

Phylogenomics

Christoph Bleidorn

Phylogenomics

An Introduction

 Springer

Christoph Bleidorn
Museo Nacional de Ciencias Naturales
Spanish National Research Council (CSIC)
Madrid
Spain

ISBN 978-3-319-54062-7 ISBN 978-3-319-54064-1 (eBook)
DOI 10.1007/978-3-319-54064-1

Library of Congress Control Number: 2017942964

© Springer International Publishing AG 2017

This work is subject to copyright. All rights are reserved by the Publisher, whether the whole or part of the material is concerned, specifically the rights of translation, reprinting, reuse of illustrations, recitation, broadcasting, reproduction on microfilms or in any other physical way, and transmission or information storage and retrieval, electronic adaptation, computer software, or by similar or dissimilar methodology now known or hereafter developed.

The use of general descriptive names, registered names, trademarks, service marks, etc. in this publication does not imply, even in the absence of a specific statement, that such names are exempt from the relevant protective laws and regulations and therefore free for general use.

The publisher, the authors and the editors are safe to assume that the advice and information in this book are believed to be true and accurate at the date of publication. Neither the publisher nor the authors or the editors give a warranty, express or implied, with respect to the material contained herein or for any errors or omissions that may have been made. The publisher remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Printed on acid-free paper

This Springer imprint is published by Springer Nature
The registered company is Springer International Publishing AG
The registered company address is: Gewerbestrasse 11, 6330 Cham, Switzerland

Preface

All life on earth shares a common ancestor, and the aim of phylogenetic systematics is to reconstruct the tree or network of life. Shortly after the availability of the first protein sequences, molecular phylogenetic approaches were developed to understand the evolutionary relationships between proteins (or genes). It became clear that gene trees will also help to unravel the phylogeny of species. The introduction of Sanger sequencing and polymerase chain reaction (PCR) paved the way that genetic approaches became available across the scientific community and contributed to the rise of molecular phylogenetics. At the end of the 1990s, results from single-gene studies challenged the century-old textbook view of evolutionary relationships of many groups (e.g. animals, plants). Fierce discussions regarding the validity of these results led to important methodological advances, and, nowadays, molecular phylogenies are broadly accepted to represent organismal relationships in textbooks. In the mid-2000s, the way of sequencing has been revolutionized, leading to a huge drop in its costs, and unprecedented amounts of sequence data became affordable for every type of study and also for non-model organisms. This development transformed the field of molecular phylogenetics to phylogenomics, where genome-scale data (genomes, transcriptomes) can be exploited. The term phylogenomics was already coined in 1998 by Jonathan Eisen (also known under his twitter handle @phylogenomics), who outlined the importance of phylogenetic methods for the annotation of genes without relying on direct (time consuming) functional studies. This underlines how deeply embedded phylogenetic methods are in the field of genomics. The theoretical background for reconstructing gene trees (functional annotations) and species trees (reconstruction of the tree of life) is broadly overlapping. In this book I will introduce the major steps of phylogenomic analyses in general. The first two chapters briefly introduce the field of genomics (► Chap. 1, «Genomes») and the evolution and peculiarities of organellar genomes (► Chap. 2, «Organellar Genomes and Endosymbionts»). In ► Chap. 3 («Sequencing Techniques»), I review the most widely used sequencing platforms, which is difficult in a print format, as the field advances so fast that many numbers describing the output of these machines might be already out of date when you read this chapter. ► Chapter 4 («Sequencing Strategies») gives an overview of different strategies to sequence complete or partial genomes and transcriptomes. The outputs of every sequencing platform are sequences which are considerably shorter than chromosomes and in the case of short-read sequencing also shorter than most genes. In ► Chap. 5 («Assembly and Data Quality»), ways to puzzle these small pieces into more complete representations of genomes and genes (called assembly) are introduced. Fundamental steps for every phylogenomic study are alignments, read mapping and finding homologous genes, which are explained in ► Chaps. 6 («Alignment and Mapping») and 7 («Finding Genes»). Based on a sequence alignment, it is possible to reconstruct phylogenetic trees, and the methods are briefly reviewed in ► Chap. 8 («Phylogenetic Analyses»). I kept this chapter on purpose rather brief, as many excellent textbooks describing these methods (and its underlying algorithms) in detail are available (see references in ► Chap. 8). Moreover, the basic theory underlying these methods did not change much in the last decade. Surprisingly, even with this vast amount of data, many phylogenetic

questions remain still difficult to resolve. Some problems of phylogenetic reconstruction get even amplified when using hundreds or thousands of genes due to the presence of systematic error. ► Chapter 9 («Sources of Error and Incongruence in Phylogenomic Analyses») gives an overview of possible sources of error, as well as recommendations on how to deal with them. Moreover, the differences in analysing gene trees and species trees and possible sources of incongruence between those are outlined. Finally, in ► Chap. 10 («Rare Genomic Changes»), I introduce further phylogenetic markers apart from plain sequence data (e.g. integrations of mobile elements, gene order) and give an overview on how these rare genomic changes are utilized for phylogenetic systematics.

