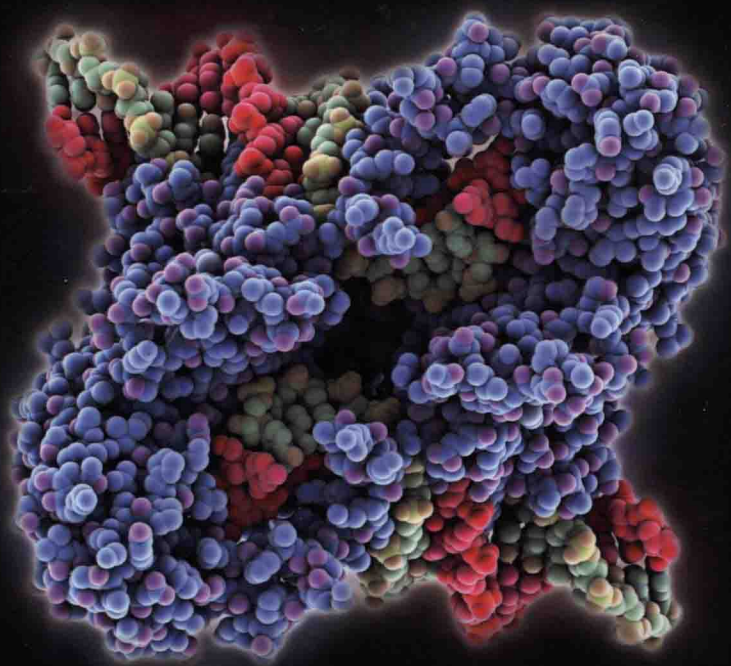
PHYLOGENOMIC DATA ACQUISITION

PRINCIPLES AND PRACTICE



W. BRYAN JENNINGS

CRC Press
Taylor & Francis Group

PHYLOGENOMIC DATA ACQUISITION

PRINCIPLES AND PRACTICE

W. BRYAN JENNINGS

CRC Press
Taylor & Francis Group
Boca Raton London New York

CRC Press is an imprint of the
Taylor & Francis Group, an **Informa** business

The cover image shows a computer model of a Tn5 synaptic complex, which is comprised of a Tn5 transposase enzyme (blue) bound to the ends of a DNA transposon (red and green). These enzymes play a vital role in some Next Generation Sequencing methods. (image credit: Laguna Design, Science Photo Library/Getty Images.)

CRC Press
Taylor & Francis Group
6000 Broken Sound Parkway NW, Suite 300
Boca Raton, FL 33487-2742

Printed on acid-free paper
Version Date: 20161021

International Standard Book Number-13: 978-1-4822-3534-0 (Hardback)

Library of Congress Cataloging-in-Publication Data

Names: Jennings, W. Bryan, author.
Title: Phylogenomic data acquisition : principles and practice / author, W. Bryan Jennings.
Description: Boca Raton : Taylor & Francis, 2017. | Includes bibliographical references.
Identifiers: LCCN 2016032138| ISBN 9781482235340 (hardback : alk. paper) | ISBN 9781482235357 (ebook)
Subjects: | MESH: Sequence Analysis, DNA--methods | Polymerase Chain Reaction--methods | Phylogeny | Biological Evolution | Data Collection
Classification: LCC QP624.5.D726 | NLM QU 500 | DDC 572.8/6--dc23
LC record available at https://lccn.loc.gov/2016032138

Visit the Taylor & Francis Web site at
http://www.taylorandfrancis.com

and the CRC Press Web site at
http://www.crcpress.com

PHYLOGENOMIC DATA ACQUISITION

PREFACE

Phylogenomics intersects and unites many areas in evolutionary biology including molecular and genomic evolution, systems biology, molecular systematics, phylogeography, conservation genetics, DNA barcoding, and others. Although these disciplines differ from each other in their study questions and methods of data analysis, they all use DNA sequence datasets. Phylogenomics is moving forward at a dizzying pace owing to advances in biotechnology, bioinformatics, and computers. This is reminiscent of what occurred two decades ago when the field of molecular systematics was coming of age. Another major factor that undoubtedly helped spur the growth of molecular systematics was the arrival of *Molecular Systematics*, 2nd edition (Hillis et al. 1996), a book that allowed me (along with countless others) to jump into this exciting field. As phylogenomics has grown substantially since the dawn of the genomics era—in large part due to the advent of next generation sequencing—the time is right for a book that presents the principles and practice of obtaining phylogenomic data.

