© Copyright, Princeton University Press. No part of this book may be distributed, posted, or reproduced in any form by digital or mechanical means without prior written permission of the publisher.

Short Contents

Preface xvii

Acknowledgments xix

List of Contributors xxi

CHAPTER 1 Introduction to Species Tree Inference 1
L. Lacey Knowles and Laura S. Kubatko

PART I ANALYTICAL AND METHODOLOGICAL DEVELOPMENTS 15

CHAPTER 2 Large-Scale Species Tree Estimation 19
Erin Molloy and Tandy Warnow

CHAPTER 3 Species Tree Estimation Using ASTRAL: Practical Considerations 43
Siavash Mirarab

CHAPTER 4 Species Tree Estimation Using Site Pattern Frequencies 68
David L. Swofford and Laura S. Kubatko

CHAPTER 5 Practical Aspects of Phylogenetic Network Analysis Using PhyloNet 89
Zhen Cao, Xinhao Liu, Huw A. Ogilvie, Zhi Yan, and Luay Nakhleh

CHAPTER 6 Network Thinking: Novel Inference Tools and Scalability Challenges 120
Claudia Solís-Lemus

For general queries, contact webmaster@press.princeton.edu

© Copyright, Princeton University Press. No part of this book may be distributed, posted, or reproduced in any form by digital or mechanical means without prior written permission of the publisher.

PART II Empirical Inference 145

CHAPTER 7 Phylogenomic Conflict in Plants 149
Joseph F. Walker and Stephen A. Smith

CHAPTER 8 Hybridization in *Iochroma* 161
Daniel J. Gates, Diana Pilson, and Stacey D. Smith

CHAPTER 9 Hybridization and Polyploidy in *Penstemon* 175
Paul D. Blischak, Coleen E. Thompson, Emiko M. Waight, Laura S. Kubatko, and Andrea D. Wolfe

CHAPTER 10 Comparison of Linked versus Unlinked Character Models for Species Tree Inference 191
Kerry Cobb and Jamie R. Oaks

PART III Beyond the Species Tree 211

CHAPTER 11 The Unfinished Synthesis of Comparative Genomics and Phylogenetics: Examples from Flightless Birds 215
Alexandria A. DiGiacomo, Alison Cloutier, Phil Grayson, Timothy B. Sackton, and Scott V. Edwards

CHAPTER 12 Phylogenetic Analysis under Heterogeneity and Discordance 232
James B. Pease and Ellen I. Weinheimer

CHAPTER 13 The Multispecies Coalescent in Space and Time 251
Patrick F. McKenzie and Deren A. R. Eaton

CHAPTER 14 Tree Set Visualization, Exploration, and Applications 260
Jeremy M. Brown, Genevieve G. Mount, Kyle A. Gallivan, and James C. Wilgenbusch

Bibliography 277

Index 317

© Copyright, Princeton University Press. No part of this book may be distributed, posted, or reproduced in any form by digital or mechanical means without prior written permission of the publisher.

Contents

Preface xvii

Acknowledgments xix

List of Contributors xxi

CHAPTER 1 Introduction to Species Tree Inference 1
1.1 Introduction 1
1.2 Background and Terminology 2
1.2.1 Definitions and Terminology 2
1.2.2 An Introduction to the Multispecies Coalescent 5
1.2.3 Data Types and Technologies for Generating Phylogenomic Data 6
1.3 Overview of Current Methods for Species Tree Inference 9
1.3.1 Controversies in the Estimation of Species Trees 11
1.4 A Look to the Future 12
1.4.1 Current Limitations and Future Prospects 12
1.4.2 Beyond the Species Tree 13
1.5 Organization of This Book 14

PART I Analytical and Methodological Developments 15

CHAPTER 2 Large-Scale Species Tree Estimation 19
2.1 Introduction 19
2.2 Species Tree Estimation Methods Addressing ILS 21
2.2.1 Overview 21
2.2.2 Summary Methods 21
2.2.3 Coestimation Methods 24
2.2.4 Site Based Methods 26
2.2.5 Evaluation of Branch Support in Species Trees 28
2.3 Species Tree Estimation under GDL 29
2.4 Parallel Implementations for Species Tree Estimation 30
2.4.1 ASTRAL-MP 30
2.4.2 Multilocus Species Tree Estimation Using Maximum Likelihood 31

© Copyright, Princeton University Press. No part of this book may be distributed, posted, or reproduced in any form by digital or mechanical means without prior written permission of the publisher.

2.5 Divide-and-Conquer Species Tree Estimation 33

2.5.1 Divide-and-Conquer Using Supertree Methods 34

2.5.2 Divide-and-Conquer Using Disjoint Tree Merger Methods 34

2.6 Choice of Method 36

2.6.1 Statistical Consistency 36

2.6.2 Empirical Performance 37

2.7 Summary, Challenges, and Future Directions 39

2.8 Appendix: Big-O Analysis 41

CHAPTER 3 Species Tree Estimation Using ASTRAL: Practical Considerations 43

3.1 Introduction 43

3.2 ASTRAL Algorithm 46

3.2.1 Motivation and History 46

3.2.2 ASTRAL Algorithm 47

3.2.3 Summary of Known Theoretical Results Related to ASTRAL 50

3.3 Accuracy 51

3.4 Running Time 54

3.5 Input to ASTRAL: Practical Considerations 54

3.5.1 Gene Tree Estimation 55

3.5.2 Filtering of Data 57

3.6 ASTRAL Output 61

3.6.1 Species Tree Topology and Its Quartet Score 61

3.6.2 Branch Lengths in Coalescent Units 61

3.6.3 Branch Support Using Local Posterior Probability (localPP) 64

3.7 Follow-up Analyses and Visualization 65

3.7.1 Tests for Polytomies 65

3.7.2 Per Branch Quartet Support (Measure of Discordance) 65

3.8 Conclusion 66

CHAPTER 4 Species Tree Estimation Using Site Pattern Frequencies 68

4.1 Introduction 68

4.2 Estimation of the Species Tree Topology Using SVDQuartets 69

4.2.1 Theoretical Basis 69

4.2.2 Accounting of Incomplete Lineage Sorting in SVDQuartets 74

4.2.3 Species Tree Inference: Quartet Sampling and Assembly 75

4.2.4 Algorithmic Details 76

4.2.5 Uncertainty Quantification 78

4.2.6 Application to Species Relationships among Gibbons 78

4.2.7 Properties of SVDQuartets 79

4.2.8 Recommendations for Using SVDQuartets 82

4.3 Estimation of Speciation Times 82

4.3.1 Theoretical Basis 83

4.3.2 Algorithmic Details 86

© Copyright, Princeton University Press. No part of this book may be distributed, posted, or reproduced in any form by digital or mechanical means without prior written permission of the publisher.

4.3.3 Uncertainty Quantification 86
4.3.4 Application to Species Relationships among Gibbons 87
4.3.5 Recommendations for Using Composite Likelihood Estimators of the Speciation Times 87
4.4 Conclusion and Future Work 87

CHAPTER 5 **Practical Aspects of Phylogenetic Network Analysis Using PhyloNet 89**
5.1 Introduction 89
5.2 Reading and Interpretation of a Phylogenetic Network 91
5.2.1 Phylogenetic Network Parameters and Their Identifiability 92
5.3 Heuristic Searches, Point Estimates, and Posterior Distributions, or, Why Am I Getting Different Networks in Different Runs? 92
5.4 Illustration of the Various Inference Methods in PhyloNet 96
5.4.1 Inference under the MDC Criterion 96
5.4.2 Maximum Likelihood Inference 98
5.4.3 Maximum Pseudolikelihood Inference 102
5.4.4 Bayesian Inference 103
5.4.5 Running Time 105
5.5 Analysis of Larger Data Sets 106
5.6 Comparison and Summarization of Networks 111
5.6.1 Displayed Trees 111
5.6.2 Backbone Networks 111
5.6.3 Tree Decompositions 112
5.6.4 Tripartitions 112
5.6.5 Major Trees 112
5.7 Reticulate Evolutionary Processes in PhyloNet 112
5.7.1 Analysis of Polyploids 114
5.8 Conclusions 117
Notes 119

CHAPTER 6 **Network Thinking: Novel Inference Tools and Scalability Challenges 120**
6.1 Introduction: The Impact of Gene Glow 120
6.2 Trees versus Networks 122
6.3 Species Networks 124
6.3.1 Explicit versus Implicit Networks 126
6.3.2 Extended Parenthetical Format 127
6.3.3 Displayed Trees and Subnetworks 128
6.3.4 Comparison of Networks 128
6.4 Fast Reconstruction of Species Networks 129
6.4.1 Maximum Pseudolikelihood Estimation 130
6.4.2 Rooting of Semidirected Networks 136
6.4.3 Goodness of Fit Tools 139
6.4.4 Bootstrap Analysis 140

© Copyright, Princeton University Press. No part of this book may be distributed, posted, or reproduced in any form by digital or mechanical means without prior written permission of the publisher.

6.5 Appendix: Installation and Use of the PhyloNetworks Julia Package 143
6.5.1 Main Functions in PhyloNetworks 143

PART II Empirical Inference 145

CHAPTER 7 Phylogenomic Conflict in Plants 149
7.1 Introduction 149
7.2 Two Examples of Gene Tree Conflict within Angiosperms 152
7.3 The Consequences of Gene Tree Conflict in Phylogenomics 154
7.3.1 Inference of Species Trees 154
7.3.2 Gene Duplication and Genome Duplication 157
7.3.3 Divergence Time and Comparative Analyses 158
7.4 Resolution of the Tree of Plant Life 160

CHAPTER 8 Hybridization in *Iochroma* 161
8.1 Introduction 161
8.2 Methods 163
8.2.1 Study System 163
8.2.2 Experimental Design 165
8.2.3 Target Capture and Assembly 166
8.2.4 Detection of Patterns of Hybridization from Gene Tree Distributions 167
8.2.5 Testing of Hybridization in Empirical Data Sets 168
8.3 Results 168
8.3.1 Addition of Hybrid Taxa Increases Discordance and Decreases Tree-Like Signal 168
8.3.2 Tests of Hybridization Support Different Relationships than Expected 170
8.4 Discussion 172
8.4.1 Effects of Hybridization on Patterns of Gene Tree Discordance 172
8.4.2 Challenges in Determining the Exact Hybrid Relationships 172
8.4.3 Hybridization in *Iochrominae* 173
8.5 Conclusions 174

CHAPTER 9 Hybridization and Polyploidy in *Penstemon* 175
9.1 Introduction 175
9.2 Approach 176
9.2.1 Calculation of Quartet Concordance Factors 177
9.2.2 Bootstrapping and Gene Tree Uncertainty 178
9.2.3 Validation of QCF Estimation 178
9.2.4 Implementation 179
9.3 Materials and Methods 179
9.3.1 Study System 179
9.3.2 Sample Collection, DNA Extraction, and Amplicon Sequencing 180

For general queries, contact webmaster@press.princeton.edu

© Copyright, Princeton University Press. No part of this book may be distributed, posted, or reproduced in any form by digital or mechanical means without prior written permission of the publisher.