During my time at German universities, I was heavily involved in teaching bachelor and master level students. This included lectures, seminars and practical courses. While the field of molecular phylogenetics changed while moving into the postgenomic era, so did my courses. Besides the introduction of phylogenetic methods (e.g. maximum parsimony, maximum likelihood), I realized that more and more background knowledge became of major importance to carry out phylogenetic analyses. This includes knowledge about genomics, sequencing techniques as well as bioinformatic approaches to handle sequence data before the actual phylogenetic analysis starts. With this book I want to give a concise overview of all major steps of a phylogenomic analyses, as well as some insights into recent advantages in the field of genomics. This book is mainly addressed to undergraduate and graduate biology students, but also postdocs newly moving to the field of phylogenomics might use it as a first overview. The chapters are written in a concise way and focus more on explaining the idea behind methods, instead of deeply digging into the algorithmic or technical background. However, I tried always to refer to the appropriate specific literature to get deeper insights into any method (or study) of interest. Furthermore, I specified widely used and important software for every step of the phylogenetic analysis. When possible, I mention several alternatives. The name of software or scripts is always written in all caps, irrespective of the original way a name is written. This book does not include instructions on how to use this software, as in most cases detailed descriptions are available in the manual. As already noted, this book is mainly addressed to biology students. Working in the field of phylogenomics needs good to excellent (bio)informatic skills. Unfortunately, in the curriculum of many bachelor and master programmes, bioinformatics are not taught. However, several international courses teaching programming skills for (evolutionary) biologists take place regularly (e.g. Cold Spring Harbor Course «Programming for Biology»; Programming for Evolutionary Biology in Leipzig), and many excellent online tutorials are available. As such I can only strongly suggest to any student interested in this field to get used to work with Linux/Unix command lines and to acquire at least basic knowledge into (scripting) languages like Python, Perl or R.

I would like to thank several colleagues who commented on earlier versions of the here published chapters. In alphabetical order, they are Maite Aguado, Marie-Theres Gansauge, Michael Gerth, Iker Irisarri, Lars Podsiadlowski and Alexander Suh. I am grateful that Eva Nowack provided a picture of the enigmatic *Paulinella*. Moreover, I want to thank Lars Vogt, Christoph Held and Andreas Schmidt-Rhaesa for introducing

me into the theoretical and practical world of molecular phylogenetics. The above-mentioned university courses, which helped me to develop the outline and content of this book, were taught at the Free University of Berlin, University of Potsdam and University of Leipzig (in collaboration with Matthias Meyer from the Max Planck Institute for Evolutionary Anthropology). I would like to thank the department heads Thomas Bartolomaeus, Ralph Tiedemann and Martin Schlegel who gave me complete freedom in filling these courses with life.

Christoph Bleidorn

Madrid, Spain, January 2017

Contents

1	Genomes	1
1.1	The Ring of Life	2
1.2	Genome Structure	4
1.3	Genome Size	7
1.4	The Genomes of Modern and Archaic Humans	10
	References.....	14
2	Organelle Genomes and Endosymbionts	21
2.1	Mitochondria	22
2.1.1	Origin and Evolution of Mitochondria.....	22
2.1.2	Animal Mitochondrial Genomes.....	25
2.1.3	Mitochondrial Genomes of Plants and Algae.....	26
2.1.4	Mitochondrial Genomes of «Other» Eukaryotes.....	28
2.2	Plastids	29
2.2.1	Origin and Evolution of Plastids.....	29
2.2.2	Plastid Genomes.....	31
2.2.3	Plastids in the Amoeba <i>Paulinella chromatophora</i>	32
2.3	Heritable Bacterial Endosymbionts	33
2.3.1	Primary Endosymbionts.....	33
2.3.2	Secondary Endosymbionts.....	35
2.4	DNA Barcoding	35
	References.....	37
3	Sequencing Techniques	43
3.1	Sanger Sequencing	44
3.2	454 Pyrosequencing	45
3.3	Reversible Terminator Sequencing (Illumina)	47
3.4	Ion Semiconductor Sequencing (Ion Torrent)	49
3.5	Single-Molecule Real-Time (SMRT) Sequencing (PacBio)	51
3.6	Nanopore Sequencing	53
3.7	Comparison of Sequencing Platforms	55
	References.....	57
4	Sequencing Strategies	61
4.1	Shotgun Sequencing	62
4.2	RADseq	67
4.3	Hybrid Enrichment	70
4.4	Expressed Sequence Tags and RNA-Seq	73
4.5	Single-Cell Genomics and Transcriptomics	75
	References.....	75

