalescence time for a random lineage sampled from species B and a random lineage sampled from species C is

$$\begin{aligned}\mathbb{E}_{\mathcal{S}}[T_{\text{BC}}] &= \tau_2 + \mathbb{E}[T \mid \text{same}]p_{\mathcal{S}}(\text{B}, \text{C}) + \mathbb{E}[T \mid \text{diff}][1 - p_{\mathcal{S}}(\text{B}, \text{C})] \\ &= \tau_2 + \mathbb{E}[T \mid \text{diff}] - p_{\mathcal{S}}(\text{B}, \text{C})(\mathbb{E}[T \mid \text{diff}] - \mathbb{E}[T \mid \text{same}]),\end{aligned}$$

where  $\tau_2$  is the divergence time of species B and C. In an island migration model with  $D$  demes,  $\mathbb{E}[T \mid \text{same}] = 2ND$  and  $\mathbb{E}[T \mid \text{diff}] = 2ND + (D - 1)/(2m)$  (*Wakeley*, 2009, p. 152, eqs. 5.26 and 5.27). It follows that

$$\begin{aligned}\mathbb{E}_{\mathcal{S}}[T_{\text{BC}}] - \mathbb{E}_{\mathcal{S}}[T_{\text{AB}}] &= (\tau_2 - \tau_3) - p_{\mathcal{S}}(\text{B}, \text{C})(\mathbb{E}[T \mid \text{diff}] - \mathbb{E}[T \mid \text{same}]) \\ &= (\tau_2 - \tau_3) - \frac{D - 1}{2m}p_{\mathcal{S}}(\text{B}, \text{C}).\end{aligned}$$

Recall from the counterexample that  $p_{\mathcal{S}}(\text{B}, \text{C}) > \beta_1$ , and that when holding  $\tau_2 - \tau_3$  and  $D$  constant and setting the migration rate  $m$  sufficiently small,  $\beta_1$  is arbitrarily close to 1. Set  $m$  sufficiently small such that

$$p_{\mathcal{S}}(\text{B}, \text{C}) > \frac{2(\tau_2 - \tau_3)}{D - 1}m. \quad (7.2)$$

Consequently,  $\mathbb{E}_{\mathcal{S}}[T_{\text{BC}}] < \mathbb{E}_{\mathcal{S}}[T_{\text{AB}}]$ . For  $\widehat{C}_L$  to be misleading, we need to obtain  $\bar{t}_{\text{BC}} < \bar{t}_{\text{AB}}$  as  $L \rightarrow \infty$ . By the Law of Large Numbers, Slutsky's Theorem (*Serfling*, 1980, Theorem 1.5.4), and Corollary B of *Serfling* (1980),  $\bar{t}_{\text{BC}} - \bar{t}_{\text{AB}} \xrightarrow{P} \mathbb{E}_{\mathcal{S}}[T_{\text{BC}}] - \mathbb{E}_{\mathcal{S}}[T_{\text{AB}}] < 0$  as  $L \rightarrow \infty$  and, hence,  $\bar{t}_{\text{BC}} < \bar{t}_{\text{AB}}$ . Because  $\bar{t}_{\text{BC}} < \bar{t}_{\text{AB}}$  as  $L \rightarrow \infty$ ,  $\mathbb{P}[\widehat{C}_L = \text{top}(\sigma) \mid \mathcal{S}] \rightarrow 0$  as  $L \rightarrow \infty$ . Therefore,  $\widehat{C}_L \not\xrightarrow{P} \text{top}(\sigma)$ , and  $\widehat{C}_L$  is a misleading estimator of  $\text{top}(\sigma)$  under  $\mathcal{S}(\sigma, \mathbf{D}, \mathbf{N}, \mathbf{M}, \Psi)$ .  $\square$

### 7.6.2 Average ranks of coalescences

Let  $\mathbb{E}_{\mathcal{S}}[R_{XY}^{\ell}]$  denote the expected rank of a coalescent event for a random lineage from species X and a random lineage from species Y in a gene tree from locus  $\ell$  under  $\mathcal{S}(\sigma, \mathbf{D}, \mathbf{N}, \mathbf{M}, \Psi)$ .

*Proof of Theorem VII.3.* We use the counterexample (assumptions 1–3 in the “Counterexample” section) to show that  $\widehat{C}_L$  is misleading. For  $\widehat{C}_L$  to not be misleading, we must have that  $\widehat{C}_L \xrightarrow{P} \text{top}(\sigma)$  as  $L \rightarrow \infty$ . Note that  $\text{top}(\sigma) = ((\lambda_A \lambda_B) \lambda_C)$ . Because the migration model is the same across loci, we will drop the superscript in  $R_{AB}^{\ell}$  and  $R_{BC}^{\ell}$  for convenience.

Define  $q_{XY} = \mathbb{P}[\tau_3 \leq T_{XY} < \tau_2 \mid \mathcal{S}]$ . The expected rank for a lineage sampled from species A and a lineage sampled from species B in species tree  $\sigma$  is

$$\begin{aligned} \mathbb{E}_{\mathcal{S}}[R_{AB}] &= \mathbb{E}_{\mathcal{S}}[R_{AB} \mid \tau_3 \leq T_{AB} < \tau_2]q_{AB} + \mathbb{E}_{\mathcal{S}}[R_{AB} \mid \tau_2 \leq T_{AB}](1 - q_{AB}) \\ &= \mathbb{E}_{\mathcal{S}}[R_{AB} \mid \tau_2 \leq T_{AB}] - q_{AB}(\mathbb{E}_{\mathcal{S}}[R_{AB} \mid \tau_2 \leq T_{AB}] - \mathbb{E}_{\mathcal{S}}[R_{AB} \mid \tau_3 \leq T_{AB} < \tau_2]). \end{aligned}$$

Note that because species B and C cannot coalesce more recently than time  $\tau_2$ ,  $q_{BC} = 0$ . The expected rank for a lineage sampled from species B and C is then

$$\begin{aligned} \mathbb{E}_{\mathcal{S}}[R_{BC}] &= \mathbb{E}_{\mathcal{S}}[R_{BC} \mid \tau_3 \leq T_{BC} < \tau_2]q_{BC} + \mathbb{E}_{\mathcal{S}}[R_{BC} \mid \tau_2 \leq T_{BC}](1 - q_{BC}) \\ &= \mathbb{E}_{\mathcal{S}}[R_{BC} \mid \tau_2 \leq T_{BC}]. \end{aligned}$$

It follows that

$$\begin{aligned} \mathbb{E}_{\mathcal{S}}[R_{BC}] - \mathbb{E}_{\mathcal{S}}[R_{AB}] &= (\mathbb{E}_{\mathcal{S}}[R_{BC} \mid \tau_2 \leq T_{BC}] - \mathbb{E}_{\mathcal{S}}[R_{AB} \mid \tau_2 \leq T_{AB}]) \\ &\quad + q_{AB}(\mathbb{E}_{\mathcal{S}}[R_{AB} \mid \tau_2 \leq T_{AB}] - \mathbb{E}_{\mathcal{S}}[R_{AB} \mid \tau_3 \leq T_{AB} < \tau_2]). \end{aligned}$$

Recall from the counterexample that when holding  $\tau_2 - \tau_3$  and  $D$  constant and setting

the migration rate  $m$  sufficiently small, the probability of zero migration events over the interval  $[\tau_3, \tau_2)$  (*i.e.*,  $\beta_1$ ) is arbitrarily close to 1. If no migration event occurs on the interval  $[\tau_2, \tau_3)$ , then there cannot be a coalescence between a lineage from A and a lineage from B. Consequently, holding  $\tau_2 - \tau_3$  and  $D$  constant and setting  $m$  sufficiently small,  $q_{AB}$  is arbitrarily close to 0. Set  $m$  sufficiently small such that

$$q_{AB} < \frac{\mathbb{E}_{\mathcal{S}}[R_{AB} \mid \tau_2 \leq T_{AB}] - \mathbb{E}_{\mathcal{S}}[R_{BC} \mid \tau_2 \leq T_{BC}]}{\mathbb{E}_{\mathcal{S}}[R_{AB} \mid \tau_2 \leq T_{AB}] - \mathbb{E}_{\mathcal{S}}[R_{AB} \mid \tau_3 \leq T_{AB} < \tau_2]}. \quad (7.3)$$

Consequently,  $\mathbb{E}_{\mathcal{S}}[R_{BC}] < \mathbb{E}_{\mathcal{S}}[R_{AB}]$ . For  $\widehat{C}_L$  to be misleading, we need to obtain  $\bar{r}_{BC} < \bar{r}_{AB}$  as  $L \rightarrow \infty$ . By the Law of Large Numbers, Slutsky's Theorem (*Serfling*, 1980, Theorem 1.5.4), and Corollary *B* of *Serfling* (1980),  $\bar{r}_{BC} - \bar{r}_{AB} \xrightarrow{P} \mathbb{E}_{\mathcal{S}}[R_{BC}] - \mathbb{E}_{\mathcal{S}}[R_{AB}] < 0$  as  $L \rightarrow \infty$  and, hence,  $\bar{r}_{BC} < \bar{r}_{AB}$ . Because  $\bar{r}_{BC} < \bar{r}_{AB}$  as  $L \rightarrow \infty$ ,  $\mathbb{P}[\widehat{C}_L = \text{top}(\sigma) \mid \mathcal{S}] \rightarrow 0$  as  $L \rightarrow \infty$ . Therefore,  $\widehat{C}_L \not\xrightarrow{P} \text{top}(\sigma)$ , and  $\widehat{C}_L$  is a misleading estimator of  $\text{top}(\sigma)$  under  $\mathcal{S}(\sigma, \mathbf{D}, \mathbf{N}, \mathbf{M}, \Psi)$ .  $\square$

### 7.6.3 Uniquely favored rooted triples

*Proof of Theorem VII.4.* We use the counterexample (assumptions 1–3 in the “Counterexample” section) to show that  $\widehat{C}_L$  is misleading. For  $\widehat{C}_L$  to not be misleading, we must have that  $\widehat{C}_L \xrightarrow{P} \text{top}(\sigma)$  as  $L \rightarrow \infty$ . Note that  $\text{top}(\sigma) = ((\lambda_A \lambda_B) \lambda_C)$ . Consider an alternative species tree  $\sigma^*$  with topology  $\text{top}(\sigma^*) = ((\lambda_B \lambda_C) \lambda_A)$ . Set the migration rate  $m$  sufficiently small such that  $P_{\mathcal{S}}[\text{top}(\sigma^*)] = P_{\mathcal{S}}[((\lambda_B \lambda_C) \lambda_A)] > (1 - \delta)^3 \beta_2$ , which is arbitrarily close to 1. By the Law of Large Numbers,  $\widehat{P}[\text{top}(\sigma^*)] \xrightarrow{P} P_{\mathcal{S}}[\text{top}(\sigma^*)]$  as  $L \rightarrow \infty$ . Because  $P_{\mathcal{S}}[\text{top}(\sigma^*)]$  has probability arbitrarily close to 1, the set of rooted triples displayed by  $\text{top}(\sigma^*)$  is the set of uniquely favored rooted triples. Because a rooted bifurcating tree topology is defined by its set of rooted triples (*Steel*, 1992, Proposition 4),  $\mathbb{P}[\widehat{C}_L = \text{top}(\sigma^*) \mid \mathcal{S}] \rightarrow 1$  as  $L \rightarrow \infty$ . Therefore,  $\widehat{C}_L \not\xrightarrow{P} \text{top}(\sigma)$ , and  $\widehat{C}_L$  is a misleading estimator of  $\text{top}(\sigma)$ .

under  $\mathcal{S}(\sigma, \mathbf{D}, \mathbf{N}, \mathbf{M}, \Psi)$ . □

#### 7.6.4 Minimizing deep coalescences

Consider fixed species tree  $\sigma$  and fixed gene tree topology  $\mathcal{T}$  and let  $\text{xl}(\text{top}(\sigma), \mathcal{T})$  denote the number of extra lineages contributed by the topology of  $\sigma$  for gene tree topology  $\mathcal{T}$ . Consider the set of all possible  $n$ -taxon rooted bifurcating tree topologies  $\mathcal{G}$ . The number of extra lineages contributed by the topology of  $\sigma$  is

$$\text{xl}(\text{top}(\sigma)) = \sum_{\mathcal{T} \in \mathcal{G}} \text{xl}(\text{top}(\sigma), \mathcal{T}) \widehat{P}[\mathcal{T}]. \quad (7.4)$$

*Proof of Theorem VII.5.* We use the counterexample (assumptions 1–3 in the “Counterexample” section) to show that  $\widehat{C}_L$  is misleading. For  $\widehat{C}_L$  to not be misleading, we must have that  $\widehat{C}_L \xrightarrow{P} \text{top}(\sigma)$  as  $L \rightarrow \infty$ . Note that  $\text{top}(\sigma) = ((\lambda_A \lambda_B) \lambda_C)$ . Consider an alternative species tree  $\sigma^*$  with topology  $\text{top}(\sigma^*) = ((\lambda_B \lambda_C) \lambda_A)$ . From the counterexample, we know that certain branch lengths of the species tree are sufficiently long such that for fixed set  $X \in \{\Gamma_A, \Gamma_B, \Gamma_C\}$  and fixed arbitrarily small  $\delta > 0$ ,  $\mathbb{P}[\text{top}(\mathcal{T}|X) = \text{top}(\sigma|X) | \mathcal{S}] > 1 - \delta$ . Using eq. 7.4, the difference in the number of extra lineages contributed by the topologies of  $\sigma$  and  $\sigma^*$  is

$$\begin{aligned} \Delta_{\text{xl}}(\text{top}(\sigma), \text{top}(\sigma^*)) &= \text{xl}(\text{top}(\sigma)) - \text{xl}(\text{top}(\sigma^*)) \\ &= \sum_{\mathcal{T} \in \mathcal{G}} [\text{xl}(\text{top}(\sigma), \mathcal{T}) - \text{xl}(\text{top}(\sigma^*), \mathcal{T})] \widehat{P}[\mathcal{T}]. \end{aligned}$$

Note that  $\text{xl}(\text{top}(\sigma), \text{top}(\sigma^*)) = 1$  and  $\text{xl}(\text{top}(\sigma^*), \text{top}(\sigma^*)) = 0$ . Set the migration rate  $m$  sufficiently small such that  $P_S[\text{top}(\sigma^*)] = P_S[((\lambda_B \lambda_C) \lambda_A)] > (1 - \delta)^3 \beta_2$ , which is arbitrarily close to 1. It follows that, for each  $\mathcal{T} \in \mathcal{G} \setminus \{\text{top}(\sigma^*)\}$ ,  $P_S[\mathcal{T}]$  is arbitrarily close to 0. For  $\widehat{C}_L$  to be misleading, we need to obtain  $\Delta_{\text{xl}}(\text{top}(\sigma), \text{top}(\sigma^*)) > 0$  as  $L \rightarrow \infty$ . For fixed arbitrarily small  $\varepsilon$  close to 0, and by the Law of Large Numbers, Slutsky’s



Theorem (*Serfling*, 1980, Theorem 1.5.4), and Corollary *B* of *Serfling* (1980), as  $L \rightarrow \infty$ ,

$$\begin{aligned}
\Delta_{\text{xl}}(\text{top}(\sigma), \text{top}(\sigma^*)) &\xrightarrow{P} \sum_{\mathcal{T} \in \mathcal{G}} [\text{xl}(\text{top}(\sigma), \mathcal{T}) - \text{xl}(\text{top}(\sigma^*), \mathcal{T})] P_{\mathcal{S}}[\mathcal{T}] \\
&= [\text{xl}(\text{top}(\sigma), \text{top}(\sigma^*)) - \text{xl}(\text{top}(\sigma^*), \text{top}(\sigma^*))] P_{\mathcal{S}}[\mathcal{T}] + \varepsilon \\
&= P_{\mathcal{S}}[\mathcal{T}] + \varepsilon \\
&> 0.
\end{aligned}$$

Because  $\Delta_{\text{xl}}(\text{top}(\sigma), \text{top}(\sigma^*)) > 0$  as  $L \rightarrow \infty$ ,  $\mathbb{P}[\widehat{C}_L = \text{top}(\sigma) | \mathcal{S}] \rightarrow 0$  as  $L \rightarrow \infty$ . Therefore,  $\widehat{C}_L \not\xrightarrow{P} \text{top}(\sigma)$ , and  $\widehat{C}_L$  is a misleading estimator of  $\text{top}(\sigma)$  under  $\mathcal{S}(\sigma, \mathbf{D}, \mathbf{N}, \mathbf{M}, \Psi)$ .  $\square$

### 7.6.5 Majority-rule

*Proof of Theorem VII.6.* We use the counterexample (assumptions 1–3 in the “Counterexample” section) to show that  $\widehat{C}_L$  is misleading. For  $\widehat{C}_L$  to not be misleading, we must have that  $\widehat{C}_L \xrightarrow{P} \text{top}(\sigma)$  as  $L \rightarrow \infty$ . Note that  $\text{top}(\sigma) = ((\lambda_A \lambda_B) \lambda_C)$ . Consider an alternative species tree  $\sigma^*$  with topology  $\text{top}(\sigma^*) = ((\lambda_B \lambda_C) \lambda_A)$ . Fix  $\alpha \in [0.5, 1)$ . Set  $\delta > 0$  arbitrarily small and recall from the counterexample that  $P_{\mathcal{S}}[\text{top}(\sigma^*)] = P_{\mathcal{S}}[((\lambda_B \lambda_C) \lambda_A)] > (1 - \delta)^3 \beta_2$ , and that when holding  $\tau_2 - \tau_3$  and  $D$  constant and setting the migration rate  $m$  sufficiently small,  $\beta_2$  is arbitrarily close to 1. For  $\widehat{C}_L$  to be misleading, all clades displayed by  $\text{top}(\sigma^*)$  must have frequency greater than  $\alpha$ . By the Law of Large Numbers,  $\widehat{P}[\text{top}(\sigma^*)] \xrightarrow{P} P_{\mathcal{S}}[\text{top}(\sigma^*)]$  as  $L \rightarrow \infty$ . Because  $P_{\mathcal{S}}[\text{top}(\sigma^*)]$  has probability arbitrarily close to 1, all clades displayed by  $\text{top}(\sigma^*)$  have a frequency greater than  $\alpha$  and so  $\mathbb{P}[\widehat{C}_L = \text{top}(\sigma^*) | \mathcal{S}] \rightarrow 1$  as  $L \rightarrow \infty$ . Therefore,  $\widehat{C}_L \xrightarrow{P} \text{top}(\sigma^*)$ , and  $\widehat{C}_L$  is a misleading estimator of  $\text{top}(\sigma)$  under  $\mathcal{S}(\sigma, \mathbf{D}, \mathbf{N}, \mathbf{M}, \Psi)$ .  $\square$

## CHAPTER VIII

# An empirical evaluation of species tree inference strategies using a multilocus dataset from North American pines

### 8.1 Introduction

In phylogenetic studies, it has become increasingly common to sequence large numbers of individuals across many loci. These multilocus datasets provide the potential to improve the accuracy of phylogeny inferences over large sets of taxa. However, for a variety of reasons, topologies of trees inferred at different loci might not match (*Rannala and Yang, 2008*). One source of this gene tree discordance is incomplete lineage sorting—the phenomenon in which sets of sampled lineages fail to coalesce in the population in which they are first capable of coalescing (*Degnan and Rosenberg, 2009*). If incomplete lineage sorting occurs, then gene trees might not match the species tree, and further, species tree inference methods can be misled by the discordance of gene trees at multiple loci (*Degnan and Rosenberg, 2006; Kubatko and Degnan, 2007; Degnan et al., 2009*).