This book enables beginners to quickly learn the essential concepts and methods of phylogenomic data acquisition so they can confidently and efficiently collect their own datasets. Directed at upper level undergraduate and graduate students, this book also benefits experienced researchers. The inference of gene trees from DNA sequence data represents one of the fundamental aspects of phylogenomic analysis. Accordingly, because robust gene tree inferences are generally made using longer DNA sequences (e.g., ~200–2,000 base pairs long), this book focuses on methods for obtaining sequences in this length range.

This book is organized as follows. Chapter 1 introduces phylogenomics within a historical context, points out connections between DNA sequence data and gene trees, discusses gene trees versus species trees, and provides an overview of the methods used today to acquire phylogenomic datasets. Chapter 2 describes the landscapes of eukaryotic genomes followed by discussion of molecular processes that govern the evolution of DNA sequences. Chapter 3 continues the discussion about properties of DNA sequence loci by reviewing six common assumptions that pertain to data characteristics before describing the different types of DNA sequence loci used in phylogenomic studies. Chapter 4 covers DNA extraction methods including high-throughput methods. Chapter 5 reviews PCR theory, discusses applications in phylogenomics, and considers high-throughput workflow. Chapter 6 describes Sanger sequencing including high-throughput sequencing. Chapter 7 explains Illumina sequencing technology and how it is used to obtain phylogenomic datasets. Chapter 8 reviews theory and methods for designing novel DNA sequence loci. Finally, Chapter 9 offers a vision of the future in phylogenomic data acquisition.

Most of the information contained in this book can be found elsewhere, but it is worthwhile to

bring it together. This synthesis provides detail including reference to the foundational papers. I hope these discussions will stimulate and direct the reader—especially students—to study these classic papers. Not only will these extra readings provide additional details about the subject at hand, but should also evoke feelings of admiration for those works and thereby generate inspiration and excitement about phylogenomics research. This book is biased toward eukaryotic organisms because of my research experience and interests in vertebrates. Therefore, an apology is in order to my colleagues who study microorganisms though they may still find at least some parts of this book useful.

I am grateful to a number of colleagues who reviewed earlier versions of chapters and provided helpful comments. For their help and encouragement I thank David Blackburn, Steve Donnellan, Andrew Gottscho, Fábio Raposo, Sean Reilly, Todd Schlenke, and especially Ryan Kerney who read four of the chapters. Any remaining errors are my own. I thank my editor Chuck Crumly at Taylor & Francis for suggesting that I undertake this project and for his constant encouragement and patience. I would also like to thank Cynthia Klivecka of Taylor & Francis as well as Mohamed Hameed and Karthick Parthasarathy of Novatechset for their tremendous help to produce the final book. Special thanks go to my doctoral advisor Eric Pianka who first encouraged me to apply molecular data to study evolutionary questions. I am indebted to all my mentors—Samuel Sweet, Jonathan Campbell, Eric Pianka, and Scott Edwards—who have helped me with my career and provided inspiration. I would also like to thank some colleagues from my graduate school days: Mark and Kris Holder, Tom Wilcox, Marty Badgett, and Todd Schlenke taught me many things in the laboratory and provided much assistance while David Hillis and James Bull generously allowed me to work in their laboratory. One of the pleasures of being a professor is having excellent graduate students who help to inspire me. I thus want to thank my students Andrew Gottscho, Sean Reilly, Igor Rodrigues da Costa, Carla Quijada, and Piero Ruschi. I would like to thank my Humboldt colleagues especially Jacob Varkey, Sharyn Marks, Brian Arbogast, Jeffrey White, John Reiss, Anthony Baker, Julie Davy, Leslie Vandermolen, and Christopher Callahan for their support of my teaching and research. I am also very grateful to my colleagues at the Federal University of Rio de Janeiro for their support and wonderful collaborations. In particular I wish to thank Ronaldo Fernandes, Paulo Buckup, Marcelo Weksler, Jose Pombal Jr, Paulo Passos, Ulisses Caramaschi, Marcos Raposo, Marcelo Britto, and Francisco Prosdocimi. This work was made possible through support from HHMI, Humboldt State University, Dean of Graduate Studies and Research (PR-2) at the Federal University of Rio de Janeiro, and the Brazilian funding agencies CAPES, CNPq, and FAPERJ. My wife Vivian Menezes Leandro provided me with constant support and encouragement, which enabled me complete this book. I dedicate this book to her.

