PHYLOGENOMIC DATA ACQUISITION

Taylor & Francis
Taylor & Francis Group
http://taylorandfrancis.com

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

© 2017 by Taylor & Francis Group, LLC
CRC Press is an imprint of Taylor & Francis Group, an Informa business

No claim to original U.S. Government works

Printed on acid-free paper
Version Date: 20161021

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

This book contains information obtained from authentic and highly regarded sources. Reasonable efforts have been made to publish reliable data and information, but the author and publisher cannot assume responsibility for the validity of all materials or the consequences of their use. The authors and publishers have attempted to trace the copyright holders of all material reproduced in this publication and apologize to copyright holders if permission to publish in this form has not been obtained. If any copyright material has not been acknowledged please write and let us know so we may rectify in any future reprint.

Except as permitted under U.S. Copyright Law, no part of this book may be reprinted, reproduced, transmitted, or utilized in any form by any electronic, mechanical, or other means, now known or hereafter invented, including photocopying, microfilming, and recording, or in any information storage or retrieval system, without written permission from the publishers.

For permission to photocopy or use material electronically from this work, please access www.copyright.com (http://www.copyright.com/) or contact the Copyright Clearance Center, Inc. (CCC), 222 Rosewood Drive, Danvers, MA 01923, 978-750-8400. CCC is a not-for-profit organization that provides licenses and registration for a variety of users. For organizations that have been granted a photocopy license by the CCC, a separate system of payment has been arranged.

Trademark Notice: Product or corporate names may be trademarks or registered trademarks, and are used only for identification and explanation without intent to infringe.

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

CONTENTS

Preface / ix

Author / xi

1 • INTRODUCTION / 1

1.1 What Is Phylogenomics? / 1

1.1.1 The Early View of Phylogenomics / 1

1.1.2 An Expanded View of Phylogenomics / 2

1.2 Anatomy of Gene Trees / 2

1.3 Gene Trees versus Species Trees / 3

1.4 Phylogenomics and the Tree of Life / 4

1.5 Sequencing Workflows to Generate Phylogenomic Data / 6

1.5.1 Sanger Sequencing Workflow / 6

1.5.2 NGS Workflow / 6

1.5.3 Is Sanger Sequencing Still Relevant in Phylogenomics? / 7

1.6 The Phylogenomics Laboratory / 8

References / 10

2 • PROPERTIES OF DNA SEQUENCE LOCI: PART I / 13

2.1 Genomic Background / 13

2.1.1 Genome Types and Sizes / 13

2.1.2 Composition of Eukaryotic Organellar Genomes / 15

2.1.3 Composition of Eukaryotic Nuclear Genomes / 18

2.1.3.1 Gene Numbers and Densities among Nuclear Genomes / 18

2.1.3.2 Intergenic DNA / 18

2.2 DNA Sequence Evolution / 21

2.2.1 Patterns and Processes of Base Substitutions / 21

2.2.1.1 Transition Bias / 21

2.2.1.2 Transition Bias and DNA Replication Errors / 23

2.2.1.3 Saturation of DNA Sites / 24

2.2.1.4 Among-Site Substitution Rate Variation / 26

2.2.2 Tandemly Repeated DNA Sequences / 26

2.2.3 Transposable Elements / 27

2.2.4 Processed Pseudogenes / 30

2.2.5 Mitochondrial Pseudogenes (“Numts”) / 30
2.2.5.1 Numt Abundance in Eukaryotic Genomes / 30
2.2.5.2 Mechanisms of Primary Numt Integration / 31
2.2.5.3 Differences between Numts and Mitochondrial DNA / 32
References / 33