Recently, three main families of approaches have been used for estimating species tree topologies from multilocus sequence data: consensus, concatenation, and Bayesian methods. Consensus methods construct species tree topologies from

gene trees according to a deterministic set of rules that are based on features of the input set of trees (*Bryant, 2003*). These methods have the property that their input is a set of gene trees that are constructed from individual loci, allowing for a separate evolutionary history to be inferred at each locus. However, because genetic lineages in different species sometimes have relatively few sequence differences, there might not be enough information in a locus to accurately infer gene trees. Concatenation methods concatenate a set of multiple alignments and, also according to a deterministic procedure, infer a species tree as the estimated tree based on the concatenated alignment. These methods have the property that they construct species trees from a large number of loci considered simultaneously, utilizing as much sequence information as is available to infer a species tree in a single step. However, concatenation methods combine all loci to form a single locus, and because different loci may have different evolutionary histories that are disregarded in the concatenation step, joining loci in this way can lead to incorrect species tree inferences. Bayesian methods infer species trees by modeling the evolution of sequences among all sampled loci. These methods have the property that in addition to inferring the evolutionary history among all sequences, they also provide a level of confidence for their species tree estimates. However, because these methods explore large spaces of evolutionary histories rather than algorithmically construct estimated trees, they are extremely computationally intensive and are typically applicable only to smaller datasets.

Simulation studies are commonly used to investigate the performance of species tree inference methods on multilocus datasets with a finite number of loci. These studies have the advantage of knowing the true species tree but the disadvantage that they can explore only a small portion of the evolutionary parameter space. An alternative approach is to evaluate the performance of methods on an empirical dataset in which the space of parameter values is defined by the actual evolutionary history of a group of species. Several recent studies have empirically investigated

the performance of phylogenetic tree construction methods from multilocus datasets. These studies have examined a variety of organisms, including birds (*Jennings and Edwards, 2005; Brumfield et al., 2008; Carling and Brumfield, 2008; Liu et al., 2008*), insects (*Carstens and Knowles, 2007; Linnen and Farrell, 2008*), newts (*Themudo et al., 2009*), plants (*Buerki et al., 2011*), primates (*Takezaki and Nei, 2008; Hird et al., 2010*), rice (*Cranston et al., 2009*), rodents (*Belfiore et al., 2008*), snakes (*Kubatko et al., 2009*), and yeast (*Rokas et al., 2003; Gatesy and Baker, 2005; Edwards et al., 2007; Liu et al., 2008*). While some of these studies have been able to construct highly-supported species trees, other studies could not do so, likely due to high levels of genealogical discordance.

In one such study, *Syring et al. (2007)* found that samples from a multilocus dataset of North American pines displayed widespread genealogical discordance. This pattern of incomplete lineage sorting is a common feature of long-lived shrubs and trees (*e.g., Bouillé and Bousquet, 2005; Ma et al., 2006; Willyard et al., 2009*), and likely has its genesis in a combination of factors such as large effective population size, long generation time, and high levels of gene flow (*Savolainen and Pyhäjärvi, 2007; Eckert and Carstens, 2008*). Because discordance is needed for different algorithms to produce different estimates, samples from a variety of North American pine species can provide a sensible dataset in which to study the performance of species tree inference methods in the presence of gene tree discordance. In this study, we take an empirical approach to the evaluation of species tree inference methods by examining the performance of 72 strategies for inferring species tree topologies using a multilocus dataset from North American pines. Each “phylogenetic inference strategy” consists of three components: a method of constructing species trees from gene trees (*e.g., consensus or concatenation*), a gene tree inference method (*e.g., maximum likelihood or neighbor-joining*), and an outgroup species. Our dataset consists of  $\sim 48$  kilobases (kb) of sequence spanning 123 nuclear loci that have been

sequenced in 120 individuals sampled from eight ingroup species of *Pinus* subsection *Strobus* and three outgroup species of *Pinus* subsection *Gerardianae*. We apply techniques from multivariate statistical analysis to sets of inferred species trees to compare and contrast characteristics of species trees estimated by different strategies and to identify groups of strategies that behave similarly. From these results, we provide recommendations for inferring species tree topologies from multilocus sequence data.

## 8.2 Methods

### 8.2.1 North American white pine dataset

A total of 120 individuals were sequenced in eight ingroup species of North American white pines from *Pinus* subsection *Strobus* (*Pinus albicaulis*, *P. ayacahuite*, *P. chiapensis*, *P. flexilis*, *P. lambertiana*, *P. monticola*, *P. strobiformis*, and *P. strobus*) and three outgroup species from *Pinus* subsection *Gerardianae* (*P. bungeana*, *P. gerardiana*, *P. squamata*), the identified sister lineage to *Pinus* subsection *Strobus* (*Syring et al.*, 2005, 2007). Sequence data were pre-processed and organized using PINESAP (*Wegrzyn et al.*, 2009), a bioinformatics pipeline that combines PHRED (*Ewing et al.*, 1998), PHRAP (*Lee and Vega*, 2004), and MUSCLE (*Edgar*, 2004a,b) to call bases and align sequencing reads. Following this pre-processing step, the data were manually assembled and aligned using CODONCODE (CodonCode Corporation, Dedham, MA). Bases were called using a minimum PHRED score (*Ewing and Green*, 1998; *Ewing et al.*, 1998) of 25 for aligned bases. All polymorphisms were visually validated. All alignments were further aligned to resequencing data from *P. taeda* (unpublished data) using the profile-profile option in MUSCLE (*Edgar*, 2004a,b). These alignments are publicly available as part of the Dendrome project (<http://loblolly.ucdavis.edu/bipod/ftp/>).

Of 245 loci sequenced initially, 37 were dropped from further consideration due to low overall quality of sequence reads. An additional 15 loci were dropped due to possible chloroplast or mitochondrial contamination, on the basis of BLAST analysis against pine organellar sequences deposited in GENBANK (*Parks et al.*, 2009). Two loci were dropped due to sequence similarity to retroelement-like proteins. This process resulted in a core set of 191 high-quality nuclear gene alignments for the 11 target species. We then eliminated 68 loci for which at least one of the 11 species contained no data. This filter reduced the dataset to 123 loci, covering  $\sim 48$  kb of aligned sequence data.

Coding regions (*i.e.*, site annotations) could confidently be identified for 112 of the 123 loci by further analysis using tBLASTx against protein-coding genes in *Arabidopsis*, *Oryza*, *Picea*, and *Populus*. For these 112 loci, the gene for the highest-scoring tBLASTx hit, in combination with the expressed sequence tag from loblolly pine, was used to identify coding regions. Site annotations for each alignment were validated with BLASTp analysis of the amino acid sequences derived from the inferred coding intervals against the gene that was used to derive the site annotations. Of the 48 kb of data available,  $\sim 62\%$  represents exonic regions,  $\sim 18\%$  represents intronic regions,  $\sim 1\%$  is from 5' UTRs, and  $\sim 19\%$  is from 3' UTRs.

### 8.2.2 Overview of the analysis

The procedure for obtaining results for each of the 72 phylogenetic inference strategies (listed in Table 8.1) is illustrated by a flow diagram in Figure 8.1. For a given strategy, we started from a dataset  $D$  with  $L$  loci. We then created a bootstrap dataset by randomly choosing with replacement  $B$  sets of  $L$  loci. Next, we applied a gene tree inference method to each bootstrap replicate dataset. Based on the set of inferred gene trees in a bootstrap replicate, we then applied a species tree construction method to estimate a species tree topology with one of the three outgroup

species from *Pinus* subsection *Gerardiana*. For each phylogenetic inference strategy, we constructed  $B = 1000$  independent bootstrap datasets, thereby estimating 1000 species tree topologies. From these species tree topologies, we created a list of clades, each with a corresponding count of its number of appearances in the 1000 bootstrap replicates. These clade lists were then analyzed to assess similarity and difference among the estimates produced by different strategies.

### 8.2.3 Creating datasets

The final set of 123 loci in our dataset contains many loci that are highly conserved across multiple species. Because of the high level of conservation, for these loci, little information exists for identifying relationships among lineages. Thus, if methods for inferring gene trees were applied to certain loci, the resulting gene trees would be highly unresolved and would therefore provide little information to species tree construction methods.

To circumvent this problem, we instead analyzed four carefully selected subsets of the initial dataset (Table 8.2). Two of these are datasets of multiple alignments that contain information on a single individual per species ( $\mathcal{D}_s$  and  $\mathcal{D}_{s,0}$ ). The other two are datasets of multiple alignments that contain information on multiple individuals per species ( $\mathcal{D}_p$  and  $\mathcal{D}_{p,0}$ ). These four datasets are constructed such that each possesses desirable properties for certain strategies in the collection of 72 phylogenetic inference strategies, providing the strategies with as much information as possible to infer resolved phylogenies. One of the two datasets with a single individual sampled per species is optimized for locus-by-locus gene tree inference ( $\mathcal{D}_{s,0}$ ), whereas the other is optimized for gene tree inference from a concatenated alignment ( $\mathcal{D}_s$ ). Similarly, one of the two datasets with multiple individuals sampled per species is optimized for locus-by-locus gene tree inference ( $\mathcal{D}_{p,0}$ ), whereas the other is optimized for gene tree inference using multiple loci simultaneously ( $\mathcal{D}_p$ ). The procedures used for

constructing these datasets are described in subsequent sections on “Datasets with one individual per species” and “Datasets with multiple individuals per species”.

Let  $S_k$ ,  $k = 1, 2, \dots, 11$ , denote the set of individuals from pine species  $k$ , considering eight ingroup species  $(S_1, S_2, \dots, S_8)$  and three outgroup species  $(S_9, S_{10}, S_{11})$ . Denote the amount of overlapping non-gap non-missing sequence between a pair of individuals  $x$  and  $y$  by  $n_{xy}$  and denote the number of non-gap non-missing nucleotide differences between a pair of individuals  $x$  and  $y$  by  $d_{xy}$  ( $0 \leq d_{xy} \leq n_{xy}$ ). Further, denote the final dataset of  $L = 123$  loci by  $\mathcal{D} = \{\mathcal{A}_1, \mathcal{A}_2, \dots, \mathcal{A}_L\}$ , where  $\mathcal{A}_\ell$  is the set of aligned sequences at locus  $\ell$  for individuals from all 11 pine species. It is from this dataset  $\mathcal{D}$  that we create the four optimized datasets as summarized in Table 8.2 and Figure 8.2.

### 8.2.3.1 Datasets with one individual per species

The first dataset, denoted  $\mathcal{D}_s$ , consists of alignments with a single individual sampled per species at each locus (not necessarily the same individual across loci). That is, we generate a dataset of multiple alignments at each of  $L$  loci with only one individual per species, thereby creating multiple alignments of 11 “individuals”. This dataset is used by phylogenetic inference strategies that utilize the concatenation-based species tree construction methods with the maximum likelihood, maximum parsimony, and neighbor-joining gene tree inference methods (see sections “Inferring gene trees” and “Inferring species trees” for method details). To create  $\mathcal{D}_s$ , we choose the subset of 11 sequences  $\mathcal{A}_\ell^s$  at locus  $\ell$  by first maximizing the total amount of overlap sequence  $n(\mathcal{A}_\ell^s) = \sum_{x,y \in \mathcal{A}_\ell^s, x \neq y} n_{xy}$  and then, if there is a tie for the overlap  $n(\mathcal{A}_\ell^s)$ , maximizing the total number of substitutions  $d(\mathcal{A}_\ell^s) = \sum_{x,y \in \mathcal{A}_\ell^s, x \neq y} d_{xy}$ . In other words, for any other set of aligned sequences  $A \subseteq \mathcal{A}_\ell$  at locus  $\ell$  with only one individual sampled per species, the amount of overlapping non-gap non-missing sequence in  $A$  is no larger than in  $\mathcal{A}_\ell^s$ , *i.e.*,  $n(A) \leq n(\mathcal{A}_\ell^s)$ . Further, for any other



set of aligned sequences  $A \subseteq \mathcal{A}_\ell$  at locus  $\ell$  with only one individual sampled per species and  $n(A) = n(\mathcal{A}_\ell^s)$ , the total number of pairwise sequence differences in  $A$  is no larger than in  $\mathcal{A}_\ell^s$ , *i.e.*,  $d(A) \leq d(\mathcal{A}_\ell^s)$ . If multiple sets of 11 individuals share the same values of  $n$  and  $d$ , then we choose the set of 11 individuals randomly among the tied sets. We choose the “optimal” set of 11 individuals at each locus in this way both to maximize the sequence contributions of individual loci to the inference of gene trees (maximizing  $n$ ) and to maximize the potential for creating resolved gene trees (maximizing  $d$ ).

The second dataset, denoted  $\mathcal{D}_{s,0}$ , is a subset of  $\mathcal{D}_s$  with  $L_{s,0} \leq L$  loci that consists of only those loci in  $\mathcal{D}_s$  for which there exists at least one nucleotide difference between each distinct pair of species (other than pairs from distinct outgroups). This condition of at least one nucleotide difference between pairs of species provides the potential to construct bifurcating gene trees. Dataset  $\mathcal{D}_{s,0}$  is used by phylogenetic inference strategies that utilize consensus methods with maximum likelihood, maximum parsimony, and neighbor-joining (see sections “Inferring gene trees” and “Inferring species trees” for method details).

### 8.2.3.2 Datasets with multiple individuals per species

The third dataset, denoted  $\mathcal{D}_p$ , is identical to our starting dataset  $\mathcal{D}$ . Similarly to dataset  $\mathcal{D}_s$ , dataset  $\mathcal{D}_p$  contains all  $L$  loci in  $\mathcal{D}$ . In specifying dataset  $\mathcal{D}_p$ , however, we do not make the restriction that at each locus, only one individual is sampled per species. Thus, strategies that use  $\mathcal{D}_p$  consider all available sampled sequences when estimating species tree topologies. Dataset  $\mathcal{D}_p$  is used by phylogenetic inference strategies that employ the concatenation-based species tree construction methods with the neighbor-joining gene tree inference method using multiple individuals (see sections “Inferring gene trees” and “Inferring species trees” for method details).

Consider a bootstrapped dataset  $D$  of  $L$  loci sampled randomly with replacement

from  $\mathcal{D}_p$ . Define

$$\mathbf{P}_{ij}^{all} = \begin{cases} 0 & , i = j \\ \frac{\sum_{\mathcal{A}_\ell \in D} \sum_{x,y \in \mathcal{A}_\ell} d_{xy} \mathbf{1}_{\{x \in S_i, y \in S_j\}}}{\sum_{\mathcal{A}_\ell \in D} \sum_{x,y \in \mathcal{A}_\ell} n_{xy} \mathbf{1}_{\{x \in S_i, y \in S_j\}}} & , i \neq j, \end{cases} \quad (8.1)$$

where  $\mathbf{1}_{\{x \in S_i, y \in S_j\}}$  is an indicator random variable that equals 1 if  $x \in S_i$  and  $y \in S_j$  and 0 otherwise. The pairwise distance matrix defined by eq. 8.1 is used to estimate gene trees for all strategies applied to  $\mathcal{D}_p$ . Given a distinct pair of species  $S_i$  and  $S_j$ , the entry  $\mathbf{P}_{ij}^{all}$  represents the  $p$ -distance (fraction of nucleotide differences; *Felsenstein*, 2004) averaged over pairs of individuals, one from species  $i$  and the other from species  $j$ .

The fourth dataset, denoted  $\mathcal{D}_{p,0}$ , is a subset of  $\mathcal{D}_p$  with  $L_{p,0} \leq L$  loci. This subset consists of only those loci in  $\mathcal{D}_p$  for which there exists a pair of individuals in each distinct pair of species (other than pairs from distinct outgroups) with at least one nucleotide difference between them. Define

$$\mathbf{P}_{ij}^\ell = \begin{cases} 0 & , i = j \\ \frac{\sum_{x,y \in \mathcal{A}_\ell} d_{xy} \mathbf{1}_{\{x \in S_i, y \in S_j\}}}{\sum_{x,y \in \mathcal{A}_\ell} n_{xy} \mathbf{1}_{\{x \in S_i, y \in S_j\}}} & , i \neq j, \end{cases} \quad (8.2)$$

where  $\mathbf{1}_{\{x \in S_i, y \in S_j\}}$  is an indicator random variable that equals 1 if  $x \in S_i$  and  $y \in S_j$  and 0 otherwise. The numerator of  $\mathbf{P}_{ij}^\ell$  represents the number of pairwise sequence differences, summed over pairs of individuals, one sampled from species  $S_i$  and the other sampled from species  $S_j$ , at locus  $\ell$ . The denominator of  $\mathbf{P}_{ij}^\ell$  represents the total sum across pairs of individuals, one from  $S_i$  and the other from  $S_j$ , of the non-gap non-missing sequence shared between pairs of individuals at locus  $\ell$ . To construct  $\mathcal{D}_{p,0}$ , we create a subset of  $\mathcal{D}_p$  that consists only of those loci in  $\mathcal{D}_p$  for which there exists a nonzero  $p$ -distance (*i.e.*,  $\mathbf{P}_{ij}^\ell > 0$ ) between each distinct pair of species (excluding pairs from distinct outgroups). This dataset is utilized by phylogenetic inference strategies that employ consensus methods with gene trees

inferred by neighbor-joining using multiple individuals (see sections “Inferring gene trees” and “Inferring species trees” for method details). Similarly to dataset  $\mathcal{D}_{s,0}$ , this condition of a nonzero  $p$ -distance between pairs of species provides the potential to construct bifurcating gene trees.

#### 8.2.4 Inferring gene trees

For each of the four datasets  $\mathcal{D}_s$ ,  $\mathcal{D}_{s,0}$ ,  $\mathcal{D}_p$ , and  $\mathcal{D}_{p,0}$ , we inferred gene trees from bootstrap replicate datasets (*Efron and Tibshirani, 1993*) that contain loci randomly sampled with replacement from the dataset. For strategies applied to datasets  $\mathcal{D}_s$  and  $\mathcal{D}_{s,0}$ , we inferred gene trees from sequence alignments by applying either maximum likelihood (ML; *Felsenstein, 2004*, ch. 9) under a general time-reversible substitution model (*Felsenstein, 2004*, ch. 13), maximum parsimony (MP; *Felsenstein, 2004*, ch. 1), or neighbor-joining (NJ; *Felsenstein, 2004*, ch. 11) to a  $p$ -distance matrix calculated between pairs of alignments. For strategies applied to  $\mathcal{D}_p$  and  $\mathcal{D}_{p,0}$ , we inferred gene trees by applying neighbor-joining to the  $\mathbf{P}^{all}$  and  $\mathbf{P}^\ell$   $p$ -distance matrices, respectively. We term the method for inferring gene trees from the  $\mathbf{P}^{all}$  and  $\mathbf{P}^\ell$   $p$ -distance matrices “neighbor-joining using multiple individuals” (NJM). All gene trees were inferred using PAUP\* (*Swofford, 2003*).

#### 8.2.5 Inferring species trees

The six species tree construction methods used in this study are Concatenation (*Rokas et al., 2003; de Queiroz and Gatesy, 2007*), SuperMatrix Rooted Triple (SMRT; *DeGiorgio and Degnan, 2010*), STEAC (*Liu et al., 2009*), STAR (*Liu et al., 2009*), Rooted Triple Consensus (RTC; *Ewing et al., 2008*), and Minimize Deep Coalescences (MDC; *Maddison, 1997; Than and Nakhleh, 2009*). Concatenation and SMRT are concatenation-based and STEAC, STAR, RTC, and MDC are consensus methods. These methods were chosen to represent a range of species tree construction

methods suitable for rapid computation in the variety of scenarios that were of interest to consider. Due to their computational burden, we did not investigate Bayesian methods.