W. BRYAN JENNINGS
Rio de Janeiro, Brazil
December 2016

AUTHOR

W. Bryan Jennings is a foreign visiting professor in the Department of Vertebrates and Post-Graduate Program in Zoology at the National Museum of Brazil and Federal University of Rio de Janeiro. He earned his BA in zoology from the University of California at Santa Barbara; MS in biology from the University of Texas at Arlington; and PhD in ecology, evolution, and behavior from the University of Texas at Austin. He was a postdoctoral fellow in the Department of Biology at the University of Washington, and in the Museum of Comparative Zoology and Department of Organismic and Evolutionary Biology at Harvard University. He was then appointed teaching fellow for one year in the Department of Molecular and Cellular Biology at Harvard before becoming an assistant professor of biology at Humboldt State University. At Humboldt, he taught genetics labs, bioinformatics, biogeography, introductory molecular biology, and introductory biology for nonbiology majors. In 2010, he moved to the National Museum of Brazil to accept a CAPES foreign visiting professorship. At the National Museum, he cofounded the Molecular Laboratory of Biodiversity Research, teaches a graduate course in phylogeography, and mentors masters students, doctoral students, and postdocs. Studies in his lab are focused on phylogenomics of vertebrates with an emphasis on phylogeographic and conservation genetics studies.

CONTENTS

CHAPTER ONE

Introduction

The great evolutionary geneticist, Theodosius Dobzhansky, famously wrote "*Nothing in biology makes sense except in the light of evolution*" (Dobzhansky 1973). One area in evolutionary biology that has shed much light on biological phenomena is the field of molecular phylogenetics. Phylogenetic trees inferred from molecular genetic data have led to quantum leaps in our understanding about molecular evolution and the *Tree of Life*. The Tree of Life Project is a worldwide collaboration of evolutionary biologists that aims to elucidate the evolutionary history for all life found on Earth (Maddison et al. 2007; http://tolweb.org/tree/). Another important initiative is the Open Tree of Life (http://opentreeoflife.org/). Advances in DNA sequencing capability, computers, and bioinformatics from the late 1970s through the 1990s spurred the rapid growth of molecular phylogenetics (Hillis et al. 1996). An outgrowth of this field, which began slowly in the 1990s but later blossomed into its own field due to the emergence and explosive growth of genomics, is the discipline of *phylogenomics*. Given that substantial overlap obviously exists between molecular phylogenetics and phylogenomics, we should ask the following questions: *What is phylogenomics and how does it differ from molecular phylogenetics?*

1.1 WHAT IS PHYLOGENOMICS?

Before we further consider a definition for phylogenomics, let's first examine a traditional definition of molecular phylogenetics. The field of molecular phylogenetics can be defined as follows: *molecular phylogenetics is the discipline concerned with using phylogenetic methodology on molecular genetic data to infer evolutionary phylogenies or "trees" to elucidate the evolutionary relationships and distances or divergence times among DNA sequences, amino acid sequences, populations, species, or higher taxa*. The vast majority of molecular phylogenetic studies have been based on DNA sequence data, typically representing one to several genes, though other types of molecular genetic data such as amino acid sequences are also used. Molecular phylogenies, especially those inferred for a single gene, are commonly called *gene trees*.

In contrast to molecular phylogenetics, "phylogenomics" is more difficult to define for two reasons. First, some methodological and conceptual overlap exists between them—namely both fields rely on phylogenetic methodology to infer phylogenies from molecular data. Secondly, researchers have used the term phylogenomics to characterize different types of studies. We will now take a closer look at how researchers have used the term phylogenomics before we settle on a definition to follow in this book.

1.1.1 The Early View of Phylogenomics

Eisen (1998a) originally coined the term phylogenomics and defined this discipline as the prediction of gene function and study of gene and genome evolution using molecular phylogenies in conjunction with modern comparative methods. For example, in an early phylogenomic study, Eisen (1998b) first inferred the gene tree among members (amino acid sequences) of the MutS family of proteins, a group of proteins important for recognition and repair of DNA mismatches caused by errors during DNA replication. He then used this tree to investigate the evolutionary diversification of this gene family by looking at MutS homologs found within and among genomes across the Tree of Life. Phylogenetic-based methods are not only superior to similarity-based methods for

predicting the functions of unknown genes, but they allow a researcher to split genes into orthologous and paralogous subfamilies and identify key events in the histories of gene families such as gene divergences, lateral gene transfers, and gene losses (Eisen et al. 1995; Eisen 1998a,b).

