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

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

Service Part

Abbreviations

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