Consider a set of  $L$  loci (multiple alignments) with  $m$  ingroup species and an outgroup species. Concatenation methods concatenate the  $L$  multiple alignments to create a single “super locus” consisting of a multiple alignment of the  $m + 1$  species across  $L$  loci. From this multiple alignment, a gene tree is inferred and this gene tree is taken as the species tree estimate. Similarly, SMRT creates a concatenated alignment of the  $m + 1$  species from a set of  $L$  multiple alignments. However, SMRT then constructs from this concatenated alignment all  $\binom{m}{3}$  concatenated alignments of three ingroup species and an outgroup species. Rooted three-taxon gene trees are then inferred from each of the  $\binom{m}{3}$  concatenated alignments. A supertree algorithm is then applied to the set of rooted three-taxon gene trees to estimate an  $m$ -taxon species tree topology. This study uses the Modified Mincut supertree algorithm implemented in the program SUPERTREE (Page, 2002) to construct a species tree from rooted three-taxon gene trees.

Consider a set of  $(m + 1)$ -taxon gene trees ( $m$  ingroup and one outgroup species) inferred at each of  $L$  loci. STEAC estimates a species tree topology by using estimated mean coalescence times. For distinct species  $S_i$  and  $S_j$ , the mean coalescence time is computed as the estimated divergence time for  $S_i$  and  $S_j$ , averaged over all  $L$  gene trees. These mean coalescence times specify a distance between each pair of species and are placed into a distance matrix. Neighbor-joining is applied to the matrix to estimate the species tree topology.

STAR estimates a species tree topology by using average coalescence ranks. STAR assumes that the rank of the root of a gene tree is equal to the number of species in the tree ( $m + 1$  in our case). Internal nodes of a gene tree are then assigned the rank of the node directly ancestral to it minus one. For distinct species  $S_i$  and  $S_j$ , the

average coalescence rank is computed as the rank of the node that connects  $S_i$  and  $S_j$ , averaged over all  $L$  gene trees. Similarly to the algorithm of STEAC, these average coalescence ranks specify a distance between each pair of species and are placed into a distance matrix. Neighbor-joining is applied to the matrix to estimate the species tree topology.

RTC estimates a species tree by using rooted three-taxon tree topologies. At each locus  $\ell$ ,  $\ell = 1, 2, \dots, L$ , RTC finds the set of  $\binom{m}{3}$  rooted tree topologies of three ingroup and one outgroup species that are displayed by the inferred gene tree at locus  $\ell$ . After applying this procedure to all loci, RTC applies the quartet puzzling algorithm (*Strimmer and von Haeseler, 1996*) to the set of  $\binom{m}{3}L$  rooted three-taxon tree topologies to estimate the species tree topology.

A coalescence event between a pair of lineages is considered “deep” if the coalescence does not occur in the first population in which the pair of lineages is capable of coalescing. Given a gene tree, the number of deep coalescences on a species tree is defined as the total number of “extra lineages”, summed across branches of the species tree topology, that is needed to fit the gene tree within the species tree topology. Here, the number of extra lineages for a branch is one fewer than the number of lineages that survive to the root of the branch; if incomplete lineage sorting does not occur, then only one lineage persists from a branch to its ancestor, and there are no extra lineages. For a set of  $L$  gene trees, the number of deep coalescences for a species tree is the total number of deep coalescences for the species tree given a gene tree, summed across the  $L$  gene trees. MDC estimates a species tree topology by minimizing the number of deep coalescences. That is, MDC finds a species tree topology for which the number of deep coalescences that will fit the set of  $L$  gene trees within the species tree topology is minimal.

### 8.2.6 Multivariate analysis

In our study, we want to determine which of the 72 phylogenetic inference strategies perform similarly. Consider a  $72 \times 145$ -dimensional data matrix  $\mathbf{S}$  in which rows represent the 72 strategies and columns represent 145 observed clades, among the  $\sum_{k=2}^{8-1} \binom{8}{k} = 246$  possible non-trivial clades of eight species. Entry  $\mathbf{S}_{ij}$  in column  $i$  and row  $j$  of  $\mathbf{S}$  is the number of times that strategy  $i$  infers clade  $j$  in 1000 bootstrap replicates across loci.

Principal components analysis (PCA) was applied to  $\mathbf{S}$  to create a  $72 \times 2$ -dimensional matrix  $\mathbf{V}$  in which the rows represent the 72 strategies and the first and second columns represent the first and second principal components, respectively. Plotting 72 strategies onto the space defined by the first and second principal components yields a two-dimensional spatial “map” of phylogenetic inference strategies.

To compare spatial maps of phylogenetic inference strategies, we used Procrustes analysis (*Dryden and Mardia, 1998; Cox and Cox, 2001; Gower and Dijksterhuis, 2004*). In particular, we compared the spatial distribution of a subset of  $72 - r$  strategies when analyzed alone to the spatial distribution for all 72 strategies. The comparison enabled us to quantify the influence that a set of  $r$  strategies with a particular feature (*i.e.*, species tree construction method, gene tree inference method, or outgroup species) has on the spatial distribution of all 72 strategies. Consider a set of strategies  $\Sigma = \{\sigma_1, \sigma_2, \dots, \sigma_{72-r}\}$  that is a proper subset of the full set of 72 strategies. Consider a  $(72 - r) \times 145$ -dimensional data matrix  $\mathbf{S}_\Sigma$  in which rows represent the  $72 - r$  strategies in set  $\Sigma$  and columns represent 145 observed clades (*i.e.*,  $\mathbf{S}_\Sigma$  is a submatrix of  $\mathbf{S}$ , in which the  $72 - r$  rows corresponding to strategies in  $\Sigma$  are selected from  $\mathbf{S}$ ). Consider a  $(72 - r) \times 2$ -dimensional target matrix  $\mathbf{X}$  and a  $(72 - r) \times 2$ -dimensional comparison matrix  $\mathbf{Y}$ .  $\mathbf{X}$  is matrix  $\mathbf{V}$  restricted to the set of strategies  $\Sigma$ .  $\mathbf{Y}$  is a matrix representing the first two principal components in the PCA

applied to matrix  $\mathbf{S}_\Sigma$ . Now consider a  $(72 - r) \times 2$ -dimensional matrix  $\mathbf{Z} = b\mathbf{Y}\mathbf{T} + \mathbf{C}$  that is a transformation of  $\mathbf{Y}$ , where  $b$  is a scaling factor,  $\mathbf{T}$  is a  $2 \times 2$ -dimensional matrix that rotates and reflects  $\mathbf{Y}$ , and  $\mathbf{C}$  is a  $(72 - r) \times 2$ -dimensional matrix that has constant columns and that is used to translate the matrix. Procrustes analysis seeks to find  $b$ ,  $\mathbf{T}$ , and  $\mathbf{C}$  to minimize the sum of squared differences between  $\mathbf{X}$  and some  $(72 - r) \times 2$ -dimensional matrix  $\mathbf{Z}^* = b\mathbf{Y}\mathbf{T} + \mathbf{C}$ . That is,  $\mathbf{Z}^*$  is formally defined as  $\mathbf{Z}^* = \operatorname{argmin}_{\mathbf{Z}} \{ \sum_{i=1}^{72-r} \sum_{j=1}^2 (\mathbf{X}_{ij} - \mathbf{Z}_{ij})^2 \}$ . Then the dissimilarity measure between  $\mathbf{X}$  and  $\mathbf{Z}^*$  is computed as

$$\frac{\sum_{i=1}^{72-r} \sum_{j=1}^2 (\mathbf{X}_{ij} - \mathbf{Z}_{ij}^*)^2}{\sum_{i=1}^{72-r} \sum_{j=1}^2 (\mathbf{X}_{ij} - \mu_j)^2}, \quad (8.3)$$

where  $\mu_j = \frac{1}{72-r} \sum_{i=1}^{72-r} \mathbf{X}_{ij}$  is the  $j$ th dimension of the centroid of  $\mathbf{X}$ , computed across all row vectors of  $\mathbf{X}$ . This measure takes the sum of squared differences between points on the spatial maps defined by  $\mathbf{X}$  and  $\mathbf{Z}^*$  and normalizes it by the sum of squared differences between the points on the spatial map defined by  $\mathbf{X}$  and the centroid of those points.

Define a cluster as a set of strategies and let the centroid of a cluster of strategies be the location in the 145-dimensional space of clades whose coordinates are the means of those of all strategies in the cluster. Hierarchical clustering was performed by first creating a matrix of Euclidean distances between all  $\binom{72}{2}$  pairs of 145-dimensional vectors represented by the matrix  $\mathbf{S}$ . Define the within-cluster sum of squared Euclidean distance as the squared Euclidean distance between a point in a cluster and the centroid of the cluster, summed over all points in the cluster. From the  $72 \times 72$ -dimensional matrix of Euclidean distances between strategies, a dendrogram relating the 72 strategies was constructed using the Ward algorithm (*Ward*, 1963). The Ward algorithm iteratively merges clusters until all points are contained within a single cluster. The nesting of clusters created by the Ward algorithm defines the

dendrogram among a set of points. For a given iteration, two clusters are merged if their merged cluster has a smaller within-cluster sum of squared Euclidean distances than any other potential merged cluster.

We performed  $K$ -means clustering on the 72 145-dimensional vectors, using  $K$  clusters,  $K = 2, 3, \dots, 9$ . Given a number of clusters  $K$ , the 72 strategies were separated into  $K$  clusters on the basis of the squared Euclidean distance between all pairs of the 72 strategies in a 145-dimensional space. For each  $K$ , we ran  $10^4$  replicates with random starting locations. For specified  $K$ , each replicate yielded a total within-cluster sum of squared distances for the set of  $K$  clusters. This total within-cluster sum of squared distances represents the within-cluster sum of squared distances between points in a cluster and the cluster centroid, summed over all  $K$  clusters. For a given  $K$ , we chose the set of cluster assignments that had the minimum total within-cluster sum of squared distances, where the minimum was taken over all  $10^4$  replicate starting locations.

To compute the correlation coefficient between a pair of strategies, we calculated the Pearson correlation coefficient between the pair by only using points in the 145-dimensional vector that were nonzero in both strategies being compared.

### 8.3 Results

We accounted for the variable outcomes of individual phylogenetic inference strategies by applying the strategies to bootstrap (*Efron and Tibshirani, 1993*) datasets instead of their respective full datasets. Our analysis identified 145 distinct clades observed in the set of 72 phylogenetic inference strategies, among 246 possible non-trivial clades on eight species, across 1000 bootstrap replicates for each strategy. From these clades, we created a  $72 \times 145$  matrix  $\mathbf{S}$  in which each row is a phylogenetic inference strategy and each column is a clade. The value of  $\mathbf{S}_{ij}$ , the cell in row  $i$  and column  $j$ , is the number of times among the 1000 bootstrap replicates that strategy



$i$  inferred a species tree with clade  $j$ . This summarized dataset  $\mathbf{S}$  of clade counts was used for all further analyses.

### 8.3.1 Clade size

We first investigated the level of balance (*Sackin*, 1972; *Colless*, 1982; *Shao and Sokal*, 1990; *Kirkpatrick and Slatkin*, 1993) in the tree topologies inferred by each phylogenetic inference strategy. The distribution of clade sizes (number of taxa within a clade) provides a basis for measuring tree topological balance. Topologies with numerous small clades tend to be more balanced than topologies with large clades. For example, consider the topologies  $T_{bal} = (((AB)(CD))((EF)(GH)))$  and  $T_{unbal} = (((((((AB)C)D)E)F)G)H)$ . Topology  $T_{bal}$  is the most balanced eight-taxon topology whereas  $T_{unbal}$  is the most unbalanced eight-taxon topology. Considering non-trivial clades,  $T_{bal}$  has four clades of size two and two clades of size four.  $T_{unbal}$  has one clade each of size two, three, four, five, six, and seven. Thus, the clades of  $T_{bal}$  are smaller than those of  $T_{unbal}$ . The mean clade size for  $T_{bal}$  is  $\sim 2.67$  and the mean clade size for  $T_{unbal}$  is 4.5.

Figure 8.3A displays the cumulative distribution of clade sizes for each of the 72 phylogenetic inference strategies, considering all 1000 bootstrap replicate species trees for each strategy. The cumulative distribution of clade sizes increases most quickly for strategies based on MDC, for which most of the distribution is located in clades of size two. In contrast, the cumulative distribution of clade sizes increases most slowly for strategies based on SMRT and STEAC, for which much of the probability distribution is located in clades of size six and seven. Figure 8.3B displays a bar graph of the mean clade size for each of the 72 phylogenetic inference strategies. This graph shows that among all six species tree construction methods, the 12 MDC strategies have the smallest mean clade size as well as the smallest variance in mean clade size across the 12 combinations of outgroup and gene tree inference method. In

contrast, SMRT and STEAC in general have the largest mean clade size. However, all 12 SMRT strategies infer trees with large mean clade size, whereas the mean clade size of STEAC varies across the 12 combinations of outgroup and gene tree inference method. Interestingly, the mean clade size averaged over all 12 strategies based on MDC is  $\sim 2.79$ , a value that is close to the mean clade size for  $T_{bal}$  of  $\sim 2.67$ .

### 8.3.2 Clustering of strategies

We next used PCA, hierarchical clustering,  $K$ -means clustering, and correlation analysis on the matrix of clades  $\mathbf{S}$  to identify phylogenetic inference strategies that perform similarly. Figure 8.4 displays plots of the first two principal components, which account for 38.94% and 18.96% of the variation across strategies, respectively. Figure 8.4A shows that separate clusters are formed by strategies that are based on Concatenation, SMRT, and STEAC, and that strategies based on STAR, RTC, and MDC form a cluster together. Further, a large “super cluster” is formed by strategies that are based on Concatenation, SMRT, and STEAC, and another large super cluster is formed by strategies that are based on STAR, RTC, and MDC. These super clusters have a nice interpretation in that one super cluster contains topologically-based strategies (STAR, RTC, and MDC) and the other super cluster contains strategies that are not strictly topologically-based (Concatenation, SMRT, and STEAC). Strategies are classified as topologically-based if they only use information on tree topologies to construct a species tree. In contrast, strategies are classified as not strictly topologically-based if they use information other than the gene tree topologies, such as sequence or branch length information, to construct a species tree. Relabeling the points in Figure 8.4A according to gene tree inference method, Figure 8.4B shows that strategies that are based on NJM form a cluster, and that there are no separate clusters for strategies that are based on ML, MP, or NJ. Figure 8.4C, which labels points according to outgroup, shows that no strategies

separate into clusters based on the choice of outgroup.

From Figure 8.4, we can see that much of the variation across the 72 phylogenetic inference strategies, as explained by PCA, is caused by NJM. Strategies based on NJM are more similar in clade outcomes to other strategies based on NJM than they are to other strategies that are not based on NJM. The magnitude of this effect can be quantified using Procrustes analysis, which demonstrates that NJM has a large influence on the spatial relationship among all other phylogenetic inference strategies (Figure 8.5). Interestingly, the cluster formed by NJM separates into two sub-clusters, one cluster of strategies based on MDC and another cluster of strategies not based on MDC. These sub-clusters are not surprising because of the distinctive bias that MDC exhibits toward balanced tree topologies (Figure 8.3).

Figure 8.6 shows the results of our cluster and correlation analyses. The main clusters formed by phylogenetic inference strategies involve strategies based on the species tree construction methods Concatenation, SMRT, STEAC, and MDC or the gene tree inference method NJM (Figure 8.6). The clusters of strategies formed by  $K$ -means and the large groupings of strategies formed by hierarchical clustering are quite similar. Additionally, the correlation coefficient between clade vectors inferred by pairs of phylogenetic inference strategies is generally higher for pairs of strategies that are placed into the same cluster by either  $K$ -means or hierarchical clustering than for pairs of strategies that are not placed into the same cluster by either  $K$ -means or hierarchical clustering (Figure 8.6).

Interestingly, the clustering of strategies found by PCA in Figure 8.4 matches well with the clusters and groupings observed in Figure 8.6. In Figure 8.6, three large clusters are formed and are represented by the subtree to the left of the root of the dendrogram and the two subtrees that are rooted by the subtree to the right of the root of the dendrogram. These clusters correspond closely to the blue, pink, and orange colors in the  $K$ -means clustering with  $K = 3$ . The two subtrees to the

right of the root (or pink and orange clusters defined by  $K$ -means clustering) involve strategies that are based on NJM (pink  $K$ -means cluster or left subtree of the right subtree of the dendrogram) or strategies that are based on species tree construction methods that are topologically-based (orange  $K$ -means cluster or right subtree on the right subtree of the dendrogram). That is, strategies that correspond to the orange cluster are based on either STAR, RTC, or MDC. In contrast, the subtree to the left of the root (or the blue cluster defined by  $K$ -means clustering) contains only strategies that use species tree construction methods that are not strictly topologically-based (*i.e.*, Concatenation, SMRT, or STEAC).

From the results of Figures 8.4 and 8.6, we find that phylogenetic inference strategies form three basic clusters: a cluster that involves strategies that are based on NJM, a cluster that involves strategies that are topologically-based, and a cluster that involves strategies that are not strictly topologically-based.

### 8.3.3 Clade flow

In this section, we identify phylogenetic inference strategies that do and do not robustly infer clades that are supported by the other strategies. Figure 8.7 displays a heat map that represents the “flow” of clades between phylogenetic inference strategies. The cell at row  $i$  and column  $j$  in the heat map represents the fraction of clades inferred by strategy  $i$  that were not inferred by strategy  $j$ . As can be seen from the mostly white and yellow boxes for rows corresponding to strategies based on NJM, the heat map shows that strategies based on NJM tend to infer clades that are supported by other strategies. That is, if a species tree topology is inferred by a strategy that is based on NJM, then clades displayed by that topology will often also be present on species tree topologies inferred by other strategies. In Figure 8.5, we found that strategies based on NJM contribute to the most variation across strategies. A possible explanation for this observation is that the flow of clades

is largely unidirectional. That is, if a strategy is based on NJM, then clades that are inferred by that strategy also tend to be supported by other strategies; however, if a strategy not based on NJM infers a clade, then that clade is not frequently supported by strategies based on NJM. Because clades that are inferred by strategies based on NJM also tend to be supported by other strategies, it follows that strategies based on NJM tend to infer clades that are also supported by other strategies that are based on NJM. This sharing of clades among strategies based on NJM causes those strategies to be more similar to each other than they are to strategies not based on NJM. In contrast to the results for NJM, as can be seen from the mostly dark boxes for rows corresponding to strategies based on MDC, strategies based on MDC tend to infer clades that are not supported by other strategies (especially when compared with strategies based on NJM).

Similarly to the behavior of MDC, strategies that are based on Concatenation, SMRT, and STEAC together with ML, MP, or NJ share more clades with other such strategies (mostly white and yellow boxes) than with the remaining strategies (mostly dark boxes). In contrast, as was observed with NJM, strategies based on STAR and RTC together with ML, MP, or NJ share similar numbers of clades among other such strategies as with the remaining strategies (mostly yellow boxes). These results suggest that strategies that are topologically-based (*i.e.*, STAR and RTC) tend to infer clades that are also supported both by other strategies that are topologically-based and by strategies that are not strictly topologically-based, whereas strategies that are not strictly topologically-based (*i.e.*, Concatenation, SMRT, and STEAC) tend to infer clades that are not supported by strategies that are strictly topologically-based (*i.e.*, STAR, RTC, and MDC).

### 8.3.4 Representative topologies

We next wanted to use a set of representative species tree topologies to highlight similarities and differences in topologies constructed by various strategies. Topologies were estimated using the Greedy Consensus algorithm (*Bryant, 2003*) applied to clade counts. Because our previous results (Figures 8.4-8.6) indicate that the choice of outgroup species does not strongly influence the overall inferred topologies, it is sensible to average across outgroups. Therefore, we first present topologies for each of the 24 species tree-gene tree inference method pairs constructed from clade counts that were averaged over the three outgroups (Figure 8.8). Next, to obtain a clearer picture of the types of topologies that are inferred by the six species tree inference methods, we present topologies for each of the six species tree inference methods, constructed from clade counts that were averaged over gene tree inference methods and outgroup species (Figure 8.9). Finally, to assess the influence that various gene tree inference methods have on the overall inferred species tree topology, we present topologies for each of the four gene tree inference methods, constructed from clade counts that were averaged over species tree inference methods and outgroup species (Figure 8.10).