Shortly thereafter, O'Brien and Stanyon (1999) used the term phylogenomics differently, as they mentioned "comparative phylogenomics" to describe studies using comparative gene maps for a number of closely related species combined with cladistic analysis to reconstruct ancestral genomes (e.g., Haig 1999). Although these two uses of the term phylogenomics were both applied to the study of molecular or genomic evolution, these studies nonetheless differed from each other in terms of data, analytical methods, and study goals.

1.1.2 An Expanded View of Phylogenomics

The purview of phylogenomics broadened further during the early 2000s soon after the genome era commenced, as the rapidly increasing volumes of genomic data—including some fully sequenced eukaryotic genomes such as the human genome—allowed researchers to dramatically scale up sizes of their datasets in phylogenetically based evolutionary studies. It was during this time that researchers could begin using phylogenetic methodology to analyze enormous genome-wide datasets for addressing problems ranging from genome evolution to reconstructing the Tree of Life (Eisen and Fraser 2003; Rokas et al. 2003; Delsuc et al. 2005; Philippe et al. 2005).

The viewpoint that phylogenomics is a discipline comprised of two main areas of inquiry—one concerned with questions in molecular and genomic evolution and the other focused on the evolutionary history of organisms—was subsequently reinforced at the first phylogenomics symposium (Philippe and Blanchette 2007) and in the first book focused on phylogenomics (Murphy 2008). Aside from the dramatic growth in phylogenomic studies since that time, little has changed regarding this dichotomy of research goals. Thus, at the present time we can think of phylogenomics as comprising two major subdisciplines: *molecular phylogenomics* and *organismal phylogenomics*. Accordingly, we may broadly define "phylogenomics" as *the field of study concerned with using genome-wide data to infer the evolution of genes, genomes, and the Tree of Life*. What primarily differentiates molecular phylogenetics from phylogenomics is that the latter field often uses much larger or "computer bursting" datasets and gene trees as independent units of analyses in evolutionary studies.

1.2 ANATOMY OF GENE TREES

Gene trees that have been inferred from DNA sequence data represent fundamental units of analysis in various types of phylogenomic studies. The basic anatomy of a gene tree is illustrated in Figure 1.1. In this figure we see the genealogical relationships among a sample of five DNA sequences labeled a through e for a single gene. Except for some specialized cases (e.g., studying microorganisms in a laboratory), the *true* gene tree cannot be known for a given set of DNA sequences. Instead, the genealogical history must be *inferred* or reconstructed using phylogenetic methodology.

The structure of an inferred gene tree is defined by its branching pattern or "topology" and length of each branch. For example, the tree in Figure 1.1 has four *nodes* (labeled 1 through 4). The bottom-most node (node 4) represents the *root* of the gene tree. The root node is particularly important because it represents the *most recent common ancestor* or "MRCA" of the five DNA sequences and therefore provides directionality or time's arrow along the tree (Figure 1.1). Similarly, we can describe node 1 as the MRCA of a and b, node 2 is the MRCA of a, b, and c, and node 3 is the MRCA of d and e. The other major structural features of gene trees are its *branches*, which represent *lines of descent*.

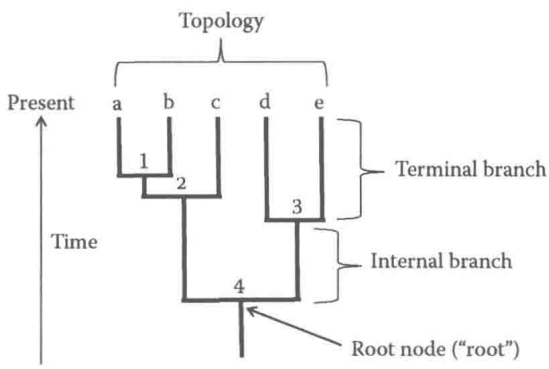


Figure 1.1. An example of a rooted gene tree for five DNA sequences labeled (a–e). In this tree there are four nodes, five terminal branches, and four internal branches. The placement of the root in this tree (node 4) gives the tree a direction with respect to time and thus ancestor–descendant relationships can be inferred.