9.3.3 Species Tree Inference 181
9.3.4 Candidate Hybridization Events from Rooted Triples 181
9.3.5 Species Network Inference 182
9.4 Results 182
9.4.1 Nuclear Amplicon Data 182
9.4.2 Species Tree Inference 182
9.4.3 Tests for Hybridization and Species Network Inference 186
9.5 Discussion 186
9.5.1 Taxonomy of Subsections *Humiles* and *Proceri* 188
9.5.2 Character Evolution and Biogeography 189
9.5.3 Phylogenetics of Hybrids and Polyploids 189
9.6 Conclusions 190

CHAPTER 10 Comparison of Linked versus Unlinked Character Models for Species Tree Inference 191
10.1 Introduction 191
10.2 Methods 192
10.2.1 Simulations of Error-Free Data Sets 192
10.2.2 Introduction of Site Pattern Errors 193
10.2.3 Assessment of Sensitivity to Errors 194
10.2.4 Project Repository 194
10.3 Results 195
10.3.1 Behavior of Linked (StarBEAST2) versus Unlinked (Ecoevolity) Character Models 195
10.3.2 Analysis of All Sites versus SNPs with Ecoevolity 195
10.3.3 Coverage of Credible Intervals 197
10.3.4 MCMC Convergence and Mixing 197
10.4 Discussion 197
10.4.1 Robustness to Character-Pattern Errors 207
10.4.2 Relevance to Empirical Data Sets 208
10.4.3 Recommendations for Using Unlinked-Character Models 209
10.4.4 Other Complexities of Empirical Data in Need of Exploration 209

PART III Beyond the Species Tree 211

CHAPTER 11 The Unfinished Synthesis of Comparative Genomics and Phylogenetics: Examples from Flightless Birds 215
11.1 Introduction 215
11.1.1 Phylogenetics of Modern Birds 216
11.1.2 Paleognathous Birds as a Test Case for Post-Genomic Phylogenetics 218
11.2 Building of a Whole-Genome Species Tree for an Ancient Radiation of Birds 218
11.3 The Unfinished Synthesis of Comparative Genomics and Genomic Heterogeneity 225
11.3.1 A Species Tree for Paleognathous Birds as a Foundation for Comparative Genomics 225

© Copyright, Princeton University Press. No part of this book may be distributed, posted, or reproduced in any form by digital or mechanical means without prior written permission of the publisher.

11.3.2 Accommodation of Uncertainty into Whole-Genome Alignments 225
11.3.3 Gene Tree Heterogeneity and Detecting Rate Variation in Genes and Noncoding Regions 228
11.3.4 Phylogenetic Analysis of Quantitative ’Omics Data: Gene Expression and Epigenetics 230
11.4 Conclusions 231

CHAPTER 12 **Phylogenetic Analysis under Heterogeneity and Discordance 232**
12.1 Introduction 232
12.2 The Origin of Discordance 232
12.2.1 A History of Systems and Methods 232
12.2.2 Concepts of Harmony and Discordance 234
12.2.3 The Species Tree 236
12.2.4 Comparison of the Incomparable 238
12.3 Characterization and Quantification of Phylogenetic Heterogeneity 238
12.3.1 Quantification and Visualization of Discordance 238
12.3.2 Quantification of Conflict and Tree Evaluation 240
12.3.3 Visualization of Conflict 241
12.4 Analysis under Phylogenetic Heterogeneity 243
12.4.1 Testing of Introgression and Hybridization under Phylogenetic Heterogeneity 243
12.4.2 Testing of Selection under Phylogenetic Heterogeneity 245
12.4.3 Testing of Traits under Phylogenetic Heterogeneity 247
12.4.4 Testing of Coevolution under Phylogenetic Heterogeneity 249
12.5 Conclusion 250

CHAPTER 13 **The Multispecies Coalescent in Space and Time 251**
13.1 Introduction 251
13.2 Coalescent Simulations 252
13.2.1 Units, Space, and Time 253
13.2.2 Tree Size, Tree Space, and Phylogenetic Decay 255
13.3 Linked Genealogies and Gene Tree Inference 256
13.4 Conclusions 258

CHAPTER 14 **Tree Set Visualization, Exploration, and Applications 260**
14.1 Introduction to Visualizing and Exploring Tree Sets 260
14.1.1 Tree Set Visualization 261
14.1.2 Detection of Structure in Tree Sets 262
14.2 Applications to Gene Trees, Species Trees, and Phylogenomics 264
14.2.1 Sensitivity to Models of Sequence Evolution 264
14.2.2 Joint versus Independent Inference of Gene Trees 268

© Copyright, Princeton University Press. No part of this book may be distributed, posted, or reproduced in any form by digital or mechanical means without prior written permission of the publisher.

14.2.3 Understanding of Variation across Genomes 271

14.2.4 Prospects for Future Development and Application 275

14.3 Appendix 275

Bibliography 277

Index 317

For general queries, contact webmaster@press.princeton.edu

© Copyright, Princeton University Press. No part of this book may be distributed, posted, or reproduced in any form by digital or mechanical means without prior written permission of the publisher.

CHAPTER 1

Introduction to Species Tree Inference

L. Lacey Knowles and Laura S. Kubatko

1.1 Introduction

Estimation of the evolutionary relationships among a collection of organisms remains a central focus of much of evolutionary and ecological study within the field of biology as these relationships provide the background for testing hypotheses in these fields. For example, support for different hypotheses about early animal evolution, and in particular the evolution of sophisticated cell types such as nerve and muscle cells, was contingent upon the phylogenetic relationships among the earliest diverging animal lineages. Especially important in addressing these questions was the placement of Ctenophora because of their shared complex cell types with bilaterians [642]. As another example, accurate time and rate estimation forms the basis for questions in ecology and evolution [468], with shifts in rates being central to tests about the drivers of diversification (e.g., [143, 596]). Clearly, accurate estimation of phylogenetic relationships that can leverage all available data within a firm inferential framework are crucial to addressing questions such as these.

Within the last 20 years, the field of phylogenetics has grown rapidly, both in the quantity of data available for inference and in the number of methods available for phylogenetic estimation. Our first book, *Estimating Species Trees: Practical and Theoretical Aspects*, published in 2010, gave an overview of the state of phylogenetic practice for analyzing multilocus sequence data at the time, but much has changed since then. Indeed, the rapid pace at which the field has advanced in the intervening time has led to the need for an updated reference. We intend this book both to serve as an update on current practice within the field and to provide a timely look toward the future.

We begin this chapter with a brief recap of the history of species tree estimation, including definitions and basic terminology. We next discuss both opportunities and challenges in the field. This discussion includes a critical look at the limitations currently imposed by data availability and computational power and how these might be expected to change in the future, but it also addresses uncertainty surrounding sampling and data analysis in the wake of the big data wave sweeping phylogenetics. We then consider inference beyond the species tree, highlighting the important problems that a genome-scale phylogeny and underlying data allow us to address in a rigorous inferential framework. We conclude with an overview of the book and its organization.

© Copyright, Princeton University Press. No part of this book may be distributed, posted, or reproduced in any form by digital or mechanical means without prior written permission of the publisher.

1.2 Background and Terminology

Prior to the routine collection of DNA sequence data, the fields of population genetics and phylogenetics were largely viewed as distinct as they addressed questions at different evolutionary time scales. Much of the mathematical and statistical development of models at the within-population scale was undertaken in the 1980s, through contributions by Kingman [364, 365, 363] and others (e.g., [746, 745]) that resulted in what is now known as *Kingman's coalescent model*, a continuous-time approximation of the Wright–Fisher (and other) population-level models. Kingman's coalescent today forms the theoretical basis for many of the methods used for species tree inference.

Following these developments, several authors noted that when Kingman's coalescent model was applied across species, inferred evolutionary relationships might vary from gene to gene. Important contributions to the development of these ideas, including mathematical details, were provided by [743], [784], [744], and [559], among others. However, much of this work went unnoticed by the phylogenetics community until the mid-1990s, when a seminal paper by Maddison [455] provided clear descriptions of the possible causes of differences in gene-level and species-level phylogenies. This coincided with a decrease in the cost of DNA sequencing, and the subsequent availability of multilocus sequence data prompted several authors to highlight the need for new inferential frameworks to accommodate these data properly [813, 538, 633, 634].

Importantly, the potential for differences between gene trees and species trees were also recognized to result not only from the coalescent process but also from other evolutionary processes, such as horizontal transfer and gene duplication and loss. By the early 2000s, several papers highlighted the possibility of variation in the evolutionary history across the genome in carefully annotated empirical data sets (e.g., [134, 630, 213]), and the need for methodology that specifically aimed to estimate species-level phylogenetic trees became well accepted by many in the community.

1.2.1 DEFINITIONS AND TERMINOLOGY

A species tree or species phylogeny can be defined as a rooted bifurcating phylogenetic tree for which the tips of the tree represent species and the internal nodes represent speciation events. The times associated with internal nodes of the tree represent the times of speciation events, and branch lengths along the species phylogeny represent the amount of time between speciation events. Speciation times are often given in coalescent units, which can be defined as the number of $2N_e$ generations, where N_e is the effective population size. The advantage of using coalescent units to describe speciation times is that a standardized unit can be discussed in such a way that characteristics associated with this unit can be "translated" to any species of interest once the generation time in years and the effective population size are specified. When N_e varies across the tree, it may be more difficult to define an appropriate unit (number of generations is a reasonable choice, see [446]). Mutation units, the unit commonly used for gene tree inference that is given by the number of substitutions per site per unit time, are also sometimes used. Figure 1.1 shows an example species phylogeny for three taxa, labeled A, B, and C (shaded, thicker tree in each panel).

A gene tree represents the evolutionary history for an individual gene, where a gene is defined as a stretch of contiguous sequence of any length. The tips of a gene tree represent sequences collected from individuals sampled from a particular species, while the internal nodes represent gene divergence times (looking forward in time) or common

© Copyright, Princeton University Press. No part of this book may be distributed, posted, or reproduced in any form by digital or mechanical means without prior written permission of the publisher.

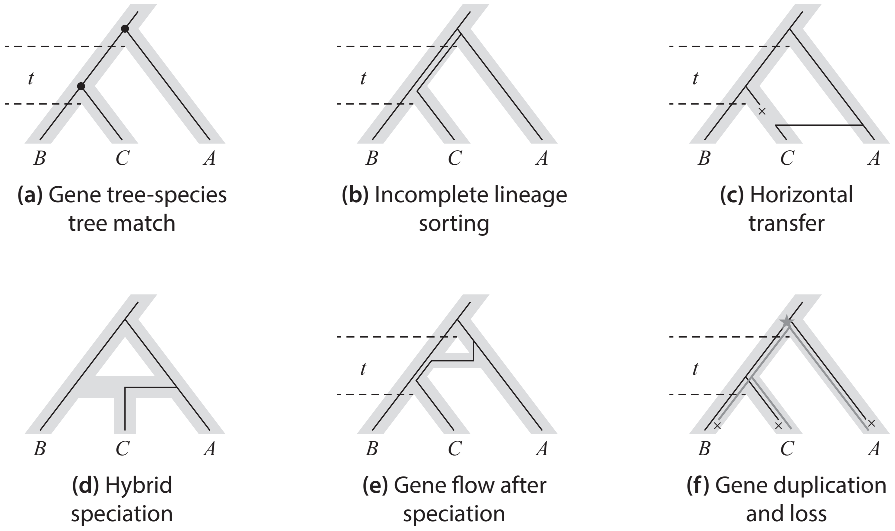
t
B
C
A
(a) Gene tree-species tree match
t
B
C
A
(b) Incomplete lineage sorting
t
B
C
A
(c) Horizontal transfer
B
C
A
(d) Hybrid speciation
t
B
C
A
(e) Gene flow after speciation
t
B
C
A
(f) Gene duplication and loss