Figure 8.8 displays 24 topologies with clade support values for each combination of a species tree construction method and a gene tree inference method. The clade  $\{P. chiapensis, P. strobilus\}$  is generally highly supported, appearing for 22 of 24 strategies, with support ranging from 382 to 982 among 1000 bootstrap replicates. The smallest support values for  $\{P. chiapensis, P. strobilus\}$  occur in strategies that use SMRT with ML, MP, and NJ, producing support values of 382, 406, and 395, respectively. The largest support values for  $\{P. chiapensis, P. strobilus\}$  occur in strategies that use NJM, with values ranging from 824 to 982. Further, although strategies based on SMRT with ML, MP, and NJ yield lower support values than other strategies, when SMRT is combined with NJM, the support for  $\{P. chiapensis, P. strobilus\}$  is

905. In addition, although two of the strategies based on STEAC do not support the clade  $\{P. chiapensis, P. strobilus\}$ , when STEAC is combined with NJM, the support for the clade is 982. Another clade that is highly supported is  $\{P. ayacahuite, P. flexilis, P. strobiformis\}$ . This clade is observed across all strategies, with support among non-MDC strategies out of 1000 bootstrap replicates ranging from 858 to 1000. Strategies that use MDC with ML, MP, and NJ yield support values for  $\{P. ayacahuite, P. flexilis, P. strobiformis\}$  of 560, 407, and 415, respectively. However, using MDC with NJM yields a support value of 933 for  $\{P. ayacahuite, P. flexilis, P. strobiformis\}$ . Across the 24 trees, the topological positions of *P. albicaulis*, *P. lambertiana*, and *P. monticola* are variable and are generally poorly supported. Each of these species is found in a variety of positions across all trees.

Figure 8.9 displays six topologies with clade support values for each species tree construction method. Similarly to Figure 8.8, the clade  $\{P. chiapensis, P. strobilus\}$  is generally highly supported across all six species tree construction methods, with support ranging from 522 to 876 among 1000 bootstrap replicates. Also, as in Figure 8.8, the clade  $\{P. ayacahuite, P. flexilis, P. strobiformis\}$  is highly supported across all six species tree construction methods, with support ranging from 579 to 999 among 1000 bootstrap replicates. From these topologies, we can also observe that in agreement with the clade size distribution, strategies based on Concatenation, SMRT, and STEAC tend to produce more unbalanced trees than strategies based on STAR, RTC, and MDC (Figure 8.3). Strategies based on Concatenation, SMRT, and STEAC support topologies in which *P. lambertiana* is on the opposite side of the root from the other seven species. In contrast, strategies based on STAR, RTC, and MDC place *P. monticola* and *P. lambertiana* as sister species. These results support the observations from Figures 8.4, 8.6, and 8.7 that strategies based on species tree construction methods that are topologically-based behave differently from strategies that are not strictly topologically-based.

Figure 8.10 displays four topologies with clade support values, considering each gene tree inference method and combining species tree construction methods for each gene tree inference method. Similar to Figures 8.8 and 8.9, the clades  $\{P. \textit{chiapensis}, P. \textit{strobilus}\}$  and  $\{P. \textit{ayacahuite}, P. \textit{flexilis}, P. \textit{strobiformis}\}$  are generally highly supported across all four gene tree inference methods, with supports among 1000 bootstrap replicates respectively ranging from 610 to 931 and from 858 to 988.

## 8.4 Discussion

In this article, we have empirically evaluated strategies for inferring species tree topologies from multilocus sequence data. We have found that MDC tends to infer balanced topologies, whereas SMRT and STEAC tend to infer more unbalanced topologies. This bias toward balanced topologies exhibited by MDC is a consequence of the nature of the criterion that MDC uses to construct species trees. Because balanced trees have fewer nodes between their root and their leaves than do unbalanced trees, they have fewer opportunities for incomplete lineage sorting events to occur. With fewer opportunities for incomplete lineage sorting, the expected number of deep coalescence events is smaller than for unbalanced trees. Thus, because MDC infers trees by minimizing the number of deep coalescences among species tree candidates, it is likely that MDC will be biased toward choosing balanced topologies.

The strategies that we have examined fall into three classes in terms of the species tree inferences they produce: strategies that use information on all available sequenced individuals (*i.e.*, NJM), strategies that are topologically-based (*i.e.*, STAR, RTC and MDC), and strategies that are not strictly topologically-based (*i.e.*, Concatenation, SMRT, and STEAC). This result is surprising because the strategies that are not strictly topologically-based take quite different approaches to construct species trees (*i.e.*, STEAC is a consensus method whereas Concatenation and SMRT are concatenation-based methods). We also found that strategies that are



topologically-based tend to support clades that are also supported by other strategies, whereas strategies that are not strictly topologically-based tend to infer clades that are not supported by strategies that are topologically-based. A possible reason for this observation could be differences in mutation rate at various loci. That is, if a phylogenetic inference strategy is not strictly topologically-based, then a single locus can have a strong influence on the overall species tree inferred by that strategy. For example, because STEAC uses branch length information to infer a species tree topology, if the mutation rate is high at a locus, then the branch lengths of the gene tree inferred at that locus will likely be large (*i.e.*, sequences will have many mutations and, hence, estimated gene trees will have large divergence times). As a consequence, these large branch lengths could skew the average branch lengths that are used by STEAC to construct a species tree topology. Also, because both Concatenation and SMRT combine all loci and treat this combination of loci as one large superlocus, a locus with many mutations could have considerable influence on the species tree that is inferred by either of these two methods.

Our analyses have highlighted several important characteristics of phylogenetic inference strategies that enable us to provide recommendations for inferring rooted phylogenies from large-scale multilocus data. First, it is beneficial to examine multiple strategies, considering some methods that use only topological information (*e.g.*, STAR, RTC, and MDC) and others that use more than just topological information (*e.g.*, Concatenation, SMRT, and STEAC). If species tree topologies returned by these different classes of species tree construction methods are identical, then an investigator can be more confident in the inferred tree topology. Second, species estimates should probably not be based solely on species tree construction methods that appear to be biased toward certain types of topologies (*e.g.*, MDC). Instead, it is preferable to utilize these types of species tree construction methods in conjunction with other methods. For example, after obtaining an unbalanced inferred tree from

an inference method, if MDC also infers the same unbalanced topology, then we might feel confident that the true species topology is actually unbalanced. Finally, it is best to utilize as much information as is available on individuals at every locus. That is, if multiple individuals are sampled within species at a given locus, then we should use all available sequence data from the species (*e.g.*, NJM). This point is supported by the observation that clades inferred by NJM tend to “flow” to other strategies (Figure 8.7).

## 8.5 Acknowledgments

We thank Zach Szpiech, Cuong Than, and Chaolong Wang for helpful discussions and Ben Figueroa, Ismael Grachico, Erik Grimstad, Brian Knaus, John Liechty, and Jill Wegrzyn for their help with data generation and analysis. This work was supported by NSF grants DEB-0716904 and DBI-0638502, NIH training grants T32 GM070449 and T32 HG000040, a University of Michigan Rackham Merit Fellowship, and the Murdock College Science Research Program.

Table 8.1: Phylogenetic inference strategies

Index	Species inference method	tree	Gene tree inference method	Outgroup species	Concatenation or consensus	Strictly topology-based
1	Concatenation		ML	<i>P. gerardiana</i>	Concatenation	No
2	Concatenation		ML	<i>P. bungeana</i>	Concatenation	No
3	Concatenation		ML	<i>P. squamata</i>	Concatenation	No
4	Concatenation		MP	<i>P. gerardiana</i>	Concatenation	No
5	Concatenation		MP	<i>P. bungeana</i>	Concatenation	No
6	Concatenation		MP	<i>P. squamata</i>	Concatenation	No
7	Concatenation		NJ	<i>P. gerardiana</i>	Concatenation	No
8	Concatenation		NJ	<i>P. bungeana</i>	Concatenation	No
9	Concatenation		NJ	<i>P. squamata</i>	Concatenation	No
10	Concatenation		NJM	<i>P. gerardiana</i>	Concatenation	No
11	Concatenation		NJM	<i>P. bungeana</i>	Concatenation	No
12	Concatenation		NJM	<i>P. squamata</i>	Concatenation	No
13	SMRT		ML	<i>P. gerardiana</i>	Concatenation	No
14	SMRT		ML	<i>P. bungeana</i>	Concatenation	No
15	SMRT		ML	<i>P. squamata</i>	Concatenation	No
16	SMRT		MP	<i>P. gerardiana</i>	Concatenation	No
17	SMRT		MP	<i>P. bungeana</i>	Concatenation	No
18	SMRT		MP	<i>P. squamata</i>	Concatenation	No
19	SMRT		NJ	<i>P. gerardiana</i>	Concatenation	No
20	SMRT		NJ	<i>P. bungeana</i>	Concatenation	No
21	SMRT		NJ	<i>P. squamata</i>	Concatenation	No
22	SMRT		NJM	<i>P. gerardiana</i>	Concatenation	No
23	SMRT		NJM	<i>P. bungeana</i>	Concatenation	No
24	SMRT		NJM	<i>P. squamata</i>	Concatenation	No

Index	Species inference method	tree method	Gene tree inference method	Outgroup species	Concatenation or consensus	Strictly topology-based
25	STEAC		ML	<i>P. gerardiana</i>	Consensus	No
26	STEAC		ML	<i>P. bungeana</i>	Consensus	No
27	STEAC		ML	<i>P. squamata</i>	Consensus	No
28	STEAC		MP	<i>P. gerardiana</i>	Consensus	No
29	STEAC		MP	<i>P. bungeana</i>	Consensus	No
30	STEAC		MP	<i>P. squamata</i>	Consensus	No
31	STEAC		NJ	<i>P. gerardiana</i>	Consensus	No
32	STEAC		NJ	<i>P. bungeana</i>	Consensus	No
33	STEAC		NJ	<i>P. squamata</i>	Consensus	No
34	STEAC		NJM	<i>P. gerardiana</i>	Consensus	No
35	STEAC		NJM	<i>P. bungeana</i>	Consensus	No
36	STEAC		NJM	<i>P. squamata</i>	Consensus	No
37	STAR		ML	<i>P. gerardiana</i>	Consensus	Yes
38	STAR		ML	<i>P. bungeana</i>	Consensus	Yes
39	STAR		ML	<i>P. squamata</i>	Consensus	Yes
40	STAR		MP	<i>P. gerardiana</i>	Consensus	Yes
41	STAR		MP	<i>P. bungeana</i>	Consensus	Yes
42	STAR		MP	<i>P. squamata</i>	Consensus	Yes
43	STAR		NJ	<i>P. gerardiana</i>	Consensus	Yes
44	STAR		NJ	<i>P. bungeana</i>	Consensus	Yes
45	STAR		NJ	<i>P. squamata</i>	Consensus	Yes
46	STAR		NJM	<i>P. gerardiana</i>	Consensus	Yes
47	STAR		NJM	<i>P. bungeana</i>	Consensus	Yes
48	STAR		NJM	<i>P. squamata</i>	Consensus	Yes

Index	Species inference method	tree method	Gene tree inference method	Outgroup species	Concatenation or consensus	Strictly topology-based
49	RTC		ML	<i>P. gerardiana</i>	Consensus	Yes
50	RTC		ML	<i>P. bungeana</i>	Consensus	Yes
51	RTC		ML	<i>P. squamata</i>	Consensus	Yes
52	RTC		MP	<i>P. gerardiana</i>	Consensus	Yes
53	RTC		MP	<i>P. bungeana</i>	Consensus	Yes
54	RTC		MP	<i>P. squamata</i>	Consensus	Yes
55	RTC		NJ	<i>P. gerardiana</i>	Consensus	Yes
56	RTC		NJ	<i>P. bungeana</i>	Consensus	Yes
57	RTC		NJ	<i>P. squamata</i>	Consensus	Yes
58	RTC		NJM	<i>P. gerardiana</i>	Consensus	Yes
59	RTC		NJM	<i>P. bungeana</i>	Consensus	Yes
60	RTC		NJM	<i>P. squamata</i>	Consensus	Yes
61	MDC		ML	<i>P. gerardiana</i>	Consensus	Yes
62	MDC		ML	<i>P. bungeana</i>	Consensus	Yes
63	MDC		ML	<i>P. squamata</i>	Consensus	Yes
64	MDC		MP	<i>P. gerardiana</i>	Consensus	Yes
65	MDC		MP	<i>P. bungeana</i>	Consensus	Yes
66	MDC		MP	<i>P. squamata</i>	Consensus	Yes
67	MDC		NJ	<i>P. gerardiana</i>	Consensus	Yes
68	MDC		NJ	<i>P. bungeana</i>	Consensus	Yes
69	MDC		NJ	<i>P. squamata</i>	Consensus	Yes
70	MDC		NJM	<i>P. gerardiana</i>	Consensus	Yes
71	MDC		NJM	<i>P. bungeana</i>	Consensus	Yes
72	MDC		NJM	<i>P. squamata</i>	Consensus	Yes

Table 8.2: Datasets

Dataset	Strategies that use the dataset	Number of strategies	Description
$\mathcal{D}_s$	Concatenation or SMRT with either ML, MP, or NJ	18	Consists of all 123 loci, with a single individual sampled from each of 11 species at each locus.
$\mathcal{D}_{s,0}$	STEAC, STAR, RTC, or MDC with either ML, MP, or NJ	36	Subset of $\mathcal{D}_s$ requiring that each locus has at least one sequence difference between each distinct pair of species other than pairs from distinct outgroups.
$\mathcal{D}_p$	Concatenation or SMRT with NJM	6	Consists of the full dataset $\mathcal{D}$ , which contains all individuals and all loci.
$\mathcal{D}_{p,0}$	STEAC, STAR, RTC, or MDC with NJM	12	Subset of $\mathcal{D}_p$ requiring that each locus has at least one sequence difference between each distinct pair of species other than pairs from distinct outgroups.

Note that the total number of strategies sums to 72.

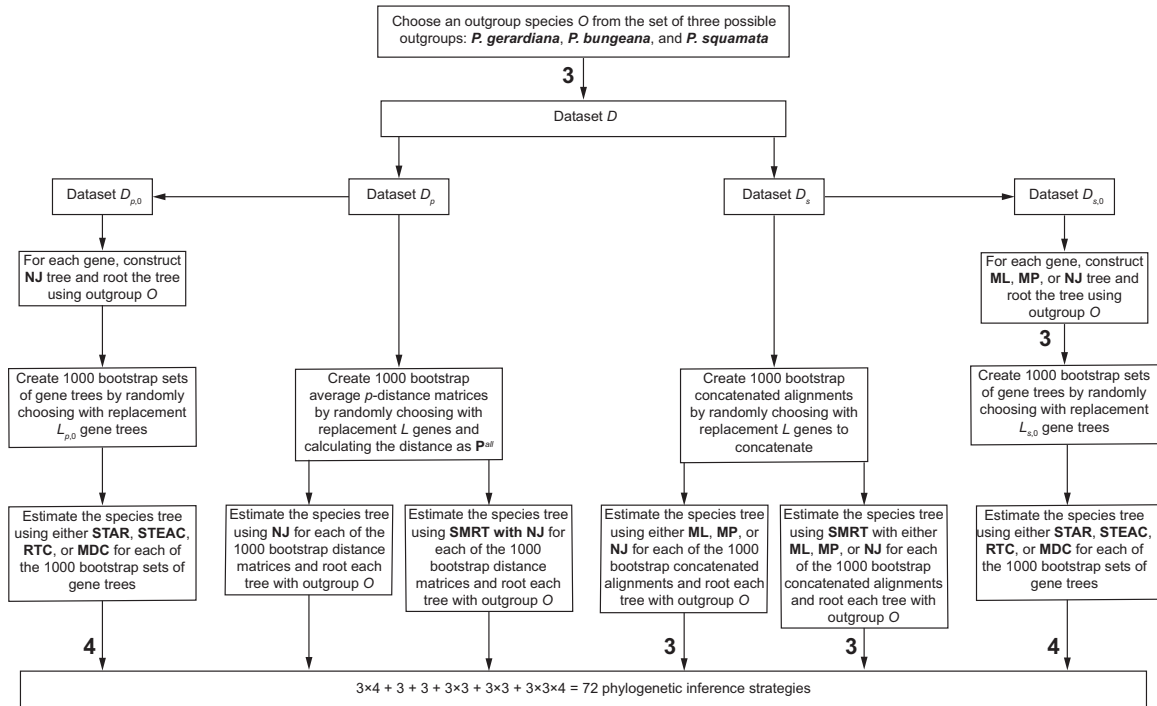


Figure 8.1: Flow diagram representing the procedure in which we obtained results on the behavior of phylogenetic inference strategies. A boldface number attached to a downward arrow indicates the number of phylogenetic inference strategies that are generated by the box immediately above the arrow. Absence of a number indicates a value of 1. The number of phylogenetic inference strategies for a particular path from the topmost box to the bottommost box of the diagram is calculated as the product of the boldface numbers visited during the traversal of the path. The number of phylogenetic inference strategies analyzed is 72, the sum over all paths from the topmost to the bottommost box in the diagram.

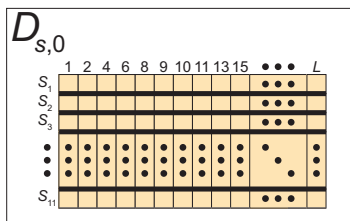
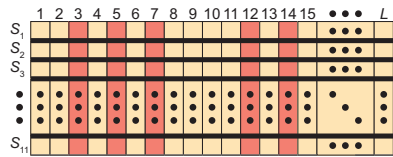
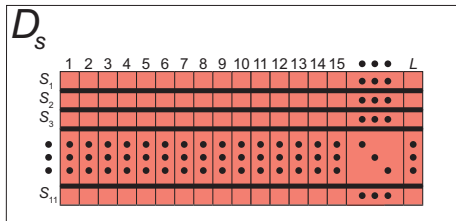
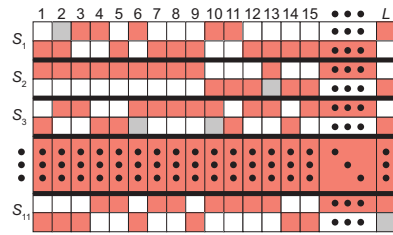
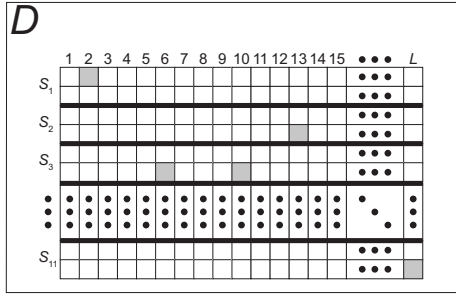
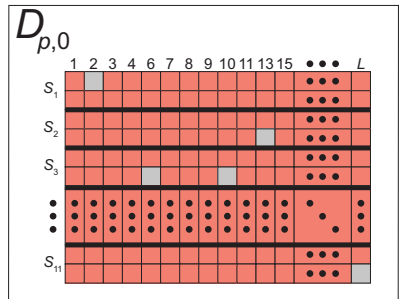
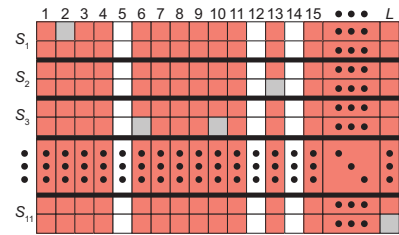
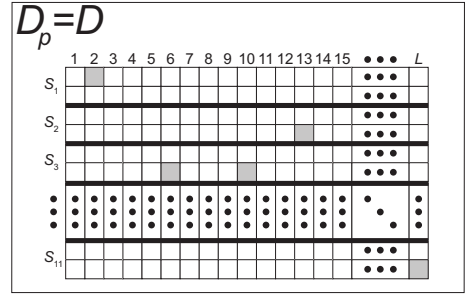
**A****B**



Figure 8.2: Schematic for creating the four subsets  $\mathcal{D}_s$ ,  $\mathcal{D}_{s,0}$ ,  $\mathcal{D}_p$ , and  $\mathcal{D}_{p,0}$  from dataset  $\mathcal{D}$  (see Table 8.2). For the matrices of datasets  $\mathcal{D}$ ,  $\mathcal{D}_s$ ,  $\mathcal{D}_{s,0}$ ,  $\mathcal{D}_p$ , and  $\mathcal{D}_{p,0}$ , each row is an individual and each column is a locus. Thick black lines in these matrices separate the individuals in different species. Gray boxes indicate a missing sequence. (A) At each locus, a single sequence from each species (indicated in red) is selected from dataset  $\mathcal{D}$ . These selected sequences are used to create  $\mathcal{D}_s$  such that there exists a single sequence sampled per species at each locus. Sequences from a subset of loci in  $\mathcal{D}_s$  (indicated in yellow) are used to create dataset  $\mathcal{D}_{s,0}$  such that each locus has at least one nucleotide difference between each distinct pair of species other than pairs from distinct outgroups. (B) Dataset  $\mathcal{D}_p$  is the full starting dataset  $\mathcal{D}$ . At each locus  $\ell$ , a distance matrix is created according to eq. 8.2. Sequences from a subset of loci (indicated in red) in  $\mathcal{D}_p$  are used to create dataset  $\mathcal{D}_{p,0}$  such that each locus has a nonzero  $p$ -distance between each distinct pair of species other than pairs from distinct outgroups. Observe that the  $\mathcal{D}_{p,0}$  matrix includes loci 3 and 7, which are not included in the  $\mathcal{D}_{s,0}$  matrix. The reason that loci 3 and 7 are included in dataset  $\mathcal{D}_{p,0}$  but not in dataset  $\mathcal{D}_{s,0}$  is that in  $\mathcal{D}_{p,0}$ , pairs of species contain at least one pair of individuals with different sequences, whereas in  $\mathcal{D}_{s,0}$ , at least one pair of the 11 selected individuals have identical sequences. Therefore, the set of loci in  $\mathcal{D}_{p,0}$  is a superset of the set of loci in  $\mathcal{D}_{s,0}$ , and the number of loci in dataset  $\mathcal{D}_{p,0}$  is always greater than or equal to the number of loci in dataset  $\mathcal{D}_{s,0}$ .