Branches (i.e., the vertical lines in the tree) can be subdivided into two categories: *terminal branches*, which connect the tips (observed DNA sequences a–e) with nodes below them. For example, the branches connecting node 1 to sequences a and b represent two terminal branches. Likewise, the branch between node 2 and sequence c is also a terminal branch. The other type of branch is known as an *internal branch*. There are a total of four internal branches in this tree and they are found between nodes 1 and 2, nodes 2 and 4, nodes 3 and 4, and below the root node. The lengths of the branches in gene trees can indicate rates of molecular evolution or evolutionary time depending on how the tree is constructed. Note that the tree in Figure 1.1 represents one of the 105 possible different topologies that could be generated for five labeled tips and a tree that is completely bifurcating (i.e., each node has exactly two descendent branches connected to it) and has a root (Felsenstein 2004). The numbers of unique rooted tree topologies becomes shockingly high as the number of sequences increases. For example, a perusal of Table 3.1 in Felsenstein (2004) shows that for only 10 sequences there are more than 34 million different rooted bifurcating trees and a mind boggling 2.75×10^{76} different trees for 50 sequences! As the focus of this book is on acquiring phylogenomic data, we will not delve into the details on how these trees are made. Readers wanting to learn about methods for inferring phylogenetic trees using molecular data should consult the following references: Hillis et al. (1996), Felsenstein (2004), and Lemey et al. (2009).

1.3 GENE TREES VERSUS SPECIES TREES

When a molecular biologist reconstructs a gene tree for a particular gene family, the interpretations of the resulting tree are clear-cut: the tree shows the inferred evolutionary relationships among the *molecules* (amino acid or DNA sequences) used to make the tree and may also display the rates or timing of lineage divergences. In other words, what is gleaned from a gene tree in this type of phylogenomic study is the evolution of the molecular sequences themselves. In contrast, when a single gene tree is used to infer a phylogeny of populations or species, as has been done innumerable times in traditional molecular phylogenetics, then the researcher is extrapolating an *organismal phylogeny* from a molecular phylogeny (Maddison 1995, 1997). In Tree of Life studies, an organismal phylogeny is more commonly referred to as a *species tree* because it shows the branching relationships and times of divergence among populations or species (Maddison 1995, 1997). Although a gene tree may to some extent mirror a species tree, it is important to realize that gene trees and organismal trees are not the same thing. Thus, a researcher who uses a single gene tree to infer a species tree hopes that they match each other (i.e., are congruent).

Schematic examples of gene and species trees are shown side-by-side in Figure 1.2. Let's first consider the meaning of the topology in each evolutionary tree. The gene tree in Figure 1.2a shows the evolutionary relationships between two DNA sequences (haplotypes) sampled from one gene in two species, whereas the species tree in Figure 1.2b shows the evolutionary relationships between the two species. In addition to the topologies, the divergence times in each type of tree are also fundamentally different from each other. Divergence times derived from a gene tree represent *gene divergence times*, whereas in the species tree the divergence times represent *population* or *species divergence times* (Figure 1.2). The MRCA for the two haplotypes (i.e., gene divergence) corresponds to a single individual in the ancestral population, while the MRCA for the two species is the ancestral species. Thus the gene tree shows the inferred genealogical history of the sampled DNA sequences, whereas the latter exhibits the evolutionary history for populations or species. It can be a perilous practice to naively equate a gene tree with a species tree because, even if a gene tree is reconstructed without error, its topology may be incongruent with the corresponding species tree (Hudson 1983, 1992; Tajima 1983; Maddison 1995, 1997; Rosenberg 2002; Felsenstein 2004).

Notice in Figure 1.2 that the widths of the branches differ between gene and species trees. The branches of a gene tree are thin lines of constant width (Figure 1.2a) while the branches of a species tree are much wider branches. In some cases, the branch widths are drawn in this manner simply to distinguish the schematic of a gene tree from a species tree. In other cases, the widths of each branch in a species tree are drawn to be proportional to the *effective population sizes* for that population at particular points in time (Figure 1.2b). Thus wider branches represent larger effective