Figure 1.1. Relationships between gene trees and species trees. In each panel, the species tree is represented by the shaded, thicker tree. Speciation events are indicated with horizontal dotted lines, and the length of time between speciation events is denoted by *t*. Gene divergence, or coalescent, events are indicated in panel (a) by black circles. Each panel shows a possible relationship between the gene tree and the species tree resulting from a specific evolutionary process: (a) The gene tree and species tree share the same topology. (b) The topologies of the gene and species trees are discordant due to incomplete lineage sorting. Tracing the lineages sampled from species *B* and species *C* back in time, we see that they fail to coalesce in the immediately ancestral population, and instead the lineage sampled from species *C* coalesces with that sampled from *A* in the common ancestral population. (c) Genetic information is transferred horizontally across the phylogeny from species *A* to species *C*, leading to a gene tree that is discordant with the species tree. (d) A species network in which species *C* is a hybrid of species *A* and *B* is shown. For the particular gene sampled, species *C* inherited its genetic material from species *A*. Owing to the hybrid speciation event, it is possible for *C* to inherit genetic information directly from either *B* or *A*, even in the absence of incomplete lineage sorting. (e) Gene tree discordance due to gene flow from *A* to *C* following speciation. (f) A gene duplication event, marked by a star, occurs after the separation of the lineage leading to *A* from the ancestor of *B* and *C*; the duplicated lineage is sampled in *A* and *C*, while the original lineage is sampled in *B*, leading to discordance between the gene tree and species tree. See also figure 7.1.

ancestor events for the sampled sequences (looking backward in time). These are sometimes also called coalescent events. A gene tree may have many more tips than a species tree because multiple individuals may be sampled within each species included in the species phylogeny. A gene tree may differ from the species tree that gives rise to it both in terms of its topology (branching pattern) and in terms of the times associated with its nodes. Differences in topology between gene trees and the species tree can result from many different evolutionary processes. For example, incomplete lineage sorting (i.e., the failure of lineages to coalesce in their immediately ancestral population) can lead to gene trees with topologies that differ from the species tree (see figure 1.1b). This form

© Copyright, Princeton University Press. No part of this book may be distributed, posted, or reproduced in any form by digital or mechanical means without prior written permission of the publisher.

of gene tree discordance is typically modeled by applying Kingman's coalescent across the phylogeny (which is then commonly referred to as the *multispecies coalescent*) and is well studied; in particular, the probability distributions of both gene tree topologies [179] and gene genealogies [601] have been derived.

Horizontal transfer (figure 1.1c) is another evolutionary process that is well-known to generate discord between gene trees and the species tree and refers to any process by which genetic information is moved from one species to another by means other than modification with descent. For example, in bacteria, horizontal transfer occurs when distinct bacterial strains recombine to generate unique sequences that include genetic material from both strains. In sexually reproducing organisms, horizontal transfer can occur when a virus or other vector moves a segment of DNA from one species' genome to another. Hybridization (figure 1.1d) and introgression/gene flow (figure 1.1e) can also be thought of as forms of horizontal transfer, in that these processes both involve the exchange of genetic material between distinct, contemporaneous species (i.e., "horizontally" along the phylogeny) rather than through a process of descent with modification within a single species. Regardless of the precise mechanism by which the horizontal transfer occurs, such processes can result in portions of the genome that are inherited differently than others. For example, introgressed loci will show a pattern of inheritance from a species different than that of the majority of the genome if the introgression occurs between non-sister taxa (e.g., figure 1.1c). In the absence of other processes, the extent of discordance due to horizontal inheritance will depend on the extent to which genetic material has been transferred from one species to another throughout the evolutionary history of the set of species under consideration.

The process of gene duplication and loss (figure 1.1f) provides another evolutionary mechanism that results in differences between gene trees and species trees. When a gene is duplicated in a genome, the two versions of the gene subsequently evolve independently of one another, and in descendent species one or both versions of the gene may be present in the genome being sampled. Depending on which copy is sampled, the gene tree for the locus under consideration may differ from the true species-level relationship. Loss of one copy of a duplicated gene may also lead to incongruence between the gene tree and the species tree, or may result in missing data for the locus under consideration, depending on the time that has passed since the duplication and loss events. Gene duplication and loss is prevalent in many species and provides an important mechanism for the generation of new gene function (e.g., a duplicated copy of the gene is under less evolutionary constraint and may evolve to provide a new function in the organism). Thus, consideration of this evolutionary process at the stage of species tree inference is crucial, and many methods have been and continue to be proposed for inference in the presence of duplication and loss.

Closely related to the concept of a species tree is that of a species network, in which relationships between species are depicted by a sequence of speciation events, as in a species tree, but in which species may arise from more than one immediately ancestral species. This may result from evolutionary processes such as hybrid speciation (figure 1.1d), extensive gene flow between distinct species (figure 1.1e), or other forms of horizontal transfer. Much recent work has focused on carefully defining species networks and developing methods of inferring such networks from phylogenomic data, often within a coalescent framework (see, e.g., [845, 841, 843, 713, 861], as well as chapters 5 and 6 of this volume).

For general queries, contact webmaster@press.princeton.edu

© Copyright, Princeton University Press. No part of this book may be distributed, posted, or reproduced in any form by digital or mechanical means without prior written permission of the publisher.

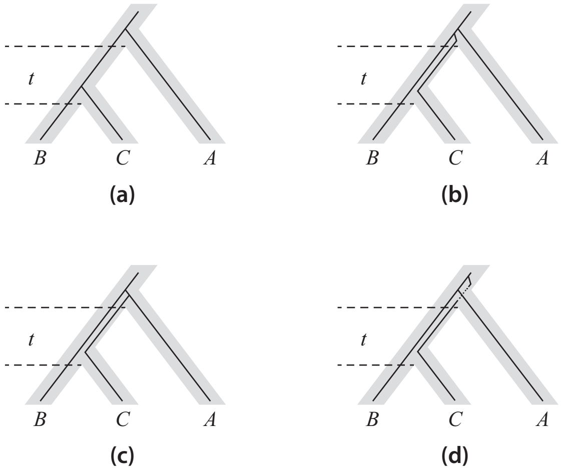

Figure 1.2. Four coalescent histories compatible with a three-taxon species tree. Note that the histories in (a) and (b) share the same topology as the species tree, while those in (c) and (d) do not.

1.2.2 AN INTRODUCTION TO THE MULTISPECIES COALESCENT

As mentioned in the previous section, the multispecies coalescent model underlies many of the methods for species tree inference that are commonly applied to multilocus data. Rather than provide a complete mathematical description of this model, we provide here an introduction to the main ideas for three-taxon trees. Readers wishing to see a more full description can consult [383, 770, 289].

Figure 1.2 shows the same three-taxon species trees as shown in figure 1.1. Embedded within the species tree are the four possible *coalescent histories* consistent with this species tree, where coalescent histories refer both to the gene tree topology and the species tree branch lengths along which coalescent events occur. Note that the history in figure 1.2a is the only one in which the first coalescent event occurs within the species branch of length t. Under Kingman's coalescent, times to coalescent events follow an exponential distribution with rate given by $\binom{n}{2}$ when n lineages are available to coalesce. Since $n = 2$ lineages are available to coalesce in the interval of length t in figure 1.2a, the probability of observing this history is the probability that an exponential random variable with rate 1 is less than t, which is $1 - e^{-t}$.

Since the probability associated with all four histories must sum to 1, this leaves e^{-t} of the probability to be distributed over the other three histories, shown in figure 1.2b–d. Note that these three histories all involve the first coalescent event occurring above the root of the species tree, and all three lineages are available to coalesce within this ancestral population. Under Kingman's coalescent, each pair of lineages is equally likely to be the first to coalesce, and thus each of these histories has probability $\frac{1}{3}e^{-t}$.

Finally, we note that the first two histories (figure 1.2a and b) have the same gene tree topology. Thus to derive the probability distribution of gene tree topologies, we can add these two probabilities. The coalescent model then specifies that for three species, the gene tree topology that matches the species tree occurs with probability $1 - \frac{2}{3}e^{-t}$, while the two nonmatching gene trees each have probability $\frac{1}{3}e^{-t}$. Noting that $1 - \frac{2}{3}e^{-t} \geq \frac{1}{3}e^{-t}$ with equality only when $t = 0$, we can identify a common pattern for which the coalescent model is a good fit: a dominant gene tree topology that occurs with highest frequency (the one matching the species tree) with the two alternative topologies occurring in lower and approximately equal frequencies. Such a pattern has

For general queries, contact webmaster@press.princeton.edu

© Copyright, Princeton University Press. No part of this book may be distributed, posted, or reproduced in any form by digital or mechanical means without prior written permission of the publisher.

been observed for empirical data [565, 145], and deviation from this pattern has been used as evidence for introgression [652].

1.2.3 DATA TYPES AND TECHNOLOGIES FOR GENERATING PHYLOGENOMIC DATA

New data collection techniques have driven shifts in not only the quantity of data but also in the types of data available for phylogenetic inference, with a variety of high-throughput phylogenetic data collection technologies to choose from (table 1.1). These range from different types of targeted sequencing technologies (e.g., hybrid enrichment strategies; [422, 611]) to random genomic sequencing (e.g., reduced representation restriction site-associated DNA sequencing [RADseq]) or targeted genotyping-by-sequencing (GBS) (e.g., RAPTURE; see [8, 60]) and whole transcriptome or genome sequencing.

One important factor in deciding among the different technologies is the differences in their costs, both in terms of the initial time investments and expense but also associated costs when expanding to large numbers of taxa (or individuals). For example, amplifying targeted amplicons involves substantial costs for setup, but it is relatively inexpensive to capture sequences, whereas random genomic sequences from RADseq technologies are economical and provide a universal approach for collecting comparative genomic data. As sequencing costs drop, whole transcriptome and genome sequencing are becoming more widely applied [853, 425]. Alternatively, RADseq can generate very large numbers of loci (i.e., in the thousands to millions of loci) while being scalable to large sample sizes [414], including hundreds of thousands of individuals with targeted genotyping-by-sequencing, and, because of the short sequence reads, they are amenable for applications to museum specimens for which DNA degradation can preclude large amplicons [837].

Another primary consideration for choosing a technology (besides the cost and ease of setup) is differences in their utility. For example, the very large numbers of loci generated by technologies like RADseq become highly desirable for estimation of phylogenetic relationships at recent time scales (e.g., [475, 465]). However, their utility drops as the evolutionary distances between taxa increase (but see [779]) because of allele dropout (but see [210]), which will result in missing data among more distantly related taxa (i.e., homologs will not be sequenced in some taxa because of mutations in the enzyme cutter sites, although new technologies guard against allelic dropout; see [84]). Decisions about what threshold of missing data to use for analysis of RADseq data is complicated. Eliminating loci with a lot of missing data can result in a biased data set with an overrepresentation of loci with low mutation rates [318], which means the data set may not contain the actual loci that are phylogenetically informative for resolving relationships among taxa that diversify rapidly—that is, loci with the highest rate of evolution. On the other hand, discordant relationships have been shown to be disproportionately represented among loci with missing data [413], suggesting that they may be less reliable for phylogenetic inference. Whole-genome or transcriptome sequencing has the appeal of providing not just a lot of data for phylogenetic inference but also information to address questions provided by the phylogenetic framework, including questions about genome evolution [428]. However, in addition to assembly challenges, such data also pose new challenges because of the potential heterogeneity of processes contributing to genomic differences among taxa, making model misspecification a more pressing problem compared with the relatively small data sets (e.g., hundreds to a few thousand loci). In contrast, targeted amplicon approaches such as

© Copyright, Princeton University Press. No part of this book may be distributed, posted, or reproduced in any form by digital or mechanical means without prior written permission of the publisher.

Table 1.1. Summary of sequencing technologies.