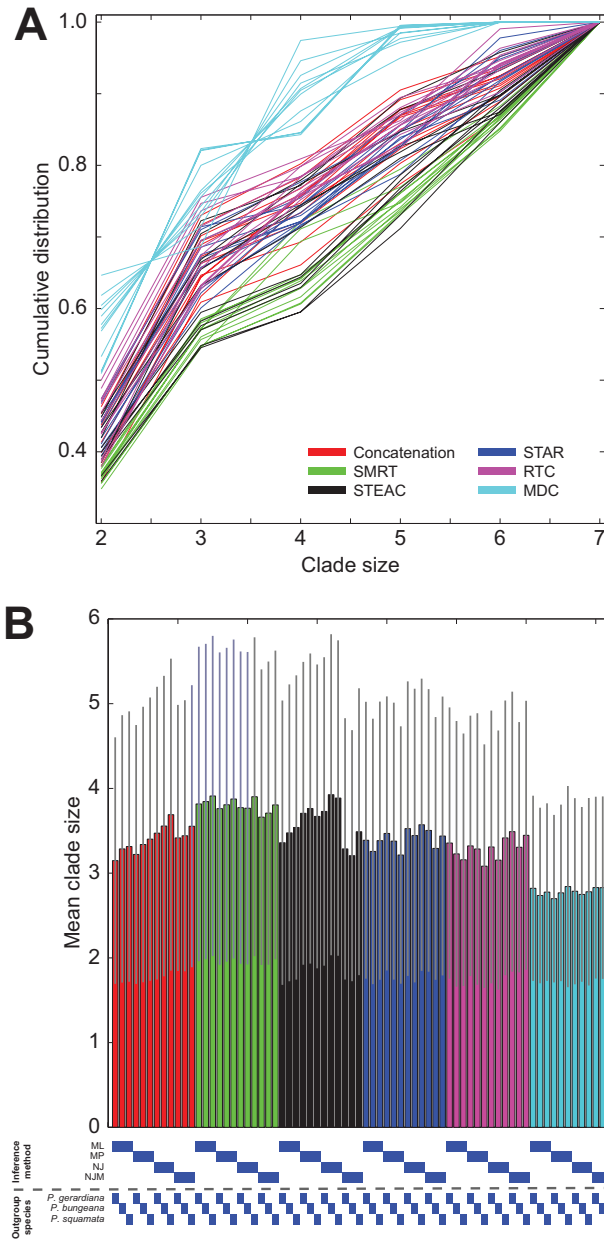


Figure 8.3: Distribution of clade size for all 72 phylogenetic inference strategies. (A) Cumulative distribution of clade sizes. Each line represents a strategy, of which there are 12 per color. (B) The mean clade size for a phylogenetic inference strategy was calculated as the mean size over all clades inferred across 1000 bootstrap replicates. Vertical lines centered at the top of each vertical bar represent the standard errors of mean clade sizes.

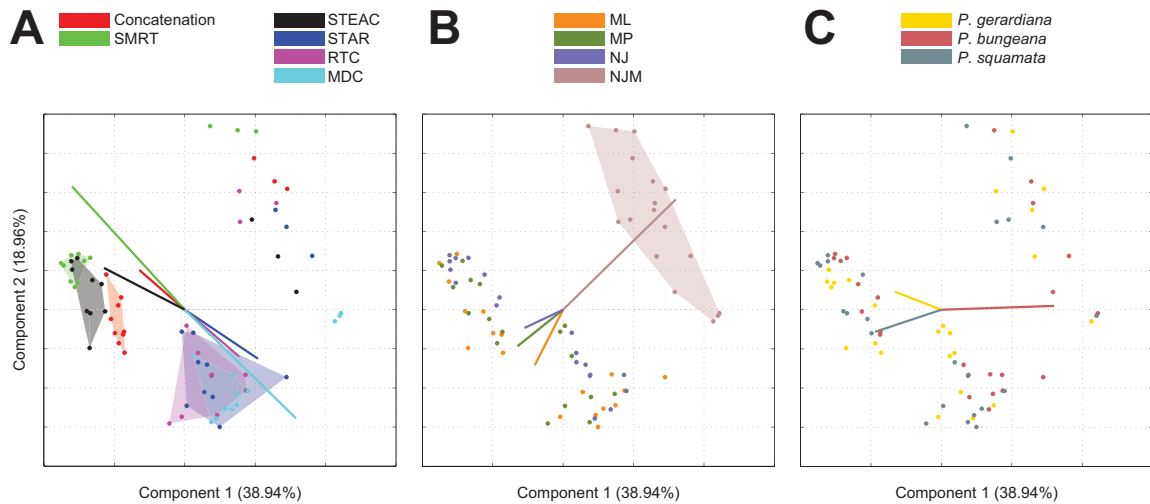


Figure 8.4: Principal components analysis of phylogenetic inference strategies. Principal components analysis was applied to 72 phylogenetic inference strategies, in which each strategy represents a point in a 145-dimensional space of clades. The plots show the first and second principal components. Each of the three plots represents the same 72 points in the space of principal components 1 and 2 with the exception that the three plots have points colored differently to highlight different features of the phylogenetic inference strategies. (A) Colors represent different methods for constructing species trees (Concatenation, SMRT, STEAC, STAR, RTC, and MDC). (B) Colors represent different gene tree inference methods (ML, MP, NJ, and NJM). (C) Colors represent different outgroups (*P. gerardiana*, *P. bungeana*, and *P. squamata*). The points on each graph represent different combinations of the three factors that form phylogenetic inference strategies. Each line in part A represents the resultant vector (scaled by a constant to lie within the span of the 72 points) for all 12 points of a certain method for constructing species trees. Each line in part B represents the resultant vector for all 18 points of a certain gene tree inference method (scaled by a constant). Each line in part C represents the resultant vector for all 24 points of a certain outgroup (scaled by a constant). Each of the constants used to scale resultant vectors in parts A, B, and C are distinct to parts A, B, and C, respectively. Each of the six shaded regions in part A and the shaded region in part B is a convex hull of the points from a particular species tree or gene tree inference method.

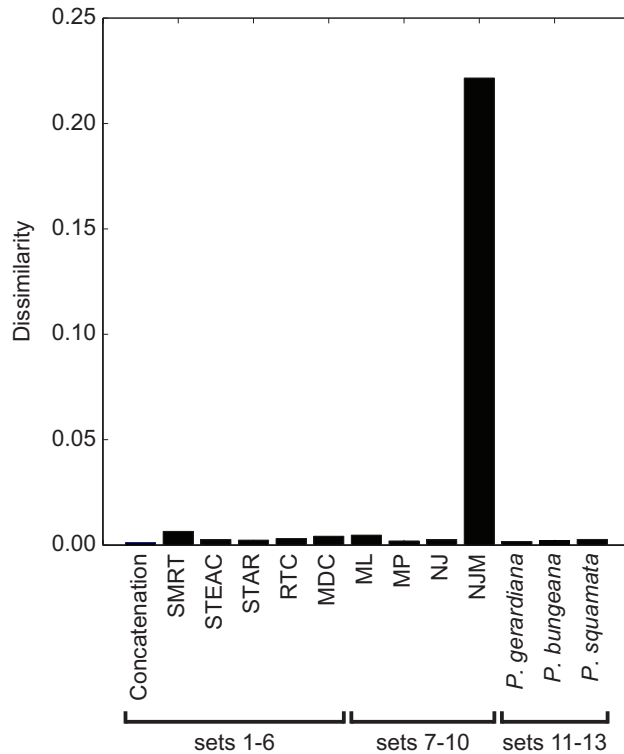


Figure 8.5: Procrustes analysis of the principal component plots in Figure 8.4. Principal components analysis of phylogenetic inference strategies was performed as in Figure 8.4 with the exception that the principal components analysis was applied to each of 13 subsets  $\mathcal{F}_1, \mathcal{F}_2, \dots, \mathcal{F}_{13}$  of phylogenetic inference strategies for which one feature of a strategy is removed.  $\mathcal{F}_1, \mathcal{F}_2, \dots, \mathcal{F}_6$  are subsets containing 60 strategies, in which one of the six species tree construction methods is not included,  $\mathcal{F}_7, \mathcal{F}_8, \dots, \mathcal{F}_{10}$  are subsets containing 54 strategies, in which one of the four gene tree inference methods is not included, and  $\mathcal{F}_{11}, \mathcal{F}_{12}, \mathcal{F}_{13}$  are subsets containing 48 strategies, in which one of the three outgroup species is not included. In other words, we performed principal components analysis on 13 different datasets: six in which a species tree construction method was removed (Concatenation, SMRT, STEAC, STAR, RTC, and MDC), four in which a gene tree inference method was removed (ML, MP, NJ, and NJM), and three in which an outgroup was removed (*P. gerardiana*, *P. bungeana*, and *P. squamata*). The results of each principal components analysis were projected into the two-dimensional plane spanned by their first and second principal components. The points in each two-dimensional plane were compared to the points in Figure 8.4A through Procrustes analysis. Each comparison gives a dissimilarity measure computed using eq. 8.3.

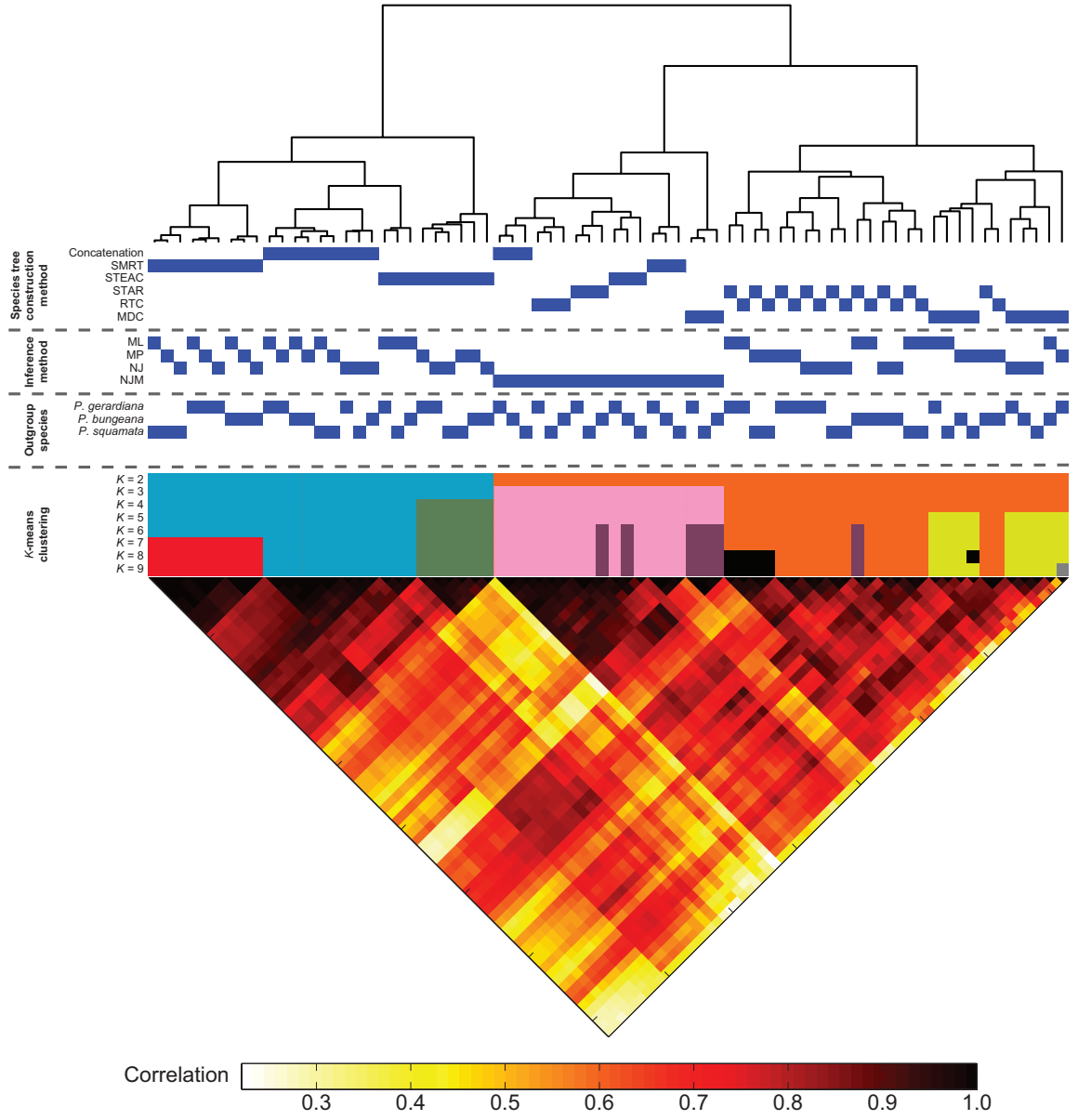


Figure 8.6: Cluster and correlation analysis of phylogenetic inference strategies. Each leaf of the dendrogram corresponds to a different phylogenetic inference strategy for obtaining the rooted phylogeny of eight ingroup pine species. Blue squares directly below the dendrogram indicate the features used to construct a rooted phylogeny for the eight pine species. The first six rows below the dendrogram represent different species tree construction methods. The next four rows below the dendrogram represent gene tree inference methods. The following three rows below the dendrogram represent the outgroup species. The dendrogram was constructed by hierarchical clustering using the Ward algorithm (*Ward*, 1963) applied to a matrix of Euclidean distances between all  $\binom{72}{2}$  pairs of 144-dimensional vectors (each dimension representing a distinct clade). The remaining nine rows below the outgroups show the results of  $K$ -means clustering applied to the 72 144-dimensional vectors with  $K$  clusters,  $K = 2, 3, \dots, 9$ . Below the cluster analysis is a heat map of the correlation coefficients between all  $\binom{72}{2}$  pairs of phylogenetic inference strategies. An entry in the heat map represents the Pearson correlation coefficient between a pair of strategies by only using points in the 144-dimensional vector that were nonzero in both strategies being compared.

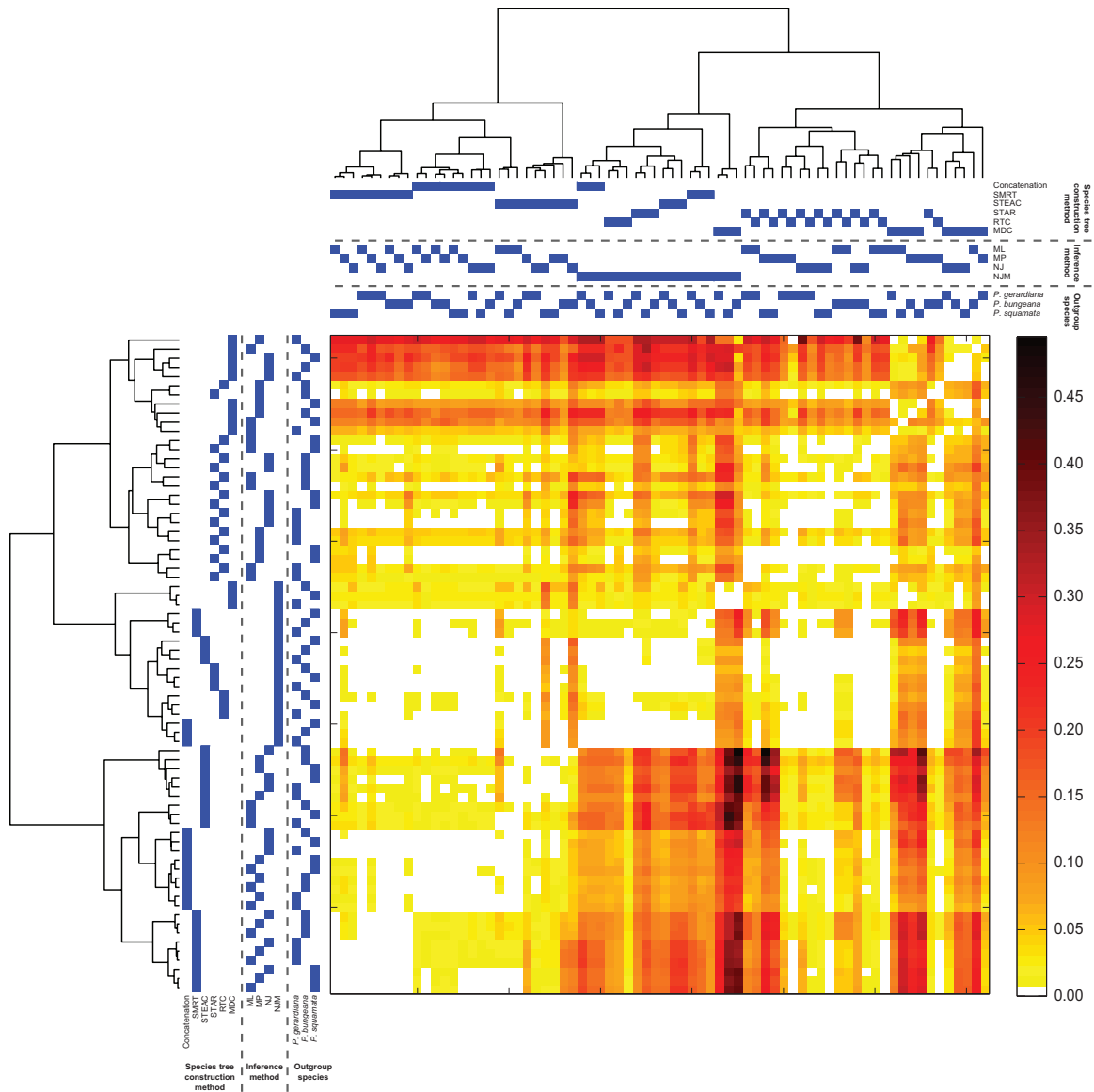


Figure 8.7: Heat map representing the “flow” of clades between phylogenetic inference strategies. We use clade flow to measure the proportion of clades inferred by one strategy that are not inferred by a different strategy (*i.e.*, the proportion of clades that do not “flow” from one strategy to another strategy). Phylogenetic inference strategies are ordered using the dendrogram. The cell at row  $i$  and column  $j$  represents the fraction of clades inferred by strategy  $i$  that were not inferred by strategy  $j$ . Darker colors indicate lower levels of “flow” from a row to a column.