5	Assembly and Data Quality	81
5.1	Data Quality and Filtering	82
5.2	Assembly Strategies	84
5.2.1	Greedy Assemblies	87
5.2.2	Overlap-Layout-Consensus (OLC) Assemblies	88
5.2.3	K-mer Assemblies Using de Bruijn Graphs	90
5.3	Comparing Assemblies	94
5.4	De Novo Assembly of Genomes	96
5.4.1	Scaffolding	96
5.4.2	Hybrid Assemblies	97
5.5	De Novo Assembly of Transcriptomes and Metagenomes	97
	References	100
6	Alignment and Mapping	105
6.1	Pairwise Alignment	106
6.2	Local Alignment and BLAST Searches	111
6.3	Multiple Sequence Alignment	114
6.4	Alignment Masking	115
6.5	Mapping Sequence Reads	117
6.6	Whole-Genome Alignments	121
	References	122
7	Finding Genes	127
7.1	What Is a Gene?	128
7.2	Gene Gain and Loss	128
7.3	Homology of Genes	130
7.4	Inferring Orthology	131
7.5	Hidden Markov Profiles	133
7.6	Gene Ontology and the Ortholog Conjecture	136
7.7	Whole-Genome Duplications	138
	References	139
8	Phylogenetic Analyses	143
8.1	Trees	144
8.2	Models of Nucleotide Substitution	147
8.3	Models of Amino Acid Substitutions	152
8.4	Model Selection and Data Partitions	155
8.4.1	Model Selection	155
8.4.2	Partition Finding	157
8.5	Inferring Phylogenies	158
8.5.1	Neighbour Joining	158
8.5.2	Maximum Parsimony	159
8.5.3	Maximum Likelihood	160
8.5.4	Heuristic Methods and Genetic Algorithms	162
8.5.5	Bayesian Inference	163
8.6	Support Measures	165
8.7	Molecular Clocks	166
	References	168

9	Sources of Error and Incongruence in Phylogenomic Analyses	173
9.1	Incongruence in Phylogenomic Analyses	174
9.2	Systematic Errors	177
9.3	Missing Data, Phylogenetic Information Content and Taxon Sampling	180
9.3.1	Missing Data.....	180
9.3.2	More Genes or More Taxa?.....	182
9.3.3	Taxon Sampling.....	182
9.3.4	Gene Sampling.....	183
9.4	Incongruence Between Gene Trees and Species Trees	186
	References.....	189
10	Rare Genomic Changes	195
10.1	The Perfect Phylogenetic Marker	196
10.2	Mobile Elements	198
10.3	MicroRNAs	201
10.4	Introns	202
10.5	Gene Order	203
10.6	Changes in the Genetic Code	206
	References.....	207
	Service Part	
	Glossary.....	214
	Index.....	219

Abbreviations

μm	Micrometre	Mb	Mega base pairs
A	Adenine	MCMC	Markov chain Monte Carlo method
AIC	Akaike information criterion	MCMCMC	Metropolis-coupled Markov chain Monte Carlo method
ATP	Adenosine triphosphate	MITE	Miniature inverted-repeat transposable element
BAC	Bacterial artificial chromosome	ML	Maximum likelihood
BI	Bayesian inference	MP	Maximum parsimony
BIC	Bayesian information criterion	mRNA	Messenger RNA
BLAST	Basic Local Alignment Search Tool	mya	Million years ago
bp	Base pairs	NCBI	National Center for Biotechnology Information
C	Cytosine	NGS	Next-generation sequencing
cDNA	Complementary DNA	NIP	Near intron pair
CI	Cytoplasmic incompatibility	NJ	Neighbour joining
CMOS	Complementary metal-oxide semiconductor	NNI	Nearest neighbour interchange
CNV	Copy number variation	OTU	Operational taxonomic unit
CRISPR	Clustered regularly interspaced short palindromic repeat	PAM	Point accepted mutations
ddNTP	Dideoxynucleoside triphosphate	PCR	Polymerase chain reaction
DNA	Deoxyribonucleic acid	PE	Paired-end sequencing
dNTP	Deoxynucleoside triphosphate	pH	Power of hydrogen
DUI	Doubly uniparental inheritance	QTL	Quantitative trait loci
G	Guanine	RNA	Ribonucleic acid
Gb	Giga base pairs	SINE	Short interspersed element
GBS	Genotyping by sequencing	SMRT	Single-molecule real-time
GTR	General time-reversible model	SNP	Single nucleotide polymorphism
GWAS	Genome-wide association study	SPR	Subtree pruning and regrafting
HGT	Horizontal gene transfer	T	Thymine
ICE	Integrative conjugative element	Tb	Tera base pairs
ILS	Incomplete lineage sorting	TBR	Tree bisection and reconnection
ISFET	Ion-sensitive field-effective transistor	TE	Transposable element
Kb	Kilo base pairs	TPRT	Target-primed reverse transcription
LBA	Long-branch attraction	tRNA	Transfer RNA
LD	Linkage disequilibrium	UCE	Ultraconserved element
LINE	Long interspersed element	wgs	Whole-genome shotgun
LRT	Likelihood ratio test	ZMW	Zero-mode waveguide
LTR	Long terminal repeat		