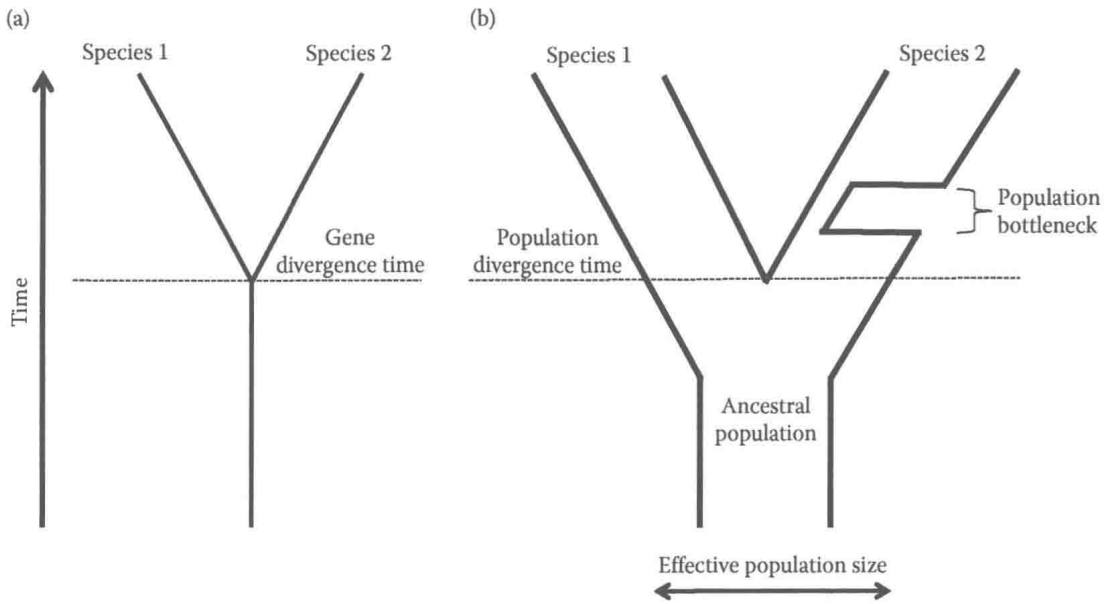


Figure 1.2. A comparison between the anatomy of a gene tree versus a species tree. (a) Shown is a gene tree that depicts the evolutionary history of two haplotypic lineages from two species. The timing of divergence between lineages is called gene divergence time. (b) A species tree showing the evolutionary history of two sister species. The timing of divergence between lineages in a species tree is referred to as the population divergence or speciation time. The branch widths in a species tree are drawn in proportion to the effective population sizes, which may vary through time. In this example, the ancestral population of species 2 is shown to have suffered a population bottleneck.

population sizes than do thinner branches. For example, a species tree with a branch having the same width from node to node means that the long-term effective population sizes have remained constant in size (e.g., ancestral populations of Species 1 in Figure 1.2b). In contrast, the ancestral populations may have undergone fluctuations in size such as from a population bottleneck (e.g., ancestral populations of Species 2 in Figure 1.2b). In contrast, line width in a gene tree has no meaning. Despite the potential lack of congruence between gene and species trees, phylogenomic methods have been developed to address these issues, which in turn, are enabling researchers to generate more accurate and robust estimates of species trees than was ever possible using single gene trees (Knowles and Kubatko 2010).

1.4 PHYLOGENOMICS AND THE TREE OF LIFE

The task of reconstructing the Tree of Life represents a monumental undertaking. In the realm of phylogenomics, there are several levels of study that are contributing to this effort. First, at the shallowest levels in the Tree of Life (i.e., "recent" speciation events), the use of phylogenomic datasets in conjunction with phylogeographic methodology is helping to enumerate the true numbers of extant species as well as provide insights into the history of their formation. *Phylogeography* is the study of how past demographic processes and environmental forces have contributed to speciation and shaped the genetic structure of contemporary populations and species (Avise 2000). Thus, phylogeography provides insights about the temporal and geographical aspects of speciation in recent species radiations and therefore this field shares a close connection to the Tree of Life. On a larger scale, the application of phylogenomic methods for inferring species trees is assisting with the reconstruction of the topology and branch lengths of the Tree of Life (Knowles and Kubatko 2010). Lastly, the use of *DNA barcoding* (Hebert et al. 2003), whether based on single organellar genes, entire organellar genomes, or even vast numbers of nuclear loci, is providing biologists with a powerful and simple tool for identifying known and possible new species. We will now briefly introduce each of these approaches.

Phylogenomics and phylogeography—During the first two decades of phylogeography beginning in the 1980s, the vast majority of phylogeographic studies relied on gene trees that were inferred