Method	Description	Data	Reference

Targeted enrichment	Also known as "hybrid enrichment," "anchored enrichment," and "enrichment"; may also be referred to as "capture." Probes or baits complementary to the sequence of interest are used to hybridize with the target sequence, which is then enriched via PCR and sequenced on a HTS platform.	Typical enrichment/capture data sets for phylogenomics contain tens to hundreds of taxa and hundreds to thousands of loci.	Mertes et al. 2011 (https://doi.org/10.1093/bfgp/elr033);
			Cronn et al. 2012 (https://doi.org/10.3732/ajb.1100356);
			Folk et al. 2015 (https://doi.org/10.3732/apps.1500039)

Transcriptome sequencing	This method involves sequencing RNA reads (typically via RNA-Seq on a short read HTS platform) and assembling the reads into the transcriptome. When transcriptomes are used for phylogenomic inference, a necessary step is orthology inference to ensure that markers have sequence homology.	This method can generate hundreds to thousands of loci. The number of taxa that can be included is more variable among studies but is typically fewer than 100 due to costs.	Wen et al. 2013 (https://doi.org/10.1371/journal.pone.0074394); Shen et al. 2017 (https://doi.org/10.1093/gigascience/gix116)
Whole-genome sequencing	Whole-genome sequencing for phylogenomics has the advantage of obtaining an unbiased representation of genomic data relative to reduced representation methods. As costs for genome sequencing fall, and computational efficiency increases, this method is becoming increasingly accessible to many researchers. Orthology assessment is also a necessary step when using loci from whole-genome sequencing.	Currently, whole-genome uses similar amounts of data as transcriptome sequencing—thousands of loci for tens to hundreds of individuals.	Allen et al. 2017 (https://doi.org/10.1093/sysbio/syw105)

(*continued*)

© Copyright, Princeton University Press. No part of this book may be distributed, posted, or reproduced in any form by digital or mechanical means without prior written permission of the publisher.

Table 1.1. (*continued*)

Method	Description	Data	Reference
RAD-Seq / GBS	This is a reduced representation library approach in which 1–3 restriction enzymes are used to fragment the genome. Adapters with built-in barcodes are ligated to the fragments, then the fragments are pooled and size selected. The pool is typically sequenced on a short-read HTS platform (e.g., Illumina).	The number of loci sequenced at a given depth depends on the genome size and number of individuals multiplexed in a sequencing run. For example, a user could reasonably expect to get 5,000–20,000 loci with 10X coverage when multiplexing 1000 individuals from a species with a 1 Gb genome in a lane of Illumina HiSeq.	Andrews et al. 2016 (https://doi.org/10.1038/nrg.2015.28)
Parallel microfluidic PCR + HTS	This approach uses primer pairs designed for PCR and utilizes microfluidic PCR to amplify loci in parallel using HTS. The primers can be designed for organelles or the nuclear genome, but they must have similar annealing temperatures since the PCR will be processed simultaneously in parallel. This method bypasses traditional library preparation.	Typical studies use 48–96 loci for approximately 50–100 taxa.	Uribe-Convers et al. 2016 (https://doi.org/10.1371/journal.pone.0148203); Kates et al. 2017 (https://doi.org/10.1016/j.ympev.2017.03.002)

Note: HTS = high-throughput sequencing, RADseq = restriction site-associated DNA sequencing, and GBS = genotyping by sequencing.

© Copyright, Princeton University Press. No part of this book may be distributed, posted, or reproduced in any form by digital or mechanical means without prior written permission of the publisher.

hybrid enrichment approaches avoid the problems of missing data by relying on conserved sets of priming sites to amplify sequences. They also present less of a challenge for assembly, modeling, and analysis compared with technologies like RADseq and whole-genome/transcriptome sequencing. However, they also result in substantially fewer loci, and because they rely on specific priming sites, they are nonrandom samples of the genome, which may make them less desirable for some questions.

These different data set properties (e.g., SNP-based information content, or inherent heterogeneity in underlying evolutionary model with genomic-scale sampling, and/or differing amounts and distributions of missing loci in data sets) are likewise driving different analytical and theoretical areas in phylogenetic inference. These new areas range from exciting new approaches for phylogenetic estimation and the evaluation of the confidence of such relationships (e.g., assessing phylogenetic signal; [423, 771]) to determination of the different processes contributing to locus-specific patterns of ancestry (e.g., [88, 371, 771]) and identification of subsets of data for phylogenetic inference from genome-scale data sets [675, 192, 319]. The analytical methods that might be applied will also differ depending on the technology used to generate the data. For example, the short sequence reads of RADseq means that they are not generally amenable to gene tree estimation but instead are analyzed as SNP data, whereas standard gene tree estimation methods are applied to sequences generated from technologies like hybrid enrichment because those technologies target specific genomic regions of longer read lengths. Likewise, with genome-scale data sets, computational challenges restrict the types of analyses that might be done [503].

The new technologies and unprecedented abundance of data they generate is changing phylogenetic inference and no doubt providing better resolved and more reliable phylogenetic inference in some cases. However, recalcitrant nodes persist (e.g., [798, 590]). Moreover, with phylogenetic estimates differing as a function of analysis, data set design, or inclusion/exclusion of loci, genome-scale data sets are raising many questions with no clear answers. For example, how might genome-scale data be analyzed to provide reliable phylogenetic estimates? If subsets of the data are to be analyzed, how should such data be identified (both in terms of loci and taxa)? These are some of the questions that are explored in this book, as researchers contend with the uncertainty surrounding sampling and data analysis in the big data era. Despite these unknowns, it is clear that along with these complicated questions come some amazing opportunities that extend beyond a focus on the species tree itself. As we look to the future, and in the following chapters, we emphasize this expanded role of genome-scale data—that is, next-generation inference, which will no doubt become the new focus of researchers as next-generation sequencing becomes routine (table 1.1).

1.3 Overview of Current Methods for Species Tree Inference

Given the processes described above, the precise mechanism by which data arise must be taken into consideration in the development of methods for inferring species-level phylogenies. Regardless of the process(es) responsible for gene tree–species tree discordance, it is usually assumed that gene trees arise from evolutionary processes occurring along the species tree, and DNA sequence data are subsequently generated from the gene trees associated with individual loci. Thus, DNA sequences observed from loci that are freely recombining can be viewed as conditionally independent of one another, where the conditioning is based on their underlying gene trees arising from a shared species phylogeny. Inference then proceeds in the “reverse” direction—that is, given a set of observed DNA sequence data from multiple loci, it is desired to obtain an

© Copyright, Princeton University Press. No part of this book may be distributed, posted, or reproduced in any form by digital or mechanical means without prior written permission of the publisher.

estimate of the species tree. Although gene trees are not directly observed, it is clear that they play an important role in the data-generation mechanism. For this reason, methods for estimating species trees are commonly categorized according to how they account for uncertainty in the gene trees in carrying out inference.

One class of methods for species tree inference is referred to as *summary statistics methods* or *summary methods* because these methods carry out species tree inference in two distinct steps, the first of which represents a summarization of the data. In this first step, a gene tree is estimated for each locus in the data set using one of the standard methods for phylogenetic tree estimation (e.g., maximum likelihood). The gene trees estimated in this first step are then used as input to the second step of the procedure, and a species tree estimate is obtained using only the information contained in these input gene trees. Such methods have the advantage of being computationally efficient. In the first step, the gene trees for the individual loci can be estimated in parallel, as each depends only on the sequence alignment for that gene under the conditional independence assumption mentioned in the previous paragraph. The second step is typically carried out by assuming some model for the relationship between the gene trees and the species tree. The features of the models that are commonly used for inference typically lead to computationally tractable algorithms for inference. The drawback of summary methods is that uncertainty in the gene tree estimates is typically unaccounted for in the second step of the procedure, making estimation of the uncertainty in the species-level phylogenetic estimate difficult to quantify. Though some suggestions have been made to remedy this (e.g., using as input bootstrap samples of gene trees rather than only a single point estimate of the gene trees), these strategies have not definitively been shown to improve the overall inference.

The second class of methods for species tree inference is referred to as *coestimation methods*, because they jointly estimate gene trees and the species tree under probabilistic models. The current methods within this class employ a Bayesian framework and use Markov chain Monte Carlo (MCMC) to carry out inference. Such methods have the advantages that fairly complex models can be fit and that estimates of all model parameters and associated measures of uncertainty are naturally obtained as part of the MCMC inference procedure. However, these methods can be computationally intensive, particularly as the size of the data and the complexity of the model increase. Most current methods cannot be feasibly run on genome-scale data for more than 20 or so species, but as computational power continues to grow, the ranges of data set sizes and models that can be successfully analyzed in this framework will continue to expand.

A third class of methods are those based on site pattern frequencies, which generally involve the use of genome-scale data to compare data features expected under various evolutionary models to those found in observed data sets. Importantly, site-pattern-based methods are distinct from summary methods as they are applied directly to the sequence data with the need to estimate gene trees first. Such methods have been shown to be computationally feasible for large data sets and thus show promise for carrying out efficient inference for large data sets for which analysis in the Bayesian coestimation framework is computationally prohibitive.

Across all of these classes of methods, the most common model assumed for the relationship between gene trees and the species tree is the coalescent process; however, other processes have been considered. For example, in the class of summary statistics methods, a model of duplication and loss can be used to carry out inference under the parsimony criterion by selecting the species tree that minimizes the number of gene duplication or loss events required to explain the set of estimated gene trees.

For general queries, contact webmaster@press.princeton.edu

© Copyright, Princeton University Press. No part of this book may be distributed, posted, or reproduced in any form by digital or mechanical means without prior written permission of the publisher.

Of course, important issues arise, such as whether duplications and losses should be weighted equally and how the space of possible species trees should be searched to find the tree that minimizes the number of duplications and losses. A challenge for the future will be the development of methods that can simultaneously account for many of the evolutionary processes known to generate gene tree–species tree discord in a computationally feasible manner.

1.3.1 CONTROVERSIES IN THE ESTIMATION OF SPECIES TREES

Careful testing that employs both simulated and empirical data and that includes comparisons between methods are important components of the development of methods for the problem of species tree inference, and work in this area has predictably led to disagreements among researchers about the appropriateness of various methods. One source of controversy has involved the utility of the multispecies coalescent model in improving the accuracy of estimated species-level phylogenies. Two observations have contributed to this viewpoint. First, when the possibility of discord in the gene trees is ignored and the data from all of the sampled loci are concatenated and analyzed with a method for gene tree inference, such as maximum likelihood, the resulting tree is often an accurate estimate of the species tree topology, particularly when the length of time between speciation events is large. The second, and related, observation is that incomplete lineage sorting is most commonly observed in empirical data for recent speciation events that have occurred in quick succession (e.g., species are often not monophyletic when multiple individuals are sequenced), which has led to the speculation that incomplete lineage sorting, and therefore the need for methods that explicitly model the coalescent process, is not relevant for nodes "deeper" in a tree. However, the coalescent model predicts that the amount of incomplete lineage sorting depends upon the length of time between speciation events, regardless of whether speciation was recent or occurred deep in the past (i.e., regardless of the depth in the species tree). Thus, it is not reasonable to conclude that the possibility of incomplete lineage sorting can be disregarded in the inference stage for deeper histories without knowledge about the time separating speciation events.

In regard to the observation that concatenated methods often infer the species tree topology with high accuracy, a more nuanced view is required to relate this observation to the performance of such methods. First, concatenating multilocus data and carrying out, for example, a maximum likelihood analysis on the resulting concatenated alignment makes the assumption that all loci share a single underlying phylogenetic tree from which data evolve according to the nucleotide substitution model specified during the analysis. Yet several empirical examinations (e.g., [134, 630, 213]) have established that the phylogenetic history varies across the phylogeny in a way that is often consistent with the multispecies coalescent. While one might argue that phylogenetic inference commonly requires simplifying assumptions (for example, methods for estimating gene trees from single-locus alignments typically assume that sites evolve independently), these assumptions are generally necessary because computationally efficient methods that implement more appropriate models are lacking. The variety of computationally tractable methods described within this book demonstrates that this is not a limiting factor in the case of modeling gene tree discordance.