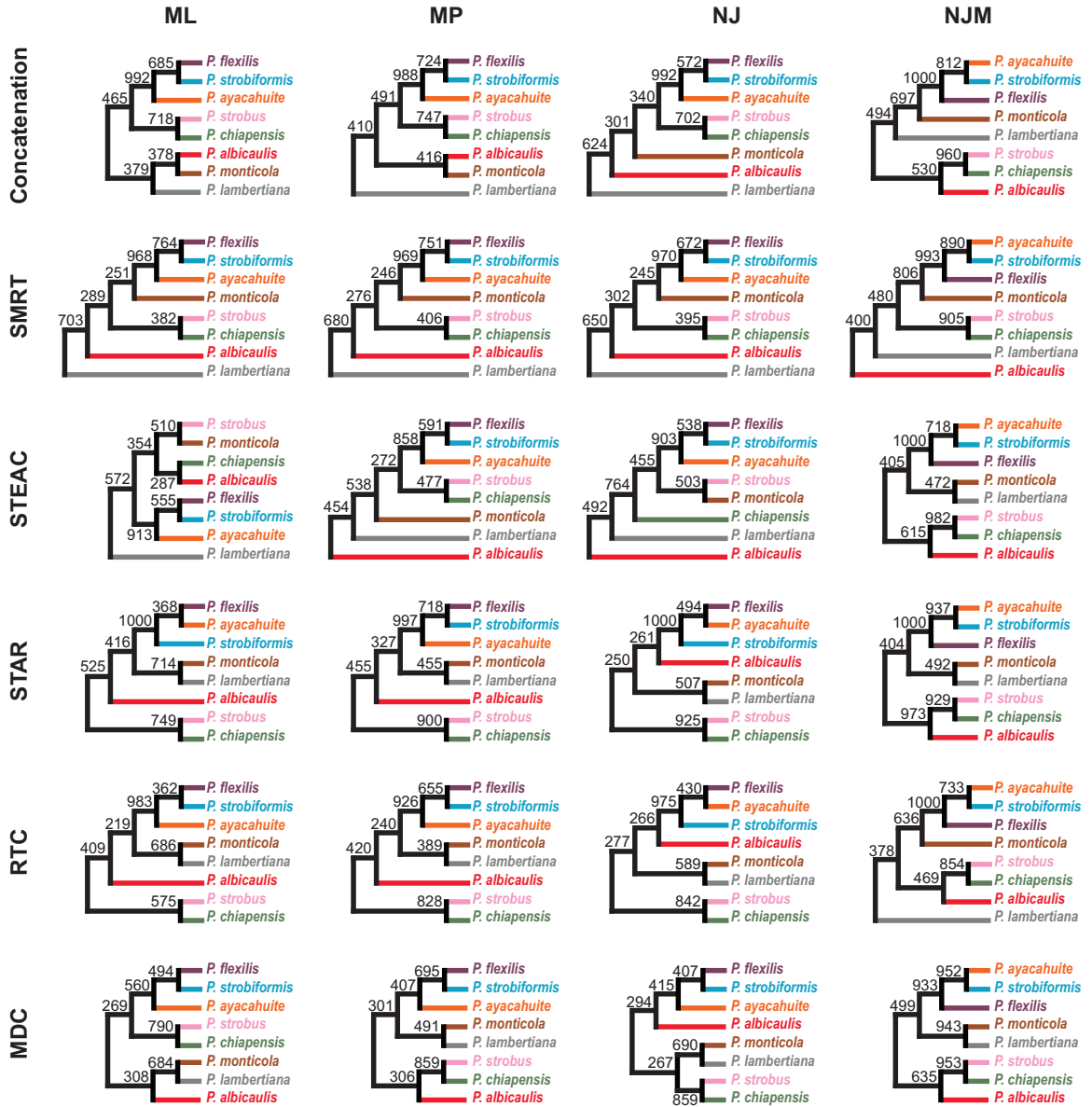


Figure 8.8: Consensus trees of phylogenetic inference strategies averaged over outgroups. For a given subset  $\mathcal{T}$  of the 72 phylogenetic inference strategies considered in this article, the bootstrap support for each of the clades that appeared in at least one tree was averaged over the set of strategies  $\mathcal{T}$  to create a set of counts  $\mathcal{C}$  for each of the clades. Greedy consensus trees (*Bryant, 2003*) were then created using the clade counts in the set  $\mathcal{C}$ . Each clade count in the set  $\mathcal{C}$  has a maximum value of 1000, because each element of  $\mathcal{C}$  is an average over values that each have a maximum value of 1000. Each consensus tree is the greedy consensus tree based on clade counts averaged over outgroup species. These trees disregard branch-length information.



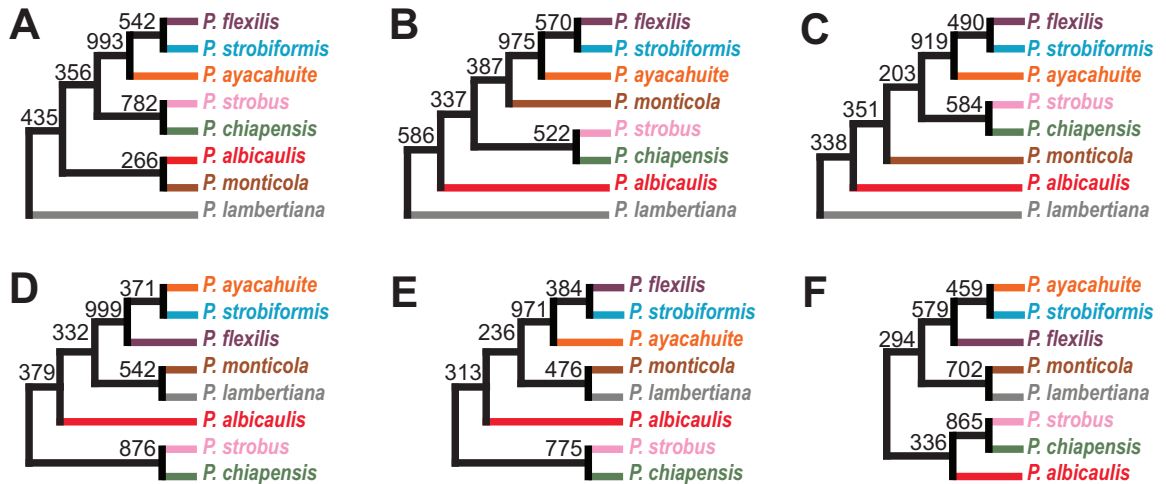


Figure 8.9: Consensus trees of phylogenetic inference strategies averaged over outgroups and gene tree inference methods. For a given subset  $\mathcal{T}$  of the 72 phylogenetic inference strategies considered in this article, the bootstrap support for each of the clades that appeared in at least one tree was averaged over the set of strategies  $\mathcal{T}$  to create a set of counts  $\mathcal{C}$  for each of the clades. Greedy consensus trees (Bryant, 2003) were then created using the clade counts in the set  $\mathcal{C}$ . Each clade count in the set  $\mathcal{C}$  has a maximum value of 1000, because each element of  $\mathcal{C}$  is an average over values that each have a maximum value of 1000. These trees disregard branch-length information. (A) Trees constructed using the 12 strategies that utilize Concatenation; (B) SMRT; (C) STEAC; (D) STAR; (E) RTC; (F) MDC.

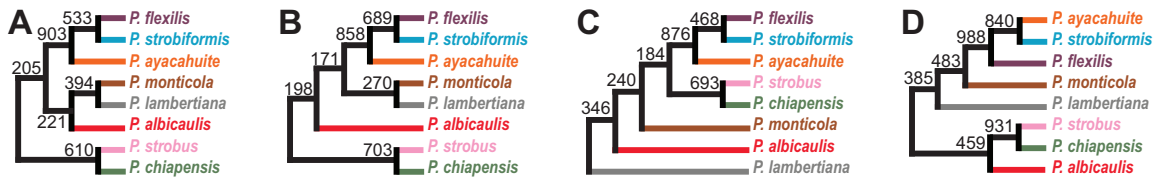


Figure 8.10: Consensus trees of phylogenetic inference strategies averaged over outgroups and species tree construction methods. For a given subset  $\mathcal{T}$  of the 72 phylogenetic inference strategies considered in this article, the bootstrap support for each of the clades that appeared in at least one tree was averaged over the set of strategies  $\mathcal{T}$  to create a set of counts  $\mathcal{C}$  for each of the clades. Greedy consensus trees (*Bryant, 2003*) were then created using the clade counts in the set  $\mathcal{C}$ . Each clade count in the set  $\mathcal{C}$  has a maximum value of 1000, because each element of  $\mathcal{C}$  is an average over values that each have a maximum value of 1000. These trees disregard branch-length information. (A) Trees constructed using the 18 strategies that utilize ML; (B) MP; (C) NJ; (D) NJM.

## CHAPTER IX

### Conclusion

In this dissertation, I have explored genetic variation at multiple levels, including variation within populations, among populations within a species, and among species. Using theory from population genetics, simulations, and genomic data, I have developed methods for measuring genetic variation, developed and analyzed methods that utilize genetic variation to infer population and species relationships, and used genetic variation to investigate the origins of anatomically modern humans. The inference tools developed and analyzed within this dissertation will enable investigators to more accurately assess genetic variation. The precise inferences of genetic variation can facilitate further understanding of evolutionary relationships both within and among populations and species. In addition, my results on modeling human origins provide compelling support for out-of-Africa hypotheses of human origins.

Chapters II and III involved the derivation of a set of unbiased estimators for a measure of genetic variation known as gene diversity (*Nei*, 1973). My results expanded upon previous work (*Nei and Roychoudhury*, 1974) by allowing for the inclusion of related individuals within samples. In addition, I derived an estimator that alleviates the bias generated by related individuals in the most general case of arbitrary ploidy, enabling investigators to apply my estimator to loci located on

haploid, diploid, and polyploid loci, and on sex chromosomes. Finally, I derived the exact variance formula for the general case of arbitrary ploidy. Given knowledge of a particular sample and locus under study, my theoretical, simulation, and empirical results suggest whether it is best (in terms of bias and mean squared error) to calculate gene diversity by correcting for related individuals, by removing related individuals from the dataset, or by neither correcting for nor removing related individuals. Given the generality of my derivations, my estimators are useful for characterizing genetic variation within a variety of organisms. In particular, to investigate patterns of worldwide human genetic variation, I applied the estimators to the human genetic datasets utilized in Chapters IV and V to correct the bias generated by related individuals within samples.

In Chapters IV and V, I investigated patterns in genetic variation predicted by various models of human evolutionary history and qualitatively compared the patterns to those observed from human genetic data. I found that a model representing the out-of-Africa hypothesis (*Relethford, 2008*), termed the serial founder model (*Ramachandran et al., 2005*), generated patterns of genetic variation that are consistent with those observed from human data. In addition, I found that variants of the serial founder model that incorporated small to moderate levels of gene flow between neighboring populations as well as small levels of archaic admixture are also consistent with the observed patterns. In contrast, I found that a model representing a version of the multiregional hypothesis (*Relethford, 2008*), termed the archaic persistence model, generated patterns of genetic variation that were opposite to those observed from human data. Additionally, by considering the instantaneous divergence model, I found that patterns of within-population genetic variation are driven primarily by a cumulative increase in genetic drift with increasing distance from a source population (*i.e.*, a population within Africa). By considering genetic variation between populations, I found that the serial founder model, in contrast to the

instantaneous divergence model, can predict observed patterns of between-population genetic variation through its incorporation of a hierarchical set of divergences. My results, therefore, provide strong support to the out-of-Africa hypothesis for the origins of anatomically modern humans.

Chapters VI, VII, and VIII took a different perspective on the topic of genetic variation by investigating phylogenetic tree reconstruction algorithms. Leveraging the coalescent process and the distribution of gene tree topologies under a species tree model (*Degnan and Salter, 2005*), I developed and evaluated methods for inferring species trees from multilocus data. In particular, due to a desirable property of the distribution of rooted three-taxon tree topologies under the coalescent (*Degnan and Rosenberg, 2006*), I developed a computationally efficient, accurate, and statistically consistent estimator of species tree topologies that performs well on genome-scale sequence datasets. My proof of statistical consistency for this method was particularly useful because it considered stochasticity due to both a coalescent and a mutation process, illustrating how our method would perform when applied to sequence data instead of known genealogies. In addition, I investigated the statistical consistency of several popular phylogenetic consensus methods (*Maddison, 1997; Degnan and Rosenberg, 2006; Ewing et al., 2008; Degnan et al., 2009; Liu et al., 2009, 2010; Mossel and Roch, 2010*) for inferring species trees from gene trees when non-random mating (or ancestral structure) exists within ancient species. I found a slightly discomfoting result for the performance of these consensus methods in that all of the methods except one, GLASS/Maximum Tree (*Liu et al., 2010; Mossel and Roch, 2010*), were statistically inconsistent under this structured population scenario; further, with simulations confirmed my theoretical predictions. In addition, when stochasticity due to the mutation process is considered, simulations showed that GLASS/Maximum Tree performs poorly even in the absence of ancestral structure, indicating that in practice, the methods may not perform well as the available data increases. However,

due to the observation of *Slatkin and Pollack* (2008) that ancestral population structure produces a certain asymmetry in the distribution of gene trees, it may be possible to leverage this asymmetry to develop methods for inferring species trees that account for ancestral population structure. Finally, I investigated the performance of phylogenetic inference strategies on a finite set of loci using an empirical dataset from North American pines. I found that phylogenetic inference strategies cluster into three distinct categories: those that utilize sequence information from all sampled individuals among a set of species, those that use only topological information to infer species trees, and those that do not strictly use topological information to infer species trees. In addition, I found that the Minimize Deep Coalescences method (*Maddison, 1997; Than and Nakhleh, 2009*) is biased toward balanced tree topologies. These three chapters will provide investigators with tools as well as guidance for accurately inferring species tree topologies from multilocus sequence data.

Through the understanding of evolutionary processes, we are capable of making inferences of within- and between-population evolutionary histories with the use of observed genetic variation. By leveraging the genetic variation encoded in our genomes and the genomes of other species, we can provide insight into our relationship with other species in the tree of life.

## BIBLIOGRAPHY

## BIBLIOGRAPHY

- Aho, A. V., Y. Sagiv, T. G. Szymanski, and J. D. Ullman (1981), Inferring a tree from lowest common ancestors with an application to the optimization of relational expressions, *SIAM J. Comput.*, *10*, 405–421.
- Almasy, L., and J. Blangero (1998), Multipoint quantitative-trait linkage analysis in general pedigrees, *Am. J. Hum. Genet.*, *62*, 1198–1211.
- Ané, C., B. Larget, D. A. Baum, S. D. Smith, and A. Rokas (2007), Bayesian estimation of concordance factors, *Mol. Biol. Evol.*, *24*, 412–426.
- Atteson, K. (1999), The performance of the neighbor-joining methods of phylogenetic reconstruction, *Algorithmica*, *25*, 251–278.
- Austerlitz, F., B. Jung-Muller, B. Godelle, and P.-H. Gouyon (1997), Evolution of coalescence times, genetic diversity and structure during colonization, *Theor. Popul. Biol.*, *51*, 148–164.
- Baum, B. R. (1992), Combining trees as a way of combining data sets for phylogenetic inference, and the desirability of combining gene trees, *Taxon*, *41*, 3–10.
- Belfiore, N. M., L. Liu, and C. Moritz (2008), Multilocus phylogenetics of a rapid radiation in the genus *Thomomys* (Rodentia: Geomyidae), *Syst. Biol.*, *57*, 294–310.
- Bininda-Emonds, O. R. P. (2004), The evolution of supertrees, *Trends Ecol. Evol.*, *19*, 315–322.
- Blum, M. G. B., and M. Jakobsson (2011), Deep divergences of human gene trees and models of human origins, *Mol. Biol. Evol.*, *28*, 889–898.
- Boehnke, M. (1991), Allele frequency estimation from data on relatives, *Am. J. Hum. Genet.*, *48*, 22–25.
- Bouillé, M., and J. Bousquet (2005), Trans-species shared polymorphisms at orthologous nuclear gene loci among distant species in the conifer *Picea* (Pinaceae): implications for the long-term maintenance of genetic diversity in trees, *Am. J. Bot.*, *92*, 63–73.
- Bourgain, C., S. Hoffjan, R. Nicolae, D. Newman, L. Steiner, K. Walker, R. Reynolds, C. Ober, and M. S. McPeck (2003), Novel case-control test in a founder population identifies P-selectin as an atopy-susceptibility locus, *Am. J. Hum. Genet.*, *73*, 612–626.



- Broman, K. W. (2001), Estimation of allele frequencies with data on sibships, *Genet. Epidemiol.*, *20*, 307–315.
- Brumfield, R. T., L. Liu, D. E. Lum, and S. V. Edwards (2008), Comparison of species tree methods for reconstructing the phylogeny of bearded manakins (Aves: Pipridae, *Manacus*) from multilocus sequence data, *Syst. Biol.*, *57*, 719–731.
- Bryant, D. (2003), A classification of consensus methods for phylogenies, in *BioConsensus*, edited by M. Janowitz, F. J. Lapointe, F. R. McMorris, B. Mirkin, and F. S. Roberts, pp. 163–183, DIMACS. AMS.
- Bryc, K., et al. (2010), Genome-wide patterns of population structure and admixture in West Africans and African Americans, *Proc. Natl. Acad. Sci. USA*, *107*, 786–791.
- Buckley, T. R., M. Cordeiro, D. C. Marshall, and C. Simon (2006), Differentiating between hypotheses of lineage sorting and introgression in New Zealand alpine cicadas (*Maoricicada dugdale*), *Syst. Biol.*, *55*, 411–425.
- Buerki, S., F. Forest, N. Salamin, and N. Alvarez (2011), Comparative performance of supertree algorithms in large data sets using the soapberry family (Sapindaceae) as a case study, *Syst. Biol.*, *60*, 32–44.
- Bulbeck, D. (2007), Where river meets sea: a parsimonious model for *Homo sapiens* colonization of the Indian Oceana rim and Sahul, *Curr. Anthropol.*, *48*, 315–321.
- Buteler, M. I., R. L. Jarret, and D. R. LaBonte (1999), Sequence characterization of microsatellites in diploid and polyploid *Ipomoea*, *Theor. Appl. Genet.*, *99*, 123–132.
- Cann, H. M., et al. (2002), A human genome diversity cell line panel, *Science*, *296*, 261–262.
- Carling, M. D., and R. T. Brumfield (2008), Integrating phylogenetic and population genetic analyses of multiple loci to test species divergence hypotheses in *Passerina* buntings, *Genetics*, *178*, 363–377.
- Carstens, B. C., and L. L. Knowles (2007), Estimating species phylogeny from gene-tree probabilities despite incomplete lineage sorting: an example from *Melanoplus* grasshoppers, *Syst. Biol.*, *56*, 400–411.
- Cavalli-Sforza, L. (2005), The Human Genome Diversity Project: past, present and future, *Nat. Rev. Genet.*, *6*, 333–340.
- Cavalli-Sforza, L. L. (2004), *Examining the Farming/Language Dispersal Hypothesis* (P. Bellwood and C. Renfrew, eds), McDonald Institute Monographs, Cambridge, UK.
- Cavalli-Sforza, L. L., and M. W. Feldman (2003), The application of molecular genetic approaches to the study of human evolution, *Nat. Genet.*, *33*, 266–275.