3 • PROPERTIES OF DNA SEQUENCE LOCI: PART II / 37
3.1 Six Assumptions about DNA Sequence Loci in Phylogenomic Studies / 37
3.1.1 Assumption 1: Loci Are Single-Copy in the Genome / 37
3.1.2 Assumption 2: Loci Are Selectively Neutral / 39
3.1.2.1 Does “Junk DNA” Exist? / 39
3.1.2.2 The Neutrality Assumption and the Indirect Effects of Natural Selection / 40
3.1.3 Assumption 3: Sampled Loci Have Independent Gene Trees / 43
3.1.3.1 How Many Independent Loci Exist in Eukaryotic Genomes? / 44
3.1.3.2 Criteria for Delimiting Loci with Independent Gene Trees / 45
3.1.4 Assumption 4: No Historical Recombination within Loci / 47
3.1.4.1 Intralocus Recombination and Gene Trees / 47
3.1.4.2 What Is the Optimal Locus Length? / 48
3.1.5 Assumption 5: Loci Evolved Like a Molecular Clock / 51
3.1.6 Assumption 6: Loci Are Free of Ascertainment Bias / 52
3.2 DNA Sequence Loci: Terminology and Types / 52
3.2.1 On Genes, Alleles, and Related Terms / 52
3.2.2 Commonly Used DNA Sequence Loci in Phylogenomic Studies / 53
3.2.2.1 Mitochondrial DNA Loci / 53
3.2.2.2 Nuclear DNA Loci / 54
References / 60

4 • DNA EXTRACTION / 67
4.1 DNA Extraction Methodology / 67
4.1.1 Summary of the DNA Extraction Process / 67
4.1.2 A Note about DNA Storage Buffers / 69
4.1.3 Extracting DNA from Plants, Fungi, and Invertebrates / 70
4.1.4 Extracting DNA from Formalin-Fixed Museum Specimens / 70
4.2 Evaluating the Results of DNA Extractions / 71
4.2.1 Agarose Gel Electrophoresis / 72
4.2.1.1 Troubleshooting / 74
4.2.2 UV Spectrophotometric Evaluation of DNA Samples / 75
4.2.2.1 UV Spectrophotometry to Determine Concentrations of Nucleic Acid Samples / 75
4.2.2.2 UV Spectrophotometry to Determine the Purity of DNA Samples / 76
4.2.3 Fluorometric Quantitation of DNA Samples / 76
4.3 The High-Throughput Workflow / 76
4.3.1 High-Throughput DNA Extractions / 77
4.3.1.1 Extracting DNA from 96 Tissue Samples / 77
4.3.1.2 High-Throughput Agarose Gel Electrophoresis / 78
4.3.1.3 High-Throughput UV Spectrophotometry / 78
4.3.1.4 Preparation of Diluted DNA Templates for High-Throughput PCR / 78
References / 79

5 • PCR THEORY AND PRACTICE / 81
5.1 Historical Overview / 81
5.2 DNA Polymerization in Living Cells versus PCR / 83
5.2.1 Brief Review of DNA Polymerization in Living Cells / 83

5.2.2 How the PCR Works / 85
5.3 PCR Procedures / 89
5.3.1 Preparation of PCR Reagents and Reaction Setup / 90
5.3.1.1 PCR Reagents / 90
5.3.1.2 Importance of Making Reagent Aliquots / 91
5.3.1.3 Setting Up PCRs / 92
5.3.2 Thermocycling / 93
5.3.3 Checking PCR Results Using Agarose Gel Electrophoresis / 94
5.4 PCR Troubleshooting / 94
5.5 Reducing PCR Contamination Risk / 97
5.6 High-Throughput PCR / 98
5.6.1 Setting Up PCRs in a 96-Sample Microplate Format / 98
5.7 Other PCR Methods / 98
5.7.1 Hot Start PCR / 99
5.7.2 Long PCR / 100
5.7.3 Reverse Transcriptase-PCR / 101
References / 102