Second, and more importantly, while the topology may be estimated fairly accurately, other important quantities may not be. One example is the quantification of uncertainty in the estimated tree, which is typically carried out by bootstrapping when a method like

© Copyright, Princeton University Press. No part of this book may be distributed, posted, or reproduced in any form by digital or mechanical means without prior written permission of the publisher.

maximum likelihood is used to analyze the concatenated alignment. Two issues arise when using the bootstrap on the concatenated alignment. First, the bootstrap assesses uncertainty in the repeated application of a particular method to data from the population of interest and thus cannot address uncertainty that results from inaccurate assumptions in the model used [222]. Given that concatenation represents an incorrect modeling assumption (i.e., that all loci share a common gene tree), the bootstrap support values are difficult to interpret. Second, bootstrap support from large concatenated alignments will tend to overestimate the actual support for a node. For example, consider a node supported by 55% of the sites in a data set while 45% of the sites favor some other arrangement, and suppose that the data contain one million sites (thus, 550,000 bp favor the node of interest, while 450,000 bp favor the other relationship). Most bootstrap samples will contain a majority of sites favoring the node of interest, and thus the bootstrap support for the node under consideration will be near 100%. Yet, the data show much more even support for the two alternative relationships. For example, significant underlying conflict was masked by high support values of concatenated analyses when conflicting phylogenetic relationships were actually strongly supported (e.g., [675, 774, 426]).

Finally, species divergence times (or speciation times) will be inaccurately estimated using a concatenation method as such a method assumes that the common ancestor for all loci in the concatenated alignment was identical. However, under a model such as the coalescent, it is clear that the gene divergence events (i.e., common ancestor times for the individual loci) must all predate the time of the speciation event in the absence of gene flow or some other form of horizontal transfer (see figure 1.1). Use of a coalescent model allows estimation of the speciation time after accounting for this, while concatenation methods do not. An example of the potential consequences of this is given in [390] (their table 5) for an empirical data set of *Sistrurus* rattlesnakes, for which the speciation times estimated for a concatenated versus coalescent-based analysis differ by as much as 70%. In addition, tree-associated parameters other than the speciation times, such as overall evolutionary rates, effective population sizes, and rates of gene flow, could potentially be affected by model misspecification when loci are concatenated.

1.4 A Look to the Future

1.4.1 CURRENT LIMITATIONS AND FUTURE PROSPECTS

The primary challenge facing current methods of species tree inference arises from the conflict between the desire to fit increasingly realistic, and therefore complex, models to the data and the computational resources required to fit such models as the size of data sets grows both in terms of taxa and genomic coverage. Overcoming this challenge will require new approaches to these problems, and these approaches will need to be designed specifically for the problem at hand. Here, we highlight several of the important issues to be addressed when the goal is to infer a species tree using phylogenomic data.

First, we note that new methods must be designed by carefully considering the data characteristics specific to phylogenomic data. In particular, properties of methods developed for inference of gene trees from single loci may need to be re-evaluated when applied to genome-scale data. As a first example, we note that when carrying out analyses of single loci, it is common to use a model selection procedure (e.g., ModelTest [584]) to choose a model of nucleotide substitution prior to inference of the gene tree. However, for a phylogenomic data set of several hundred or several thousand genes, selection of specific models for individual genes to be specified in the downstream species tree

© Copyright, Princeton University Press. No part of this book may be distributed, posted, or reproduced in any form by digital or mechanical means without prior written permission of the publisher.

inference procedure might yield little increase in the statistical power for tree inference at the expense of computational time in the model-fitting stage.

As a second example, consider the case of *phylogenetic invariants*, which were proposed in the late 1980s by [398], [124], and [125] as a possible method for inferring gene trees for samples of three or four taxa. Although promising from a theoretical standpoint, later work by [333] showed that invariants-based methods lacked power for gene tree inference and were outperformed by other methods in common use at the time. One explanation for the result of Huelsenbeck and Hillis is that phylogenetic invariants are formed from polynomials in the site pattern frequencies, where the site pattern frequencies are estimated from the data. For single loci that may be only a few hundred to a few thousand base pairs in length, accurate estimation of site pattern frequencies may be difficult, and polynomials formed from these estimates may have high variance, making them ineffective at differentiating among trees. For genome-scale data, however, a wealth of data are available, and estimates of site pattern frequencies are expected to be much more accurate, making invariants a reasonable tool for examining species-level phylogenetic relationships. The success of invariants-based methods, such as the ABBA-BABA test [208], in addressing complex problems (i.e., hybridization) indicates that the performance of such methods warrants new examination in light of the very different data structure provided by multiple loci.

A second challenge in developing methodology for species tree inference is that the availability of large quantities of sequence data provides the opportunity to use subsets of the data selectively to address specific questions of interest. Using only portions of the data can clearly result in increased precision in the resulting inferences (e.g., excluding loci with errors in alignment, assembly, or orthology detection; see [95]), but it also risks the introduction of bias when the data are not appropriately sampled. As described above, sequencing technologies in current use often result in large quantities of data, but the quality of the data for phylogenetic inference may differ. For example, some loci might be characterized by large amounts of missing data, or some loci may be involved either in a disproportionate amount of discord throughout the tree or discord in parts of the tree may arise by processes other than those captured by the phylogenetic model used for inference (e.g., [700, 95]). Depending on the method used for species tree inference, the investigator may be required to make a decision about whether to exclude taxa and/or loci because of the pattern of missing data (see, e.g., [638]). However, few methods currently incorporate a model of missing data explicitly in the inference procedure. Similarly, although data filtering is also becoming more common during the preprocessing of data in preparation for species tree inference, methods that explicitly incorporate the filtering step into formal inference model are lacking. The filtering may involve inclusion/exclusion of genes due to level of variation, the possibility of horizontal transfer or gene duplication, or numerous other reasons. However, when the process of sampling data is not correctly incorporated into the underlying models used for inference, the inferred phylogeny may be biased. It is unknown how much bias may be introduced by the nonrandom inclusion of data, and there is as yet no consensus on how filtering decisions should be made [371].

1.4.2 BEYOND THE SPECIES TREE

Even though a major challenge to species tree resolution is the gap between the data we collect for phylogenetic analyses (i.e., large-scale transcriptomic and genomic data) and the methods that accommodate the inherent complexity of big data (i.e., processes in addition to incomplete lineage sorting that contribute to discord among loci), such

© Copyright, Princeton University Press. No part of this book may be distributed, posted, or reproduced in any form by digital or mechanical means without prior written permission of the publisher.

data is also an unprecedented opportunity to better illuminate the processes that shape the tree of life. That is, big data and all its complexities when studied in a phylogenetic framework open new opportunities to address questions beyond the primary goal of resolving the species tree.

The characterization of patterns of discordance and the contribution of different processes to this discordance is itself of interest for generating hypotheses about the role of lateral gene transfer, gene duplication, and incomplete lineage sorting during the divergence of different taxa [627, 697]. Such hypotheses include those about the distribution of duplication across the tree of life and its potential association with the shifts in diversification rates or its concentration at the origin of major clades (as opposed to being dispersed across taxa). Likewise, by focusing on the processes that lead to gene tree discord, we can test whether lateral gene transfer is commonly associated with hypothesized ecological transitions versus evidence of convergent molecular evolution (e.g., [249]).

In the future, and as comparative genomics expands [425], the species tree and all its ancillary applications will only grow (e.g., dissecting evolution and disease; [491]). Such novel biological questions depend upon seeking evolutionary explanations for the distribution of discord in the gene trees, which means that the heterogeneity of comparative phylogenetic data sets should be embraced (as it is arguably the key to accurate resolution of recalcitrant phylogenetic relationships).

1.5 Organization of This Book

In the book we highlight, by example, not only how species tree estimation differs from a phylogeny estimated from concatenated multilocus data but also the issues that arise more generally from a mismatch between phylogenomic data, with all its complexities, and the models used for inference. This includes both conceptual and practical issues related to improving species tree estimates, as well as inferences that are nonbifurcating (i.e., networks). The book devotes five chapters to methodological developments, whereas the latter portion of the book, a total of eight chapters, focuses on empirical applications, including those that consider questions beyond the species tree. Some of the contributors were participants in a workshop offered at the 2018 Society of Systematic Biology Standalone meeting, whereas others were invited to cover topics identified by workshop attendees from questionaires they completed to assure that the book reflects the experiences, interests, and concerns of the diverse community that is engaged in species tree inference.

Through the set of chapters (authors representing their own perspective on aspects of species tree estimation relevant to their individual research programs), a diversity of perspectives and backgrounds are represented. This diversity means that the book speaks to people with varying levels of familiarity with the topic of species tree estimation, but it does not (nor is it intended to) provide a comprehensive overview of the subject. The combination of theoretical and empirical work is meant to provide readers with a level of knowledge of both the advances and limitations of species tree inference that can guide researchers in applying the methods while also inspiring future advances among those researchers with an interest in methodological development. Such cross talk (between empiricists and theoreticians/mathematicians) is vital to the growth of phylogenomics as it refocuses attention on the biological history of diversification (i.e., the timing and pattern of species divergence), the processes generating the observed patterns of genetic variation (e.g., sorting of ancestral polymorphism and gene flow, in addition to mutation models of nucleotide evolution), and the vast opportunities of study that includes and goes beyond a focus on the species tree itself.

For general queries, contact webmaster@press.princeton.edu

© Copyright, Princeton University Press. No part of this book may be distributed, posted, or reproduced in any form by digital or mechanical means without prior written permission of the publisher.

Index

ABBA-BABA test, 13, 122, 244; patterns, 168; statistics, 88
accuracy: ASTRAL, 51, 52, 53–54; concatenation versus, 51
Akaike Information Criterion (AIC), 90
Amborella: data set for, 155–56; patterns of conflict, 152–54; placement of, 145
analytical developments, 15–17
Ancistrocladus, 153
angiosperms, gene tree conflict examples in, 152–54
Anomalopteryx didiformis (bush moa), 218
anomalous gene trees (AGT), gene flow and, 120–21
anthophyte hypothesis, 149
ASTRAL, 66–67; accuracy, 51, 53–54; algorithm, 46, 47–50; branch lengths in coalescent units, 61–64; branch support using local posterior probability (LOCALPP), 64–65; clusters, 49; coalescent-based tree method, 120, 121; constrained MQSST (maximum quartet support species tree), 47; constrained set, 50; divergence time estimates, 159, 160; dynamic programming, 23, 47–50; estimating gene trees, 29; follow-up analyses and visualization, 65–66; input to, 54–61; insights behind, 49; internal node in unrooted tree, 48; MQSST, 47; motivation and history, 46–47; multiple individuals, 50; output, 61–65; per-branch quartet frequencies, 62–63; phylogenomic subsampling, 219, 221; practical considerations, 54–61; published versions of, 47; quartet score, 61; running time, 54; simulation accuracy, 52; software package, 15, 16; species tree estimation, 21, 22–24, 40; species trees, 220; species tree topology, 61, 222; theoretical results related to, 50–51; weight calculation, 47–50
ASTRAL-II, species tree method, 219
ASTRAL-III, species network inference, 182, 183, 185, 186
ASTRAL-DISCO, 30
ASTRAL-MP: running time, 54, species tree estimation, 30–31
ASTRAL-Pro, 29–30
ASTRID: gene tree summary method, 24; software package, 15; species tree estimation, 21, 40
ASTRID-DISCO, 30
ATAC-seq, epigenetic approach, 230
average gene tree internode distance (AGID) matrix, ASTRID, 24
avian lineages, paleognaths, 216–18. *See also* birds
Avian Phylogenomics Project, 33

backbone networks, phylogenetic network, 111
BaliPhy, software, 225
Bayesian inference, phylogenetic network, 103–5
Bayesian Information Criterion (BIC), 90
Bayesian phylogenetics and phylogeography (BPP), 17, 240–41; species tree estimation, 40–41, 79
Bayes's rule, 64
BBCA, scalability to large numbers of loci, 25–26
BEAST, 158
*BEAST, 17, 37: coestimation method, 24–26, 45; gene tree heterogeneity and rate variation, 228; scaling to large numbers of loci, 25–26; species tree estimation, 21, 40
Big-O running time analysis, 41–42
birds: classes of noncoding nuclear markers, 219, 224; conserved nonexonic elements (CNEEs), 219, 220, 224, 229; introgression under uncertainty in, 244–45; introns, 219, 220, 224; lineages of, 216; paleognathous, as test case for post-genomic phylogenetics, 218; phylogenetics of, 211, 216–18; phylogenomic analysis of paleognathous, 231; phylogenomic subsampling, 219, 221; rate heterogeneity, 229; retroelement insertions, 221, 222; rooted triplets for emu + cassowary, kiwi and rhea lineages, 223, 224; species tree for paleognathous, 225; species trees versus concatenation methods, 219, 220; ultraconserved elements (UCEs), 219, 220, 224; whole-genome species tree for ancient radiation of, 218–23
bootstrap analysis, 28; calculation of, 178; hybrids in, 142; nonparametric bootstrap (NBS), 240–41; origin of gene flow, 142–43; resampling procedures, 219; species networks, 140–43; summarization of support for hybrid edges and hybrid nodes, 141–43; summarization of support for tree edges, 140, 141; tree edges, 141

For general queries, contact webmaster@press.princeton.edu

© Copyright, Princeton University Press. No part of this book may be distributed, posted, or reproduced in any form by digital or mechanical means without prior written permission of the publisher.