- Chen, D., L. Diao, O. Eulenstein, D. Fernández-Baca, and M. Sanderson (2003), *Flipping: a supertree construction method. Pages 135-160 in Bioconsensus (Vol. 61) (M. F. Janowitz et al., eds.)*, American Mathematical Society.
- Chor, B., and T. Tuller (2005), Maximum likelihood of evolutionary trees: hardness and approximation, *Bioinformatics*, *21*, i97–i106.
- Chor, B., M. Hendy, and S. Snir (2006), Maximum likelihood Jukes-Cantor triplets: analytic solutions, *Mol. Biol. Evol.*, *23*, 626–632.
- Chor, B., M. Hendy, and D. Penny (2007), Analytic solutions for three taxon ML trees with variable rates across sites, *Discrete Appl. Math.*, *155*, 750–758.
- Chung, K. L. (1974), *A course in probability theory*, 2nd ed., 365 pp., Academic Press, San Diego, CA.
- Clayton, D. G., et al. (2005), Population structure, differential bias and genomic control in a large-scale, case-control association study, *Nat. Genet.*, *37*, 1243–1246.
- Cockerham, C. C. (1971), Higher order probability functions of identity of alleles by descent, *Genetics*, *69*, 235–246.
- Colless, D. H. (1982), Phylogenetics, the theory and practice of phylogenetic systematics (book review), *Syst. Zool.*, *31*, 100–104.
- Cormen, T. H., C. E. Leiserson, R. L. Rivest, and C. Stein (2001), *Introduction to Algorithms*, 2nd ed., The MIT Press, Cambridge, Massachusetts.
- Cotterman, C. (1940), A calculus for statistico-genetics. Ph.D. thesis, Ohio State University, in *Genetics and social structure (Published 1974)*, edited by P. Ballonoff, Dowden, Hutchinson, and Ross, Stroudsburg, Pennsylvania.
- Cox, T. F., and M. A. A. Cox (2001), *Multidimensional scaling*, 2nd ed., Chapman and Hall, Boca Raton, Florida.
- Cranston, K. A., B. Hurwitz, D. Ware, L. Stein, and R. A. Wing (2009), Species trees from highly incongruent gene trees in rice, *Syst. Biol.*, *58*, 489–500.
- Currat, M., and L. Excoffier (2004), Modern humans did not admix with Neanderthals during their range expansion in Europe, *PLoS Biol.*, *2*, 2264–2274.
- Day, W., D. Johnson, and D. Sankoff (1986), The computational complexity of inferring rooted phylogenies by parsimony, *Math. Biosci.*, *81*, 33–42.
- de Queiroz, A., and J. Gatesy (2007), The supermatrix approach to systematics, *Trends Ecol. Evol.*, *22*, 34–31.
- DeGiorgio, M., and J. H. Degnan (2010), Fast and consistent estimation of species trees using supermatrix rooted triples, *Mol. Biol. Evol.*, *27*, 552–569.

- DeGiorgio, M., and N. A. Rosenberg (2009), An unbiased estimator of gene diversity in samples containing related individuals, *Mol. Biol. Evol.*, *26*, 501–512.
- DeGiorgio, M., M. Jakobsson, and N. A. Rosenberg (2009), Explaining worldwide patterns of human genetic variation using a coalescent-based serial founder model of migration outward from Africa, *Proc. Natl. Acad. Sci. USA*, *106*, 16,057–16,062.
- DeGiorgio, M., I. Jankovic, and N. A. Rosenberg (2010), Unbiased estimation of gene diversity in samples containing related individuals: exact variance and arbitrary ploidy, *Genetics*, *186*, 1367–1387.
- Degnan, J. H., and N. A. Rosenberg (2006), Discordance of species trees with their most likely gene trees, *PLoS Genet.*, *2*, e68.
- Degnan, J. H., and N. A. Rosenberg (2009), Gene tree discordance, phylogenetic inference, and the multispecies coalescent, *Trends Ecol. Evol.*, *24*, 332–340.
- Degnan, J. H., and L. A. Salter (2005), Gene tree distributions under the coalescent process, *Evolution*, *59*, 24–37.
- Degnan, J. H., M. DeGiorgio, D. Bryant, and N. A. Rosenberg (2009), Properties of consensus methods for inferring species trees from gene trees, *Syst. Biol.*, *58*, 35–54.
- Depaulis, F., and M. Veuille (1998), Neutrality tests based on the distribution of haplotypes under an infinite-site model, *Mol. Biol. Evol.*, *15*, 1788–1790.
- Derricourt, R. (2005), Getting “Out of Africa”: sea crossings, land crossings and culture in the Hominin migrations, *J. World Prehist.*, *19*, 119–132.
- Deshpande, O., S. Batzoglou, M. W. Feldman, and L. L. Cavalli-Sforza (2009), A serial founder effect model for human settlement out of Africa, *Proc. R. Soc. B*, *276*, 291–300.
- Driscoll, C. A., M. Menotti-Raymond, G. Nelson, D. Goldstein, and S. J. O’Brien (2002), Genomic microsatellites as evolutionary chronometers: A test in wild cats, *Genome Res.*, *12*, 414–423.
- Dryden, I. L., and K. V. Mardia (1998), *Statistical shape analysis*, Wiley, Chichester.
- Eckert, A. E., and B. C. Carstens (2008), Does gene flow destroy phylogenetic signal? The performance of three methods for estimating species phylogenies in the presence of gene flow, *Mol. Phyl. Evol.*, *49*, 832–842.
- Edgar, R. C. (2004a), MUSCLE: a multiple sequence alignment method with reduced time and space complexity, *BMC Bioinformatics*, *5*, 113.
- Edgar, R. C. (2004b), MUSCLE: multiple sequence alignment with high accuracy and high throughput, *Nucleic Acids Res.*, *32*, 1792–1797.

- Edmonds, C. A., A. S. Lillie, and L. L. Cavalli-Sforza (2004), Mutations arising in the wave front of an expanding population, *Proc. Natl. Acad. Sci. USA*, *101*, 975–979.
- Edwards, S. V. (2009), Is a new and general theory of systematics emerging?, *Evolution*, *63*, 1–19.
- Edwards, S. V., L. Liu, and D. K. Pearl (2007), High-resolution species trees without concatenation, *Proc. Natl. Acad. Sci. USA*, *104*, 5936–5941.
- Efromovich, S., and L. S. Kubatko (2008), Coalescent time distributions in trees of arbitrary size, *Stat. Appl. Genet. Mol.*, *7*, 2.
- Efron, B., and R. J. Tibshirani (1993), *An introduction to the bootstrap*, Chapman and Hall, NY.
- Eswaran, V. (2002), A diffusion wave out of Africa: The mechanism of the modern human revolution?, *Curr. Anthropol.*, *43*, 749–774.
- Evetts, I. W., and B. S. Weir (1998), *Interpreting DNA Evidence: Statistical Genetics for Forensic Scientists*, Sinauer Associates, Sunderland, Massachusetts, USA.
- Ewing, B., and P. Green (1998), Base-calling of automated sequencer traces using *phred*. II. Error probabilities, *Genome Res.*, *8*, 186–194.
- Ewing, B., L. Hiller, M. C. Wendl, and P. Green (1998), Base-calling of automated sequencer traces using *phred*. I. Accuracy assessment, *Genome Res.*, *8*, 175–185.
- Ewing, G. B., I. Ebersberger, H. A. Schmidt, and A. von Haeseler (2008), Rooted triple consensus and anomalous gene trees, *BMC Evol. Biol.*, *8*, doi:10.1186/1471-2148-8-118.
- Excoffier, L., and N. Ray (2008), Surfing during population expansions promotes genetic revolutions and structuration, *Trends Ecol. Evol.*, *23*, 347–351.
- Fagundes, N. J. R., N. Ray, M. Beaumont, S. Neuenschwander, F. M. Salzano, S. L. Bonatto, and L. Excoffier (2007), Statistical evaluation of alternative models of human evolution, *Proc. Natl. Acad. Sci. USA*, *104*, 17,614–17,619.
- Felsenstein, J. (1978), The number of evolutionary trees, *Syst. Zool.*, *27*, 27–33.
- Felsenstein, J. (2004), *Inferring phylogenies*, Sinauer Associates, Sunderland, MA.
- Field, J. S., M. D. Petraglia, and M. M. Lahr (2007), The southern dispersal hypothesis and the South Asian archaeological record: Examination of dispersal routes through GIS analysis, *J. Anthropol. Archaeol.*, *26*, 88–108.
- Friedlaender, J. S., et al. (2008), The genetic structure of Pacific Islanders, *PLoS Genet.*, *4*, 173–190.

- Gadagkar, S. R., M. Rosenberg, and S. Kumar (2005), Inferring species phylogenies from multiple genes: Concatenated sequence tree versus consensus gene tree, *J. Exp. Zool.*, *304B*, 64–74.
- Garrigan, D., and M. F. Hammer (2006), Reconstructing human origins in the genomic era, *Nat. Rev. Genet.*, *7*, 669–680.
- Garrigan, D., Z. Mobasher, S. B. Kingan, J. A. Wilder, and M. F. Hammer (2005), Deep haplotype divergence and long-range linkage disequilibrium at Xp21.1 provide evidence that humans descend from a structured ancestral population, *Genetics*, *170*, 1849–1856.
- Gatesy, J., and R. H. Baker (2005), Hidden likelihood support in genomic data: can forty-five wrongs make a right?, *Syst. Biol.*, *54*, 483–492.
- Gibbs, J. P., and W. T. Martin (1962), Urbanization, technology, and the division of labor: international patterns, *Am. Sociol. Rev.*, *27*, 667–677.
- Gillois, M. (1965), Relation d’identité en génétique, *Ann. Inst. H. Poincaré Sect. B*, *2*, 1–94.
- Gini, C. W. (1912), *Variabilita e mutabilita*, Studi Economico-Giuridici della R. Università di Cagliari 3.
- Goebel, T., M. R. Waters, and D. H. O’Rourke (2008), The late Pleistocene dispersal of modern humans in the Americas, *Science*, *319*, 1497–1502.
- Gower, J. C., and G. B. Dijkstrahuis (2004), *Procrustes problems*, Oxford University Press, New York.
- Gradshteyn, I. S., and I. M. Ryzhik (2007), *Table of Integrals, Series, and Products (A. Jeffrey and D. Zwillinger, eds)*, Academic Press, Burlington, MA, USA.
- Green, R. E., et al. (2006), Analysis of one million base pairs of Neanderthal DNA, *Nature*, *444*, 330–336.
- Green, R. E., et al. (2010), A draft sequence of the Neandertal genome, *Science*, *328*, 710–722.
- Hallatschek, O., and D. R. Nelson (2008), Gene surfing in expanding populations, *Theor. Popul. Biol.*, *73*, 158–170.
- Harris, D. L. (1964), Genotypic covariances between inbred relatives, *Genetics*, *50*, 1319–1348.
- Hedtke, S. M., T. M. Townsend, and D. M. Hillis (2006), Resolution of phylogenetic conflict in large data sets by increased taxon sampling, *Syst. Biol.*, *55*, 522–529.
- Hellenthal, G., A. Auton, and D. Falush (2008), Inferring human colonization history using a copying model, *PLoS Genet.*, *4*, e1000078.

- Hendy, M. D., and D. Penny (1989), A framework for the study of evolutionary trees, *Syst. Zool.*, *38*, 297–309.
- Herrera, K. J., J. A. Somarelli, R. K. Lowery, and R. J. Herrera (2009), To what extent did Neanderthals and modern human interact, *Biol. Rev.*, *84*, 245–257.
- Hird, S., L. Kubatko, and B. Carstens (2010), Rapid and accurate species tree estimation for phylogenetic investigations using replicated subsampling, *Mol. Phylogenet. Evol.*, *57*, 888–898.
- Hirschhorn, J. N., and M. J. Daly (2005), Genome-wide association studies for common diseases and complex traits, *Nat. Rev. Genet.*, *6*, 95–108.
- Hirschhorn, J. N., K. Lohmueller, E. Byrne, and K. Hirschhorn (2002), A comprehensive review of genetic association studies, *Genet. Med.*, *4*, 45–61.
- Hoelzel, A. R., R. C. Fleischer, C. Campagna, B. J. Le Boeuf, and G. Alvord (2002), Impact of a population bottleneck on symmetry and genetic diversity in the northern elephant seal, *J. Evol. Biol.*, *15*, 567–575.
- Holland, B. R., S. Benthin, P. J. Lockhart, V. Moulton, and K. T. Huber (2008), Using supernetworks to distinguish hybridization from lineage-sorting, *BMC Evol. Biol.*, *8*, 202.
- Hudson, R. R. (1983), Properties of a neutral allele model with intragenic recombination, *Theor. Popul. Biol.*, *23*, 183–201.
- Hudson, R. R. (1990), *Gene genealogies and the coalescent. Page 1-44 in Oxford Surveys in Evolutionary Biology (D. Futuyma and J. Antonovics, eds).*
- Hudson, R. R. (2002), Generating samples under a Wright-Fisher neutral model, *Bioinformatics*, *18*, 337–338.
- Hunley, K. L., M. E. Healy, and J. C. Long (2009), The global pattern of gene identity variation reveals a history of long-range migrations, bottlenecks, and local mate exchange: implications for biological race, *Am. J. Phys. Anthropol.*, *139*, 35–46.
- International HapMap 3 Consortium (2010), Integrating common and rare genetic variation in diverse human populations, *Nature*, *467*, 52–58.
- International HapMap Consortium (2005), A haplotype map of the human genome, *Nature*, *437*, 1299–1320.
- International HapMap Consortium (2007), A second generation human haplotype map of over 3.1 million SNPs, *Nature*, *449*, 851–861.
- Jacquard, A. (1974), *The Genetic Structure of Populations*, Springer, New York, USA.

- Jakobsson, M., et al. (2008), Genotype, haplotype, and copy-number variation in worldwide human populations, *Nature*, *451*, 998–1003.
- Jankovic, I., B. M. vonHoldt, and N. A. Rosenberg (2010), Heterozygosity of the Yellowstone wolves, *Mol. Ecol.*, *19*, 3246–3249.
- Jennings, W. B., and S. V. Edwards (2005), Speciation history of Australian grassfinches (*Poephila*) inferred from thirty gene trees, *Evolution*, *59*, 2033–2047.
- Jesus, F. F., J. F. Wilkins, V. N. Solferini, and J. Wakeley (2006), Expected coalescence times and segregating sites in a model of glacial cycles, *Genet. Mol. Res.*, *5*, 466–474.
- Jostins, L., K. I. Morley, and J. C. Barrett (2011), Imputation of low-frequency variants using the HapMap3 benefits from large, diverse reference sets, *Eur. J. Hum. Genet.*, *19*, 662–666.
- Kayser, M. (2010), The human genetic history of Oceania: near and remote views of dispersal, *Curr. Biol.*, *20*, R194–R201.
- Kingman, J. F. C. (1982), The coalescent, *Stochastic Process. Appl.*, *13*, 235–248.
- Kirkpatrick, M., and M. Slatkin (1993), Searching for evolutionary patterns in the shape of a phylogenetic tree, *Evolution*, *47*, 1171–1181.
- Klein, R. G. (2008), Out of Africa and the evolution of human behavior, *Evol. Anthropol.*, *17*, 267–281.
- Klopfstein, S., M. Currat, and L. Excoffier (2006), The fate of mutations surfing on the wave of a range expansion, *Mol. Biol. Evol.*, *23*, 482–490.
- Kolaczkowski, B., and J. W. Thornton (2004), Performance of maximum parsimony and likelihood phylogenetics when evolution is heterogeneous, *Nature*, *431*, 980–984.
- Kruglyak, L. (1999), Prospects for whole-genome linkage disequilibrium mapping of common disease genes, *Nat. Genet.*, *22*, 139–144.
- Kubatko, L., H. L. Gibbs, and E. W. Bloomquist (2009), Inferring species-level phylogenies using multi-locus data for a recent radiation of sistrus rattlesnakes, *Syst. Biol.*
- Kubatko, L. S., and J. H. Degnan (2007), Inconsistency of phylogenetic estimates from concatenated data under coalescence, *Syst. Biol.*, *56*, 17–24.
- Lange, K. (2002), *Mathematical and Statistical Methods for Genetic Analysis*, 2nd ed., Springer, New York, USA.
- Lao, O., et al. (2008), Correlation between genetic and geographic structure in Europe, *Curr. Biol.*, *18*, 1241–1248.

- Le Corre, V., and A. Kremer (1998), Cumulative effects of founding events during colonisation on genetic diversity and differentiation in an island and stepping-stone model, *J. Evol. Biol.*, *11*, 495–512.
- Lee, W. H., and V. B. Vega (2004), Heterogeneity detector: finding heterogeneous positions in Phred/Phrap assemblies, *Bioinformatics*, *20*, 2863–2864.
- Li, C. C., and D. G. Horvitz (1953), Some methods of estimating the inbreeding coefficient, *Am. J. Hum. Genet.*, *5*, 107–117.
- Li, J. Z., et al. (2008), Worldwide human relationships inferred from genome-wide patterns of variation, *Science*, *319*, 1100–1104.
- Linnen, C. R., and B. D. Farrell (2008), Comparison of methods for species-tree inference in the sawfly genus *Neodiprion* (Hymenoptera: Diprionidae), *Syst. Biol.*, *57*, 876–890.
- Liu, H., F. Prugnolle, A. Manica, and F. Balloux (2006), A geographically explicit genetic model of worldwide human-settlement history, *Am. J. Hum. Genet.*, *79*, 230–237.
- Liu, L. (2008), BEST: Bayesian estimation of species trees under the coalescent model, *Bioinformatics*, *24*, 2542–2543.
- Liu, L., and S. V. Edwards (2009), Phylogenetic analysis in the anomaly zone, *Syst. Biol.*, *58*, 452–460.
- Liu, L., and D. K. Pearl (2007), Species trees from gene trees: reconstructing Bayesian posterior distributions of a species phylogeny using estimated gene tree distributions, *Syst. Biol.*, *56*, 504–514.
- Liu, L., D. K. Pearl, R. Brumfield, and S. V. Edwards (2008), Estimating species trees using multiple-allele DNA sequence data, *Evolution*, *62*, 2080–2091.
- Liu, L., L. Yu, D. K. Pearl, and S. V. Edwards (2009), Estimating species phylogenies using coalescence times among sequences, *Syst. Biol.*, *58*, 1–10.
- Liu, L., L. Yu, and D. K. Pearl (2010), Maximum tree: a consistent estimator of the species tree, *J. Math. Biol.*, *60*, 95–106.
- Ma, X.-F., A. E. Szmidt, and X.-R. Wang (2006), Genetic structure and evolutionary history of a diploid hybrid pine *Pinus densata* inferred from the nucleotide variation at seven gene loci, *Mol. Biol. Evol.*, *23*, 807–816.
- Maddison, W. P. (1997), Gene trees in species trees, *Syst. Biol.*, *46*, 523–536.
- Maddison, W. P., and L. L. Knowles (2006), Inferring phylogeny despite incomplete lineage sorting, *Syst. Biol.*, *55*, 21–30.