6 • SANGER SEQUENCING / 105
6.1 Principles of Sanger Sequencing / 105
6.1.1 The Sanger Sequencing Concept / 105
6.1.2 Modern Sanger Sequencing / 107
6.1.2.1 Cycle Sequencing Reaction / 107
6.1.2.2 Gel Electrophoresis of Extension Products / 108
6.1.2.3 Sequence Data Quality / 109
6.2 Sanger Sequencing Procedures / 112
6.2.1 Purification of PCR Products / 112
6.2.1.1 Exo-SAP Treatment of PCR Products / 112
6.2.1.2 Spin Column and Vacuum Manifold Kits for PCR Product Purification / 112
6.2.1.3 20% PEG 8000 Precipitation of PCR Products / 113
6.2.1.4 Solid-Phase Reversible Immobilization Beads / 113
6.2.1.5 Gel Purification of PCR Products / 114
6.2.1.6 Which PCR Product Purification Method Is Best? / 115
6.2.2 Setting Up Cycle Sequencing Reactions / 115
6.2.3 Purification of Extension Products / 115
6.2.4 Sequencing in a Capillary Sequencer: Do-It-Yourself or Outsource? / 116
6.3 High-Throughput Sanger Sequencing / 116
6.3.1 Sequencing 96 Samples on Microplates / 116
6.3.2 Adding Sequencing Primer "Tails" to PCR Primers / 117
6.3.2.1 How an M13-Tailed Primer Functions in PCR / 118
6.3.2.2 Cycle Sequencing and M13 Primer Tails / 118
6.3.2.3 On the Importance of Matching Sequencing Primers / 121
6.3.2.4 Benefits of Using M13-Tailed Primers / 123
6.4 Haplotype Determination from Sanger Sequence Data / 123
6.4.1 PCR Amplification and Sanger Sequencing of Diploid or Polyploid Loci / 123
6.4.2 Multiple Heterozygous SNP Sites and Haplotype Sequences / 126
6.4.3 Methods for Obtaining Nuclear Haplotype Sequences from Sanger Sequence Data / 127
6.4.3.1 Physical Isolation of PCR Haplotypes prior to Sequencing / 128
6.4.3.2 Statistical Inference of Haplotypes from Sanger Sequence Data / 128
References / 129

7 • ILLUMINA SEQUENCING / 131
7.1 How Illumina Sequencing Works / 131
7.1.1 Construction of Indexed Sequencing Libraries / 133
7.1.2 Generation of Clusters on a Flow Cell / 133
7.1.3 Sequencing of Clusters / 135
7.2 Methods for Obtaining Multiplexed Hybrid Selection Libraries / 143
7.2.1 Library Preparation Approaches / 145
7.2.1.1 Traditional Illumina Library Approach / 145
7.2.1.2 Meyer and Kircher Library Approach / 156
7.2.1.3 Rohland and Reich Library Approach / 163
7.2.1.4 Nextera Library Approach / 165
7.2.2 In-Solution Hybrid Selection / 175
7.2.3 Indexing, Pooling, and Hybrid Selection Efficiency Revisited / 185
7.3 Cost-Effective Methods for Obtaining Multiplexed Targeted-Loci Libraries / 187
7.3.1 Sequence Capture Using PCR-Generated Probes (SCPP) / 187
7.3.2 Parallel Tagged Amplicon Sequencing / 190
References / 191

8 • DEVELOPING DNA SEQUENCE LOCI / 195
8.1 Primer Design Theory / 196
8.1.1 Rules of Primer Design / 196
8.1.2 Final Comments about Primer Design Rules / 204
8.1.3 Testing New Primers in the Lab / 205
8.2 Primer and Probe Design Approaches / 205
8.2.1 Single Template Approaches for Developing PCR-Based Loci / 206
8.2.1.1 Single Template Methods Using Genomic Cloning Methods / 206
8.2.1.2 Single Template Methods Using Available Genomics Resources / 210
8.2.1.3 Single Template Methods Using NGS Partial Genome Data / 210
8.2.1.4 Single Template Methods Using Whole Genome Sequences / 211
8.2.2 Multiple Homologous Template Approaches for Designing PCR-Based and Anchor Loci / 211
8.2.2.1 Designing Universal Primers by Comparative Sequence Analysis / 212
8.2.2.2 Multiple Homologous Template Approaches Using Whole Genome Sequences / 215
8.2.2.3 Designing Anchor Loci Probes Using Whole Genome Sequences / 216
References / 217

9 • FUTURE OF PHYLOGENOMIC DATA ACQUISITION / 221
9.1 The Impending Flood of Genomes / 221
9.2 *In Silico* Acquisition of Phylogenomic Datasets / 222
References / 224

Index / 225

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.

Taylor & Francis
Taylor & Francis Group
http://taylorandfrancis.com