BPP. *See* Bayesian phylogenetics and phylogeography (BPP)

BUCKy: estimating concordance factors (CFs), 131, 134, 140, 143, 167, 181; method, 46; software, 175, 176; species tree estimation, 21

Caryophyllales, 145, 152, 153, 154

character models: assessment of sensitivity to errors, 194; complexities of empirical data in need of exploration, 209–10; illustration of species tree model, 193; linked versus unlinked, 191–92; methods, 192–94; project repository, 194; recommendations for using unlinked, 209; relevance to empirical data sets, 208; robustness to character-pattern errors, 207–8; simulations of error-free data sets, 192–93; site pattern errors, 193–94. *See also* ecoevolity; StarBEAST2

chromosomes, 212: accuracy of gene tree inference, 257; genetic linkage and spatial distribution of genealogies, 253; NDLR visualizations of trees from, 272; simulations for genealogical variation among, 252–56; spatial distribution of genealogies, 253

CloudForest, 261, 262, 263, 275

Cloudograms, 239, 241, 242

coalescent events, 3

coalescent histories, 5

coalescent independent sites (CIS), 74, 84, 87, 147; SVDQuartets, 69, 79–81

coestimation: species tree estimation, 45; species tree inference, 10

coevolution, 212; figs and wasp pollinators, 249; multiple genes and traits in *Jaltomata*, 248; oyster genes with bacterial virulence, 248; testing under phylogenetic heterogeneity, 249

communities of trees, 212–13

community detection methods, tree sets, 262–64, 273

comparative genomics: accommodation of uncertainty into whole-genome alignments, 225, 226, 227–28; detecting rate variation in genes and noncoding regions, 228–30; rooted triplets for emu + cassowary, kiwi and rhea lineages, 223, 224; screenshots of ratite genome browser, 226; species tree for paleognathous birds as foundation for, 225; study, 211–12; whole-genome species for ancient radiation of birds, 218–23

concatalescence, 255

concatenation: accuracy versus, 51; method, 80; sequences, 43; species tree estimation, 45

concordance factors (CFs), 122, 167; BUCKy for estimating, 131, 134, 140, 143, 167, 181

concordance scores, 241

conflicting gene histories, 149, 151

consensus species tree (CST) model, 235

consistency/consistent, 20, 31, 40–41, 89; ASTRAL, 50–51; divide-and-conquer methods, 35; multilocus species tree estimation, 36–37; SVDQuartets, 81

continuous-time Markov chain (CTMC), 193

convergent traits, 216

Ctenophora, 1

curvilinear components analysis (CCA), 261

DACTAL, divide-and-conquer using supertree method, 34, 39

Darwin, Charles, 216, 233

DendroPy, Python package, 193

Densitree program, 241

developments, analytical and methodological, 15–17

Dirichlet distribution, 123

discordance, 234; comparison of incomparable, 238; concepts of harmony and, 234–36; history of systems and methods, 232–34; mereocentric stance, 237–38; origin of, 232–38; per branch quartet support, 65–66; quantification of, 238–40; relaxed speciocentric stance, 237–38; species tree, 236–38; strict speciocentric stance, 236, 237; visualization of, 238–40; visualization of conflict, 241–43

disjoint tree merger methods, divide-and-conquer, 34–36

displayed trees: phylogenetic network, 111; subnetworks and, 128, 129

DisvoVista, 66

divergence time: accuracy and precision of estimates, 195, 196, 198, 199; estimates for simulated data, 159; gene tree conflict, 157, 158–60

divide-and-conquer: analysis of larger data sets, 109, 110; disjoint tree merger methods, 34–36; species tree estimation, 33–36; supertree methods, 34

DNA amplification, polymerase chain reaction (PCR), 192

Drosera, 153

Drosophyllaceae, 154

Drosophyllum, 153–54

D-statistics, 168, 244; determining exact hybrid relationships, 172–73; identification of hybrid relationships, 170, 171; introgression under uncertainty in birds, 244; testing of hybridization, 168

ecoevolity: accuracy and precision of divergence time estimates, 196, 198, 199; accuracy and precision of effective size of descendant populations, 195, 203, 204, 205; accuracy and precision of root effective population size, 195, 200, 201, 202; analysis of all sites versus SNPs with, 195, 197; assessing sensitivity to errors, 194; behavior of StarBEAST2 (linked) versus unlinked, character models, 195; coverage of credible intervals, 197; data sets simulated with unlinked characters, 195, 206; empirical data in need of exploration, 209–10; MCMC convergence and mixing, 197; recommendations for using character models, 209; relevance to empirical data sets, 208; robustness to character-pattern errors, 207–8

empirical inference, 145–47

ENCODE project, 215

Enlightenment, 233

Estimating Species Trees (Knowles and Kubatko), 1, 251

© Copyright, Princeton University Press. No part of this book may be distributed, posted, or reproduced in any form by digital or mechanical means without prior written permission of the publisher.

eudicots, 152, 153
evolution, 213; genomic, 216–17; sensitivity to models of sequence, 264–68
evolutionary biology, 216, 231
ExaML: maximum likelihood, 32, 33; phylogenomic subsampling, 219, 221; species tree, 219; species trees, 220; species tree topology, 222
explicit networks, 126–27
extended Newick format, networks, 127–28
extended parenthetical format, networks, 127–28

FastMulRFS, 29
FastTree, 57
Fermi, Enrico, 119n6
flattening matrices: rank reductions for, 71–73; SVDQuartets, 69–70
fragmentary data, 60

genealogies: linked, and gene tree inference, 256–58; Robinson-Foulds (RF) distance between, 255, 256
gene(s): coevolution of oyster, with bacterial virulence, 248; coevolution of traits and, in *Jaltomata*, 248; duplication, 13; evolution model, 44; expression and epigenetics, 230–31; gene tree conflict, 157–58; information, 3
gene duplication and loss (GDL), 3, 19, 20, 213; gene tree conflict, 149, 151; gene tree heterogeneity, 37; PhyloNet, 118; process of, 4; species tree estimation under, 29–30
gene flow, 3, 4, 146; four-taxon network, 121; horizontal gene transfer (HGT), 126; impact of, 120–22; origin of, 142–43
general heterogeneous evolution on a single topology (GHOST), 23
generalized time reversible (GTR) model, 19, 193; sequence evolution model, 265, 266, 267, 268; SVDQuartets, 70
gene-to-tree map, concept of, 175
gene tree(s), 2, 3, 213; calculation of uncertainty, 178; filtering of entire, 58–60; filtering of leaves from, 58; heterogeneity and detecting rate variation, 228–30; joint versus independent inference of, 268–71; understanding, 43, 46; visualizations of variation, 242
gene tree conflict, 149–52; causes of, 149, 151; divergence time analyses, 157, 158–60; examples within angiosperms, 152–54; experimental data sets, 169; gene and genome duplication, 157–58; illustration of consequences of, 157; inference of species trees, 154–56; species tree reconstruction, 157
gene tree estimation, 9, 33, 66, 212; accuracy of, along chromosome, 257, 258; ASTRAL, 55, 57; gene tree uncertainty, 55, 57; inference tools and models, 57; linked genealogies and, 256–58; methods for, 11, 13; nonlinear dimensionality reduction (NLDR) visualization, 269; sliding window analyses, 252, 257, 259; units for, 2
gene tree estimation error (GTEE), 23, 28–29, 40, 222–23; gene tree conflict, 151, 155; GTR+GAMMA model, 23; sensitivity to models of sequence evolution, 264–65
gene tree heterogeneity: multiple biological sources of, 38–39; phylogeny and, 217
gene tree inference. *See* gene tree estimation
gene tree space, nonlinear dimensionality reduction (NLDR), 271
genomes, 215; evolution, 43; hierarchical model, 44; regulatory network, 215; understanding variation across, 271–75
genomics, 215; integration of comparative and functional, 216
genotype-to-phenotype map, 215
genotyping by sequencing (GBS), 8
GHOST. *See* general heterogeneous evolution on a single topology (GHOST)
Gnetales, anthophyte hypothesis, 149
goodness of fit tools: choice of number of hybridizations, 139; concordance factor (CF), 122–24; visualization of comparisons between observed and expected CF, 139–40
guenomu, Bayesian supertree method, 40, 41
guide tree merger (GTM), disjoint tree merger method, 35, 36

harmony, concepts of discordance and, 234–35
Hasegawa-Kishino-Yano (HKY), sequence evolution model, 265, 266, 267, 268
Heliconius butterflies, 271
heterogeneity: phylogenetic, 233–34. *See also* phylogenetic heterogeneity
Homo sapiens, 78
Hoolock leuconedys, 78, 79
horizontal gene transfer (HGT), 19; gene flow, 126; gene tree conflict, 149, 151; gene tree heterogeneity, 37
horizontal transfer, 3, 4, 13
Humiles subsection, 146. See also *Penstemon*
hybridization, 3, 4, 146; approaches for studying, 161–63; bootstrap support for hybrid edges, 141–43; bootstrap support for hybrid nodes, 141–43; challenges in determining exact relationships, 172–73; detection of patterns of, 167; experimental design, 165, 166; gene tree conflict, 149, 151; gene tree discordance, 172; hybrid enrichment, 6; hybrid speciation and, 112, 113, 120; impact on evolution and speciation, 161; introgression and, 212; in *Iochrominae*, 163, 173–74; known and hypothesized hybrid relationships, 164; study of *Iochroma*, 165; study system, 163–64; target capture and assembly, 166–67; testing in empirical data sets, 168; testing under phylogenetic heterogeneity, 243–45; tests supporting relationships, 170–72. See also *Iochroma*; *Penstemon*
hybrid speciation, 3, 4, 19, 40, 89; hybridization and, 112, 113, 120

© Copyright, Princeton University Press. No part of this book may be distributed, posted, or reproduced in any form by digital or mechanical means without prior written permission of the publisher.