- Marchini, J., L. R. Cardon, M. S. Phillips, and P. Donnelly (2004), The effects of human population structure on large genetic association studies, *Nat. Genet.*, *36*, 512–517.
- Marth, G. T., E. Czabarka, J. Murvai, and S. T. Sherry (2004), The allele frequency spectrum in genome-wide human variation data reveals signals of differential demographic history in three large world populations, *Genetics*, *166*, 351–372.
- McCarthy, M. I., G. R. Abecasis, L. R. Cardon, D. B. Goldstein, J. Little, J. P. A. Ioannidis, and J. N. Hirschhorn (2008), Genome-wide association studies for complex traits: consensus, uncertainty and challenges, *Curr. Biol.*, *9*, 356–369.
- McPeck, M. S., X. Wu, and C. Ober (2004), Best linear unbiased allele-frequency estimation in complex pedigrees, *Biometrics*, *60*, 359–367.
- McVean, G. (2010), A genealogical interpretation of principal components analysis, *PLoS Genet.*, *5*, e1000686.
- Meltzer, D. J. (2009), *First Peoples in a New World: Colonizing Ice Age America*, University of California Press, Berkeley, CA, USA.
- Meng, C., and L. S. Kubatko (2009), Detecting hybrid speciation in the presence of incomplete lineage sorting using gene tree incongruence: a model, *Theor. Pop. Biol.*, *75*, 35–45.
- Mossel, E., and S. Roch (2010), Incomplete lineage sorting: consistent phylogeny estimation from multiple loci, *IEEE/ACM Trans. Comput. Biol. Bioinform.*, *7*, 166–171.
- Mossel, E., and E. Vigoda (2005), Phylogenetic MCMC algorithms are misleading on mixtures of trees, *Science*, *309*, 2207–2209.
- Nei, M. (1973), Analysis of gene diversity in subdivided populations, *Proc. Natl. Acad. Sci. USA*, *70*, 3321–3323.
- Nei, M. (1987), *Molecular Evolutionary Genetics*, Columbia University Press, New York.
- Nei, M., and A. K. Roychoudhury (1974), Sampling variances of heterozygosity and genetic distance, *Genetics*, *76*, 379–390.
- Neyman, J. (1971), Molecular studies in evolution: a source of novel statistical problems, in *Statistical Decision Theory and Related Topics*, edited by S. S. Gupta and J. Yackel, pp. 1–27, Academic Press, NY.
- Nielsen, R. (2010), In search of rare human variants, *Nature*, *467*, 1050–1051.
- Noonan, J. P., et al. (2006), Sequencing and analysis of Neanderthal genomic DNA, *Science*, *314*, 1113–1118.

- Novembre, J., et al. (2008), Genes mirror geography in Europe, *Nature*, *456*, 98–101.
- Ohta, T. (1980), Linkage disequilibrium between amino acid sites in immunoglobulin genes and other multigene families, *Genet. Res.*, *36*, 181–197.
- Page, R. D. M. (2002), Modified mincut supertrees. pages 537-552 in lecture notes in computer science, in *Algorithms in Bioinformatics, Second International Workshop, WABI, 2002, Rome, Italy, September 17-21, 2002, Proceedings (Vol. 2452)*, edited by R. Guigó and D. Gusfield, pp. 537–552, Springer, Berlin.
- Page, R. D. M., and M. A. Charleston (1997), From gene to organismal phylogeny: Reconciled trees and the gene tree/species tree problem, *Mol. Phylogenet. Evol.*, *7*, 231–240.
- Pamilo, P., and M. Nei (1988), Relationships between gene trees and species trees, *Mol. Biol. Evol.*, *5*, 568–583.
- Parks, M., R. Cronn, and A. Liston (2009), Increasing phylogenetic resolution at low taxonomic levels using massively parallel sequencing of chloroplast genomes, *BMC Biology*, *7*, 84.
- Pemberton, T. J., C. Wang, J. Z. Li, and N. A. Rosenberg (2010), Inference of unexpected genetic relatedness among individuals in HapMap Phase III, *Am. J. Hum. Genet.*, *87*, 457–464.
- Plagnol, V., and J. D. Wall (2006), Possible ancestral structure in human populations, *PLoS Genet.*, *2*, 972–979.
- Poe, S., and D. L. Swofford (1999), Taxon sampling revisited, *Nature*, *398*, 299–300.
- Price, A. L., A. Helgason, S. Palsson, H. Stefansson, D. St. Clair, O. A. Andreassen, D. Reich, A. Kong, and K. Stefansson (2009), The impact of divergence time on the nature of population structure: an example from Iceland, *PLoS Genet.*, *5*, e1000505.
- Prugnolle, F., A. Manica, and F. Balloux (2005), Geography predicts neutral genetic diversity of human populations, *Curr. Biol.*, *15*, R159–R160.
- R Development Core Team (2008), *R: A Language and Environment for Statistical Computing*, R Foundation for Statistical Computing, Vienna, Austria, ISBN 3-900051-07-0.
- Ragan, M. A. (1992), Phylogenetic inference based on matrix representation of trees, *Mol. Phylogenet. Evol.*, *1*, 53–58.
- Ramachandran, S., O. Deshpande, C. C. Roseman, N. A. Rosenberg, M. W. Feldman, and L. L. Cavalli-Sforza (2005), Support from the relationship of genetic and geographic distance in human populations for a serial founder effect originating in Africa, *Proc. Natl. Acad. Sci. USA*, *102*, 15,942–15,947.

- Ramachandran, S., N. A. Rosenberg, M. W. Feldman, and J. Wakeley (2008), Population differentiation and migration: coalescence times in a two-sex island model for autosomal and X-linked loci, *Theor. Pop. Biol.*, *74*, 291–301.
- Rambaut, A., and N. C. Grassly (1997), Seq-Gen: An application for the Monte Carlo simulation of DNA sequence evolution along phylogenetic trees, *Comput. Appl. Biosc.*, *13*, 235–238.
- Rannala, B., and Z. Yang (2003), Bayes estimation of species divergence times and ancestral population sizes using DNA sequences from multiple loci, *Genetics*, *164*, 1645–1656.
- Rannala, B., and Z. Yang (2008), Phylogenetic inference using whole genomes, *Annu. Rev. Genomics Hum. Genet.*, *9*, 217–231.
- Ray, N., M. Currat, P. Berthier, and L. Excoffier (2005), Recovering the geographic origins of early modern humans by realistic and spatially explicit simulations, *Genome Res.*, *15*, 1161–1167.
- Reich, D., et al. (2010), Genetic history of an archaic hominin group from Denisova Cave in Siberia, *Nature*, *468*, 1053–1060.
- Reiland, J., S. Hodge, and M. A. F. Noor (2002), Strong founder effect in *Drosophila pseudoobscura* colonizing New Zealand from North America, *J. Hered.*, *93*, 415–420.
- Relethford, J. H. (1998), Genetics of modern human origins and diversity, *Annu. Rev. Anthropol.*, *27*, 1–23.
- Relethford, J. H. (2008), Genetic evidence and the modern human origins debate, *Heredity*, *100*, 555–563.
- Ritland, K. (1996), Estimators for pairwise relatedness and individual inbreeding coefficients, *Genet. Res. Camb.*, *67*, 175–185.
- Roch, S. (2006), A short proof that phylogenetic tree reconstruction by maximum likelihood is hard, *IEEE ACM T. Comput. BI.*, *3*, 92–94.
- Rokas, A., and S. B. Carroll (2005), More genes or more taxa? The relative contribution of gene number and taxon number to phylogenetic accuracy, *Mol. Biol. Evol.*, *22*, 1337–1344.
- Rokas, A., B. L. Williams, N. King, and S. B. Carroll (2003), Genome-scale approaches to resolving incongruence in molecular phylogenies, *Nature*, *425*, 798–804.
- Rosenberg, M. S., and S. Kumar (2001), Incomplete taxon sampling is not a problem for phylogenetic inference, *Proc. Natl. Acad. Sci. USA*, *98*, 10,751–10,756.

- Rosenberg, N. A. (2006), Standardized subsets of the HGDP-CEPH Human Genome Diversity Cell Line Panel, accounting for atypical and duplicated samples and pairs of close relatives, *Ann. Hum. Genet.*, *70*, 841–847.
- Rosenberg, N. A. (2007), Counting coalescent histories, *J. Comp. Biol.*, *14*, 360–377.
- Rosenberg, N. A., and R. Tao (2008), Discordance of species trees with their most likely gene trees: the case of five taxa, *Syst. Biol.*, *57*, 131–140.
- Rosenberg, N. A., J. K. Pritchard, J. L. Weber, H. M. Cann, K. K. Kidd, L. A. Zhivotovsky, and M. W. Feldman (2002), Genetic structure of human populations, *Science*, *298*, 2381–2385.
- Rosenberg, N. A., S. Mahajan, S. Ramachandran, C. Zhao, J. K. Pritchard, and M. W. Feldman (2005), Clines, clustes, and the effect of study design on the inference of human population structure, *PLoS Genet.*, *1*, 660–671.
- Rosenberg, N. A., L. Huang, E. M. Jewett, Z. A. Szpiech, I. Jankovic, and M. Boehnke (2010), Genome-wide association studies in diverse populations, *Nat. Rev. Genet.*, *11*, 356–366.
- Rousset, F. (2002), Inbreeding and relatedness coefficients: what do they measure?, *Heredity*, *88*, 371–380.
- Sabatti, C., and N. Risch (2002), Homozygosity and linkage disequilibrium, *Genetics*, *160*, 1707–1719.
- Sabeti, P. C., et al. (2002), Detecting recent positive selection in the human genome from haplotype structure, *Nature*, *419*, 832–837.
- Sackin, M. J. (1972), “Good” and “bad” phenograms, *Syst. Zool.*, *21*, 225–226.
- Savolainen, O., and T. Pyhäjärvi (2007), Genomic diversity in forest trees, *Curr. Opin. Plant Biol.*, *10*, 162–167.
- Schaffner, S. F., C. Foo, S. Gabriel, D. Reich, M. J. Daly, and D. Alschuler (2005), Calibrating a coalescent simulation of human genome sequence variation, *Genome Res.*, *15*, 1576–1583.
- Semple, C., and M. Steel (2000), A supertree method for rooted trees, *Discrete Appl. Math.*, *105*, 147–158.
- Serfling, R. J. (1980), *Approximation Theorems of Mathematical Statistics*, 3rd ed., McGraw-Hill, New York.
- Serre, D., A. Langaney, M. Chech, M. Teschler-Nicola, M. Paunovic, P. Menecier, M. Hofreiter, G. Possnert, and S. Paabo (2004), No evidence of Neandertal mtDNA contribution to early modern humans, *PLoS Biol.*, *2*, 313–317.
- Shao, K.-T., and R. R. Sokal (1990), Tree balance, *Syst. Zool.*, *39*, 266–276.

- Shete, S. (2003), Uniformly minimum variance unbiased estimation of gene diversity, *J. Hered.*, *94*, 421–424.
- Simpson, E. H. (1949), Measurement of diversity, *Nature*, *163*, 688–688.
- Sjodin, P., I. Kaj, S. Krone, M. Lascoux, and M. Nordborg (2005), On the meaning and existence of an effective population size, *Genetics*, *169*, 1061–1070.
- Slatkin, M. (1991), Inbreeding coefficients and coalescence times, *Genet. Res.*, *58*, 167–175.
- Slatkin, M. (2008), Linkage disequilibrium – understanding the evolutionary past and mapping the medical future, *Nat. Rev. Genet.*, *9*, 477–485.
- Slatkin, M., and J. L. Pollack (2008), Subdivision in an ancestral species creates asymmetry in gene trees, *Mol. Biol. Evol.*, *25*, 2241–2246.
- Steel, M. (1992), The complexity of reconstructing trees from qualitative characters and subtrees, *J. Classification*, *9*, 91–116.
- Steel, M., and A. Rodrigo (2008), Maximum likelihood supertrees, *Syst. Biol.*, *57*, 243–250.
- Strimmer, K., and A. von Haeseler (1996), Quartet puzzling: a quartet maximum-likelihood method for reconstructing tree topologies, *Mol. Biol. Evol.*, *13*, 964–969.
- Stringer, C. (2002), Modern human origins: progress and prospects, *Phil. Trans. R. Soc. Lond. B*, *357*, 563–579.
- Swofford, D. L. (2003), PAUP\*. Phylogenetic analysis using parsimony (\* and other methods). Version 4. Sinauer Associates, Sunderland, Massachusetts.
- Syring, J., A. Willyard, R. Cronn, and A. Liston (2005), Evolutionary relationships among *Pinus* (Pinaceae) subsections inferred from multiple low-copy nuclear loci, *Am. J. Bot.*, *92*, 2086–2100.
- Syring, J., K. Farrell, R. Businsky, R. Cronn, and A. Liston (2007), Widespread genealogical nonmonophyly in species of *Pinus* subgenus *Strobus*, *Syst. Biol.*, *56*, 163–181.
- Szpiech, Z. A., M. Jakobsson, and N. A. Rosenberg (2008), ADZE: a rarefaction approach for counting alleles private to combinations of populations, *Bioinformatics*, *24*, 2498–2504.
- Tajima, F. (1983), Evolutionary relationship of dna sequences in finite populations, *Genetics*, *105*, 437–460.
- Takahata, N. (1993), Allelic genealogy and human evolution, *Mol. Biol. Evol.*, *10*, 2–22.

- Takahata, N., Y. Satta, and J. Klein (1995), Divergence time and population size in the lineage leading to modern humans, *Theor. Popul. Biol.*, *48*, 198–221.
- Takahata, N., S.-H. Lee, and Y. Satta (2001), Testing multiregionality of modern human origins, *Mol. Biol. Evol.*, *18*, 172–183.
- Takezaki, N., and M. Nei (2008), Empirical tests of the reliability of phylogenetic trees constructed with microsatellite DNA, *Genetics*, *178*, 385–392.
- Tang, H., D. O. Siegmund, P. Shen, P. J. Oefner, and M. W. Feldman (2002), Frequentist estimation of coalescence times from nucleotide sequence data using a tree-based partition, *Genetics*, *161*, 447–459.
- Tavaré, S. (1984), Line-of-descent and genealogical processes, and their applications in population genetics models, *Theor. Popul. Biol.*, *26*, 119–164.
- Thalmann, O., A. Fischer, F. Lankester, S. Pääbo, and L. Vigilant (2007), The complex evolutionary history of gorillas: insights from genomic data, *Mol. Biol. Evol.*, *24*, 146–158.
- Than, C., and L. Nakhleh (2009), Species tree inference by minimizing deep coalescences, *PLoS Comput. Biol.*, *5*, e1000501.
- Than, C., D. Ruths, H. Innan, and L. Nakhleh (2007), Confounding factors in HGT detection: Statistical error, coalescent effects, and multiple solutions, *J. Comput. Biol.*, *14*, 517–535.
- Than, C. V., and N. A. Rosenberg (2011), Consistency properties of species tree inference by minimizing deep coalescences, *J. Comput. Biol.*, *18*, 1–5.
- The 1000 Genomes Project Consortium (2010), A map of human genome variation from population-scale sequencing, *Nature*, *467*, 1061–1073.
- Themudo, G. E., B. Wiestra, and J. W. Arntzen (2009), Multiple nuclear and mitochondria genes resolve the branching order of a rapid radiation of crested newts (*Triturus*, Salamandridae), *Mol. Phylogenet. Evol.*, *52*, 321–328.
- Thompson, E. A. (1974), Gene identities and multiple relationships, *Biometrics*, *30*, 667–680.
- Thomson, R., J. Pritchard, P. Shen, P. Oefner, and M. Feldman (2000), Recent common ancestry of human Y chromosomes: evidence from DNA sequence data, *Proc. Natl. Acad. Sci. USA*, *97*, 7360–7365.
- Tishkoff, S. A., and K. K. Kidd (2004), Implications of biogeography of human populations for ‘race’ and medicine, *Nat. Genet.*, *11*, S21–S27.
- Tishkoff, S. A., and B. C. Verrelli (2003), Patterns of human genetic diversity: implications for human evolutionary history and disease, *Annu. Rev. Genomics Hum. Genet.*, *4*, 293–340.

- Tishkoff, S. A., et al. (2009), The genetic structure and history of Africans and African Americans, *Science*, *324*, 1035–1044.
- Wakeley, J. (1996a), The variance of pairwise nucleotide differences in two populations with migration, *Theor. Popul. Biol.*, *49*, 39–57.
- Wakeley, J. (1996b), Distinguishing migration from isolation using the variance of pairwise differences, *Theor. Popul. Biol.*, *49*, 369–386.
- Wakeley, J. (1996c), Pairwise differences under a general model of population subdivision, *J. Genet.*, *75*, 81–89.
- Wakeley, J. (2009), *Coalescent Theory: An Introduction*, Roberts and Company Publishers, Greenwood Village, Colorado.
- Wall, J. D., K. E. Lohmueller, and V. Plagnol (2009), Detecting ancient admixture and estimating demographic parameters in multiple human populations, *Mol. Biol. Evol.*, *26*, 1823–1827.
- Wang, C., Z. A. Szpiech, J. H. Degnan, M. Jakobsson, T. J. Pemberton, J. A. Hardy, A. B. Singleton, and N. A. Rosenberg (2010), Comparing spatial maps of human population-genetic variation using Procrustes analysis, *Stat. Appl. Genet. Mol. Biol.*, *9*, 13.
- Wang, S., et al. (2007), Genetic variation and population structure in Native Americans, *PLoS Genet.*, *3*, 2049–2067.
- Ward, J. H. (1963), Hierarchical grouping to optimize an objective function, *J. Am. Stat. Assoc.*, *58*, 236–244.
- Watterson, G. A. (1978), The homozygosity test of neutrality, *Genetics*, *88*, 405–417.
- Weaver, T. D., and C. C. Roseman (2008), New developments in the genetic evidence for modern human origins, *Evol. Anthropol.*, *17*, 69–80.
- Wegrzyn, J. L., J. M. Lee, J. Liechty, and D. B. Neale (2009), PineSAP—sequence alignment and SNP identification pipeline, *Bioinformatics*, *25*, 2609–2610.
- Weir, B. S. (1989), *Sampling properties of gene diversity*. In: *Plant Population Genetics, Breeding and Genetic Resources (Brown AHD, Clegg MT, Kahler AL, and Weir BS, eds)*, Sinauer Associates, Sunderland, Massachusetts, USA.
- Weir, B. S. (1996), *Genetic Data Analysis II*, Sinauer Associates, Sunderland, Massachusetts, USA.
- Weir, B. S., A. D. Anderson, and A. B. Hepler (2006), Genetic relatedness analysis: modern data and new challenges, *Nat. Rev. Genet.*, *7*, 771–780.

- White, M. A., C. Ané, C. N. Dewey, B. R. Larget, and B. A. Payseur (2009), Fine-scale phylogenetic discordance across the house mouse genome, *PLoS Genet.*, *5*, e1000729.
- Wilkins, J. F., and J. Wakeley (2002), The coalescent in a continuous, finite, linear population, *Genetics*, *161*, 873–888.
- Willson, S. J. (2009), Robustness of topological supertree methods for reconciling dense incompatible data, *IEEE-ACM Trans. Comput. Biol. Bioinf.*, *6*, 62–75.
- Willyard, A., R. Cronn, and A. Liston (2009), Reticulate evolution and incomplete lineage sorting among the ponderosa pines, *Mol. Phyl. Evol.*, *52*, 498–511.
- Wolpoff, M. H., J. Hawks, and R. Caspari (2000), Multiregional, not multiple origins, *Am. J. Phys. Anthropol.*, *112*, 129–136.
- Xing, J., et al. (2010), Toward a more uniform sampling of human genetic diversity: a survey of worldwide populations by high-density genotyping, *Genomics*, *96*, 199–210.
- Yang, Z. (2000), Complexity of the simplest phylogenetic estimation problem, *Proc. R. Soc. Lond. B*, *267*, 109–116.
- Zwickl, D. J., and D. M. Hillis (2002), Increased taxon sampling greatly reduces phylogenetic error, *Syst. Biol.*, *51*, 588–598.