HyDe, 243; determining exact hybrid relationships, 172–73; hybridization tests with, 188; identification of hybrid relationships, 170, 171; test, 122

Hylobates moloch, 78

Hylobates pileatus, 78

IC/TC. *See* internode and tree certainty (IC/TC)

implicit networks, 126–27

incomplete lineage sorting (ILS), 19, 20, 43, 89, 175, 216–17; accounting, in SVDQuartets, 74–75; empirical performance, 37–38; gene flow and, 120; gene tree conflict, 149, 151; gene tree heterogeneity, 37; SPILS (substitutions produced by ILS), 227, 228; whole genomes, 227

incongruence, 234

inconsistency, with ASTRAL, 51

inheritance probabilities, 91–92; branch lengths and, 100, 130, 133, 136, 143; definition, 119n1; vector of, 125, 127

internode and tree certainty (IC/TC), 146, 241; testing of hybridization, 167, 168

introgression, 4, 212; postspeciation, 243; testing under phylogenetic heterogeneity, 243–45; uncertainty in birds, 244–45

Iochroma: accessions for study of, 165; challenges in determining hybrid relationships, 172–73; decreasing tree-like signal, 168–70; detection of patterns of hybridization from gene tree distributions, 167; experimental design, 165, 166; gene tree conflict and signal across experimental data, 169; genus as model clade, 162; hybridization and gene tree discordance, 172; hybridization in *Iochrominae*, 163, 173–74; hybrid taxa increasing discordance, 168–70; identification of hybrid relationships, 171; study system, 163–64, 165; target capture and assembly, 166–67; testing of hybridization in empirical data sets, 168; tests of hybridization support, 170–72

Iochroma arborescens, 163, 164, 165, 166, 170, 171

Iochroma ayabacense, 163, 164, 165, 166, 170, 171

Iochroma confertiflorum, 163, 164

Iochroma cornifolium, 163, 164

Iochroma cyaneum, 163, 164, 165, 166, 170, 171

Iochroma lehmannii, 163, 164, 165, 166, 170, 171

Iochroma stenanthum, 163, 164, 165, 166, 171

Iochrominae, 146

IQ-Tree, 230

Jaltomata (Solanaceae), coevolution of multiple genes and traits, 248

Jules-Cantor (JC), sequence evolution model, 265, 266, 267, 268

Julia programming language, PhyloNetworks, 133, 143–44

jungles, 249

Kingman's coalescent, 2, 4–5

Lamarck, Jean-Baptiste, 232, 233

LAPACK Fortran library, 76

Leucanthemopsis, reticulate evolutionary history, 115–17

linked genealogies, 252, 253–56; gene tree inference and, 256–58

local posterior probability (LOCALPP), 28, 64–65

Locke, John, 233

McDade, Lucinda, 162

major trees, 112, 129, 140

Markov chain Monte Carlo (MCMC), 68; analysis, 270; assessing sensitivity to errors, 194; Bayesian, 43, 95; Bayesian coestimation, 24; heuristic search versus MCMC sampling, 92–93, 94; inference procedure, 10; posterior probability distributions, 17

maximum a posteriori (MAP), 79, 119n2; speciation time estimation, 85–86; tree sets, 266, 267, 268

maximum likelihood (ML), 43; analysis, 11–12; estimation, 90; ExaML, 32; methods, 20; multilocus species tree estimation, 31–33; phylogenetic network inference under, 98–102; phylogeny estimation, 20; RAxML, 32; recent developments, 32–33

maximum pseudolikelihood estimation: analysis of larger data sets, 106–7; flow chart of procedure estimating network, 131; identifiability of model, 134–36; phylogenetic network, 102–3, 104, 105; rooted four-taxon network and semidirected version, 132; species networks, 130–36

maximum pseudolikelihood inference. *See* maximum pseudolikelihood estimation

maximum quartet support species tree (MQSST), 47, 49–50

mereotic trees (MTs), 235; genome-wide sets of, 235–36; heterogeneity, 249

METAL, site-based method, 21

minimizing deep coalescence (MDC): inference under MDC criterion, 96–98, 99; phylogenetic network, 89

missing genes, 60

model selection procedure, 12–13

morphology, plant classifications, 149

Morrison, David, 113

MP-EST: phylogenomic subsampling, 219, 221; species tree method, 219; species trees, 220; species tree topology, 222

MrBayes, 52, 53, 57, 131, 143

multilocus bootstrapping (MLBS), 55, 57; branch support, 64–65

multispecies coalescent (MSC), 43, 251–52; coalescent simulations, 252–56; introduction to, 5–6; model, 175, 251, 260, 268; modeling incomplete lineage sorting (ILS), 20; phylogenetic linkage (PL), 255–56; simulations examining distribution of linked versus unlinked genealogies on species tree, 254; species tree inference, 251; tree size, tree space, and phylogenetic decay, 255–56; units, space, and time, 253–55

multispecies network coalescent (MSNC), phylogenetic network, 90

© Copyright, Princeton University Press. No part of this book may be distributed, posted, or reproduced in any form by digital or mechanical means without prior written permission of the publisher.

MULTIZ, aligning whole genomes, 225
mutation units, 2

Nepenthes, 153
networks: rooting of semidirected, 136–39; species, 124–29; trees versus, 122–24. *See also* species networks
next-generation sequencing (NGS), 150–51
NJMerge, disjoint tree merger (DTM) method, 35, 36
NJst, coalescent-based tree method, 120
Nomascus leucogenys, 78, 79
nonlinear dimensionality reduction (NLDR), 261, 269; tree set visualization, 261, 269, 270, 271, 275
nonparametric bootstrap (NBS), 240–41. *See also* bootstrap analysis
Nymphaeales, 152, 154, 156
Nyphaea colorata, 152

omni-gene hypothesis, 247
On the Origin of Species (Darwin), 233
oyster genes, coevolution of, 248

pairwise disjoint, 34
paleognathous birds: phylogenomic analysis, 216–18, 231. *See also* birds
PAML software, 229, 245, 247
parallel microfluidic PCR + HTS, 8
ParGenes, 33
partitioned coalescence support (PCS), 60
PASTA, software, 38, 225
PAUP*, 27, 28, 68; SVDQuartets, 87; SVDQuartets for quartet sampling and assembly, 75–76
Penstemon, 146; approach to study, 176–79; best maximum pseudolikelihood (ML) networks by SNaQ, 187; bootstrapping, 178; calculation of QCFs, 177–78; candidate hybridization events from rooted triples, 181–82; character evolution and biogeography, 189; gene tree uncertainty, 178; hybridization discussion, 186–90; hypothesis of allopolyploid formation in *P. attenuatus*, 180; materials and methods, 179–82; nuclear amplicon data, 182; phylogenetics of hybrids and polyploids, 189–90; phylogeny of subsections by ASTRAL-III, 183; phylogeny of subsections by qcf and QuartetMaxCut, 184; phylogeny of subsections by RAxML, 185; sample collection, DNA extraction and amplicon sequencing, 180–81; species network inference, 182, 186; species tree inference, 181, 182, 185–86; study system, 179–80; subsections *Humiles* and *Proceri*, 176; taxonomy of subsections *Humiles* and *Proceri*, 188–89; tests for hybridization, 186; validation of QCF estimation, 178–79
PHAST program, 228
PHYLDOG, 29, 155
PhyloAcc, 228–29
PhyloG2P paradigm, 215
phylogenetic(s), 1; birds, 211, 216–18; gene expression and epigenetics, 230–31; genome-scale inference, 212; invariants, 13; paleognathous birds as test case for post-genomic, 218; phylogenetic linkage (PL), 255–56; rooted triplets for emu + cassowary, kiwi and rhea lineages, 223, 224; technologies for relationships, 215–16; whole-genome species for ancient radiation of birds, 218–23
phylogenetic analysis, 232, 250; approaches for studying hybridization, 161–63; gene expression and epigenetics, 230–31; gene flow, 120. *See also* phylogenetic heterogeneity; phylogeny
phylogenetic heterogeneity: analysis under, 243–50; coevolution of figs and wasp pollinators, 249–50; coevolution of multiple genes and traits in *Jaltomata*, 248; coevolution of oyster genes with bacterial virulence, 248; gene tree distributions, 239; introgression under uncertainty in birds, 244–45; quantification and visualization of discordance, 238–40; quantification of conflict and tree evaluation, 240–41; selection under heterogeneity in wild tomatoes, 246–47; testing introgression and hybridization, 243–45; testing of coevolution under, 249; testing of selection under, 245–48; testing of traits under, 247–48; theme of studying, 212; visualization of conflict, 241–43
phylogenetic network analysis, 89–90, 213; heuristic search versus MCMC sampling, 93, 94; isomorphic, 113; parameters and identifiability, 92, 93; reading and interpreting, 91–92. *See also* PhyloNet
phylogenomic(s), 14; consequences of gene tree conflict in, 154–60; data types and technologies for generating, 6, 9; discordance, 234; filtering of data, 57–61; goal of analysis, 258–59; pipeline, 56, species tree estimation, 55, 56; studies of per branch quartet support, 65–66; subsampling, 219, 221; SVDQuartets, 81–82. *See also* gene tree conflict
phylogeny, 55, 99, 120, 145, 231, 235–36, 250; of birds, 216, 218–19, 229; branch tests, 246; coalescent theory and, 215–16; consensus, with node concordance, 242; cophylogeny, 250; data generation for species-level, 74; fixed-species, 82–83; flowering plant, 152, 153, 158; gene tree heterogeneity and, 217; hybridization across, 161, 163, 167, 174; introgression, 243, 244–45; of New World kingsnakes, 189; of *Penstemon*, 179, 180, 181, 183, 184, 185, 190; sequence alignment and, 225–26; speciation, 242; species tree estimation, 211–13, 236, 239; traits, 247–48; unlinked character models, 209
PhyloNet, 40, 176, 190, 244; analysis of larger data sets, 106–11; analysis of polyploids, 114–17; ancestral recombination graph, 113; backbone networks, 111; Bayesian inference, 103–5; comparison and summarizing networks, 111–12; computational requirements for methods other than MCMC_SEQ, 107, 108–9; displayed trees, 111; divide-and-conquer approach, 109, 110; hill climbing, 95–96; illustrating inference methods in, 96–106; inference under MDC criterion, 96–98, 99; major trees, 112; Markov chain Monte Carlo (MCMC), 92–93, 94; maximum likelihood inference, 98–102; maximum pseudolikelihood inference, 102–3, 104, 105, 106–7;

© Copyright, Princeton University Press. No part of this book may be distributed, posted, or reproduced in any form by digital or mechanical means without prior written permission of the publisher.

PhyloNet (cont.)
optimization problems, 92; reticulate evolutionary processes in, 112–17; reticulation, 90; running time, 105–6, 107; software, 16–17, 117–19; tree-based augmentation, 107, 109; tree decompositions, 112; tripartitions, 112

PhyloNetworks, 146, 176; Julia programming language, 133, 143–44; main functions in, 143–44

Physalis peruviana, 165, 166. *See also* hybridization; *Iochroma*

Plant and Fungal Tree of Life Project (PAFTOL), 150

plant classification, 149–50

plant life, resolution of tree of, 160

polymerase chain reaction (PCR), 7, 8; DNA amplification, 192, 208; genetic data, 150; microfluidic, 180

polyploids: analysis of, 114–17; phylogenetics of hybrids and, 189–90

polyploidy, 146, 180; events, 114, 117; hybridization and, 190. See also *Penstemon*

polytomies, 48, 50, 151; ASTRAL for, 65; tests for, 65

PoMo method, 41

Potts model, 263; community detection, 273

primary concordance tree (PCT), 175

probability distribution: branch lengths in gene tree histories, 74; gene tree topologies, 4, 5, 95; gene trees, 90; posterior, 17; site pattern, 69; SVDQuartets, 80

Proceri subsection, 146. See also *Penstemon*

ProgressiveCactus, 225, 227

pseudolikelihood, 124; composite likelihood, 84; gene trees, 92; measure, 90; methods, 129–30; model, 134, 139. *See also* maximum pseudolikelihood estimation; maximum pseudolikelihood inference

Python: DendroPy, 193; *ipcoal* package, 252; *msprime* program, 252; *scipy* package in, 256; *toytree* package, 252

QCF. *See* quartet concordance factors (QCF)

quartet concordance factors (QCFs), 146; calculation of, 177–78; SNaQ (species networks applying quartets) using, 176; validation of QCF estimation, 178–79

Quartet MaxCut, 133; species network inference, 182, 184, 185, 186; taxonomy of *Penstemon* subsections, 188

random genomic sequencing, 6

RAxML, 57, 167; gene tree inference, 257; introgression under uncertainty in birds, 244; maximum likelihood, 32, 33; phylogeny of *Penstemon* subsections with, 185; species tree inference, 181, 185; species tree method, 219

Ray, John, 232–33

regulatory network, genomes, 215

resolution of tree of life, 160

restriction site-associated DNA sequencing (RADseq), 6, 8, 9, 39, 80, 146–47

reticulate evolution, 112; ancestral recombination graph, 113; inference results for *Leucanthemopsis* representatives, 116–17, 118; *Leucanthemopsis* test data, 115; processes in PhyloNet, 112–17

reticulation: ASTRAL, 51; direction of, 136–37; PhyloNet, 90; reticulation edges, 91; reticulation nodes, 91

revPoMo method, 40–41

Robinson-Foulds distance: between genealogies, 255–56; between true and estimated gene trees, 51, 53; computing, 128–29; to guide tree, 35; topological discordance, 249; weighted, 265

Robinson-Foulds supertree (RFS) problem, 29; FastMulRFS, 29

running time, 16, 21; ASTRAL, 22–24, 47, 49, 50, 54; ASTRID, 24; Bayesian coestimation method, 25; BBCA analysis, 25–26; Big-O analysis, 41–42; divide-and-conquer approach, 34, 35; phylogenetic analysis, 105–6, 107, 110; PhyloNet, 90; species tree estimation, 28, 30, 31, 33; SVDQuartets, 27

selection: heterogeneity in wild tomatoes, 246–47; phylogenetic tests of, 245–47

sequence evolution model, 44

sequencing technologies, 6, 7, 13, 192, 215

Ship of Theseus, 234–35, 250

Simpson, George Gaylord, 236

simulations: accuracy of ASTRAL in, 52; genealogical variation among chromosomes, 252–56

single-nucleotide polymorphism (SNP), 146, 192; ABBA-BABA-tests for, 122 ; analysis with eco-evolity, 195–96, 198, 199, 200, 201, 202, 203, 204, 205; data, 9, 16, 27–28, 70; SVDQuartets, 78–80, 82

Singular Value Decomposition: SVDQuartets, 27; SVDquest, 27–28

siskins (*Spinus*), introgression under uncertainty in birds, 244–45

site pattern frequencies, 10, 13; HyDe tests, 181; quartet concordance factors (QCFs), 176; simplified model, 71, 72; species tree estimation, 68–70, 82–88; SVDQuartets, 80, 82

sliding window analyses, gene tree inference, 252, 257, 259

SNAPP, 16, 21, 244

SNaQ (species networks applying quartets): best networks estimated by, 137–38; choice of number of hybridizations, 139; determining exact hybrid relationships, 172–73; flow chart of procedure, 131; identification of hybrid relationships, 170, 171; maximum pseudolikelihood (ML) networks by, 187; pseudolikelihood method, 129–30; rooting of, 136–38; species network inference, 186; taxonomy of *Penstemon* subsections, 188–89; testing of hybridization, 168; using quartet concordance factors (QCFs), 176

Society of Systematic Biology, 14

Solanum sect. *Lycopersicon*, selection under heterogeneity in wild tomatoes, 246–47

© Copyright, Princeton University Press. No part of this book may be distributed, posted, or reproduced in any form by digital or mechanical means without prior written permission of the publisher.

speciations, 236
speciation times, 2, 12; algorithm details, 86; application to species relationships among gibbons, 87; composite likelihood, 84, 85; estimation of, 82–87; maximum a posteriori (MAP) estimation, 85–86; maximum composite likelihood estimates (MCLEs), 84; pseudolikelihood, 84; recommendations for composite likelihood estimators of, 87; symmetric species tree with, 83; theoretical basis, 83–86; uncertainty quantification, 86
species: concepts of, 236; divergence times, 12, 146, 228
species network(s), 124–29; bootstrap analysis, 140–43; candidate networks compatible with known outgroup, 137–39; comparison of, 128–29; displayed trees and subnetworks, 128, 129; explicit versus implicit, 126–27; extended Newick format, 127, 128; extended parenthetical format, 127–28; fast reconstruction of, 129–43; goodness of fit tools, 139–40; identifiability of pseudolikelihood model, 134–36; level-1 and level-2 networks, 126; maximum pseudolikelihood estimation, 130–36; notation for, 124; rooted and semidirected phylogenetic networks, 124, 125; rooted four-taxon network and semidirected version, 130, 132; rooting of semidirected networks, 136–39; SNaQ (species networks applying quartets), 129–33; visual artifacts misleading interpretation, 126
species network inference: analysis of larger data sets, 106–11; divide-and-conquer, 109; heuristic searches versus MCMC (Markov chain Monte Carlo) sampling, 93, 94; identifiability of network's parameters, 93; illustration of methods in PhyloNet, 96–106; maximum likelihood inference, 98–102; maximum pseudolikelihood inference, 102–3, 104, 105, 106–7; MCMC simulations, 92–93, 95; MDC (minimizing deep coalescence) criterion, 96–98; phylogeny of *Penstemon* subsections using ASTRAL-III, 182, 183; phylogeny of *Penstemon* subsections using QuartetMaxCut, 182, 184; phylogeny of *Penstemon* subsections using RAxML, 182, 185; PhyloNet, 117–19; PhyloNetworks Julia package installation and use, 143–44; reading and interpreting a phylogenetic network, 91–92; reticulate evolutionary processes in PhyloNet, 112–17; summarizing and comparing networks, 111–12; tests for hybridization and, 186; tree-based augmentation, 107, 109; trees versus networks, 122–24. *See also* species network(s)
species phylogeny, 2, 3. *See also* phylogeny
species tree(s), 2, 3, 213, 236–38; concept of, 4; controversies in estimation of, 11–12; illustration of model, 193; reconstruction and gene tree conflict, 154–56, 157; three-axon, 5; topology, 253
species tree estimation, 1, 14, 19–21, 68–69; approaches to, 45; ASTRAL algorithm, 22–24; ASTRAP-MP, 30–31; ASTRID, 24; BBCA, 25–26; coestimation methods, 24–26; divide-and-conquer, 33–36; empirical data in need of exploration, 209–10; empirical performance for method choice, 37–39; evaluation of branch support in species trees, 28; ExaML, 32; future prospects for, 12–13; gene duplication and loss (GDL), 20, 29–30; incomplete lineage sorting (ILS), 20; introduction of site pattern errors, 193–94; maximum likelihood, 20; methods addressing ILS, 21–28; model-based methods of, 191–92; multilocus, using maximum likelihood, 31–33; multispecies coalescent (MSC), 20; overview of current methods for, 9–12; parallel implementations for, 30–33; phylogeny of *Penstemon* subsections with RAxML, 181, 185; quartet sampling and assembly, 75–76; RAxML, 32; recommendations for unlinked-character models, 209; sensitivity to models of sequence evolution, 264–68; simulations of error-free data sets, 192–93; site-based methods, 26–28; statistical consistency for method choice, 36–37; summary, challenges and future directions, 39–41; summary methods, 21–24; SVDQuartets, 27; SVDquest, 27–28; topology using SVDQuartets, 69–82. *See also* ecoevolity; StarBEAST2; SVDQuartets
Spinus (siskins), introgression under uncertainty in birds, 244–45
StarBEAST2: accuracy and precision of divergence time estimates, 196, 198, 199; accuracy and precision of effective size of descendant populations, 195, 203, 204, 205; accuracy and precision of root effective population size, 195, 200, 201, 202; assessing sensitivity to errors, 194; behavior of linked, versus ecoevolity (unlinked) character models, 195; coverage of credible intervals, 197; empirical data in need of exploration, 209–10; MCMC-based method, 16, 17; MCMC convergence and mixing, 197; recommendations for using character models, 209; robustness to character-pattern errors, 207–8; species tree estimation, 21, 25
STRUCTURE plots, 241
subnetworks, 119n4: displayed trees and, 128, 129; four-taxon, 130, 132, 139; three-taxon, 111
summary methods: species tree estimation, 21–24; species tree inference, 10
supertree methods, 53, 150; Bayesian, 41; divide-and-conquer, 34–35, 39; Robinson-Foulds (RFS), 29
SVDQuartets, 68; accounting of incomplete lineage sorting (ILS), 74–75; algorithmic details, 76; application to species relationships among gibbons, 78–79; assumptions and misconceptions, 79–81; data filtering, 81–82; data generation process for species-level phylogeny with three species, 74; estimation of speciation times, 82–87; estimation of species tree topology using, 69–82; flattening matrices, 69–70, 73; future work of, 87–88; introgression under uncertainty in birds, 244; method, 15, 16, 40; missing data, 81–82; possible site patterns under simplified model, 71, 72; properties of, 79–82;

© Copyright, Princeton University Press. No part of this book may be distributed, posted, or reproduced in any form by digital or mechanical means without prior written permission of the publisher.

SVDQuartets (cont.)
quartet sampling and assembly, 75–76; rank reductions for flattening matrices, 71–73; recommendations for using, 82; schematic of estimation procedure, 77; site-based methods, 21; species tree inference, 75–76; statistical consistency, 81; SVD-Score, 70, 73, 75, 76, 77; theoretical basis, 69–73; uncertainty quantification, 78
SVDquest, 27–28, 40
SVDScore, 70, 73, 75, 76, 77
Symphalangus syndactylus, 78, 79
system, 232, 250; history of, 232–34

Tanglegrams, 249
targeted enrichment, sequencing technology, 6, 7
targeted genotyping-by-sequencing (GBS), 6, 8
10kp (Ten Thousand Genomes Project), 150
terminology, 2–9
TICR (tree incongruence checking in R), 122, 124, 146, 167
tomatoes, selection under heterogeneity in wild, 246–47
traits: coevolution of genes and, in *Jaltomata*, 248; testing, under phylogenetic heterogeneity, 247–48
transcriptome sequencing technology, 6, 7
tree-based augmentation, analysis of larger data sets, 107, 109
TreeMerge, disjoint tree merger method, 35, 36
TreeScaper software, 213 261
tree sets, 212–13; bipartition covariance networks, 264, 266, 273–75; community detection methods, 262–64, 273; curvilinear components analysis (CCA), 261, 270; detection of structure in, 262–64; *erato-sara* clade, 272, 274; exploring, 260–64; joint versus independent inference of gene trees, 268–71; Jukes-Cantor (JC) model, 265, 266, 267, 268; maximum a posteriori (MAP), 266, 267, 268; nonlinear dimensionality reduction (NLDR), 261, 269, 270, 271, 275; prospects for future development and application, 275; sequence evolution models, 265, 266, 267, 268; topological affinity network with trees in distinct communities, 263; visualization of, 260–64
TreeSetViz module, 261
TreeShrink, 58, 59
tripartitions, 31: ASTRAL, 49; phylogenetic network, 111, 112; weight of, 47, 48

unlinked genealogies, 252, 253–56

Vibrio, 248; coevolution with oyster larvae, 248
von Neumann, John, 119n6

Wallace, Alfred Russel, 233
whole-genome sequencing, 6, 7, 15, 218

zones, 233