"Phylogenetics in the Genomic Era" brings together experts in the field to present a comprehensive synthesis

Robert Waterhouse, Karen Meusemann

Peer Community In Genomics

“Phylogenetics in the Genomic Era” brings together experts in the field to present a comprehensive synthesis

Robert Waterhouse and **Karen Meusemann**

A recommendation of:

Phylogenetics in the Genomic Era

Céline Scornavacca, Frédéric Delsuc, Nicolas Galtier *(2021), HAL, PGE* **https://hal.inria.fr/PGE/**

Open Access

Submitted: 15 March 2022, Recommended: 08 April 2022

Cite this recommendation as:
Robert Waterhouse and Karen Meusemann (2022) “Phylogenetics in the Genomic Era” brings together experts in the field to present a comprehensive synthesis. *Peer Community in Genomics, 100015.* **https://doi.org/10.24072/pci.genomics.100015**

Published: 2022-04-08

Recommendation

E-book: Phylogenetics in the Genomic Era (Scornavacca et al. 2021)

This book was not peer-reviewed by PCI Genomics. It has undergone an internal review by the editors.

Accurate reconstructions of the relationships amongst species and the genes encoded in their genomes are an essential foundation for almost all evolutionary inferences emerging from downstream analyses. Molecular phylogenetics has developed as a field over many decades to build suites of models and methods to reconstruct reliable trees that explain, support, or refute such inferences. The genomic era has brought new challenges and opportunities to the field, opening up new areas of research and algorithm development to take advantage of the accumulating large-scale data. Such ‘big-data’ phylogenetics has come to be known as phylogenomics, which broadly aims to connect molecular and evolutionary biology research to address questions centred on relationships amongst taxa, mechanisms of molecular evolution, and the biological functions of genes and other genomic elements. This book brings together experts in the field to present a comprehensive synthesis of Phylogenetics in the Genomic Era, covering key conceptual and methodological aspects of how to build accurate phylogenies and how to apply them in molecular and evolutionary research. The paragraphs below briefly summarise the five constituent parts of the book, highlighting the key concepts, methods, and applications that each part addresses. Being organised in an accessible style, while presenting details to provide depth where necessary, and including guides describing real-world examples of major phylogenomic tools, this collection represents an invaluable resource, particularly for students and newcomers to the field of phylogenomics.

Part 1: Phylogenetic analyses in the genomic era

Modelling how sequences evolve is a fundamental cornerstone of phylogenetic reconstructions. This part of the book introduces the reader to phylogenetic inference methods and algorithmic optimisations in the contexts of Markov, Maximum Likelihood, and Bayesian models of sequence evolution. The main concepts and theoretical considerations are mapped out for probabilistic Markov models, efficient tree building with Maximum Likelihood methods, and the flexibility and robustness of Bayesian approaches. These are supported with practical examples of phylogenomic applications using the popular tools RAxML and PhyloBayes. By considering theoretical, algorithmic, and practical

aspects, these chapters provide readers with a holistic overview of the challenges and recent advances in developing scalable phylogenetic analyses in the genomic era.

Part 2: Data quality, model adequacy

This part focuses on the importance of considering the appropriateness of the evolutionary models used and the accuracy of the underlying molecular and genomic data. Both these aspects can profoundly affect the results when applying current phylogenomic methods to make inferences about complex biological and evolutionary processes. A clear example is presented for methods for building multiple sequence alignments and subsequent filtering approaches that can greatly impact phylogeny inference. The importance of error detection in (meta)barcode sequencing data is also highlighted, with solutions offered by the MACSE_BARCODE pipeline for accurate taxonomic assignments. Orthology datasets are essential markers for phylogenomic inferences, but the overview of concepts and methods presented shows that they too face challenges with respect to model selection and data quality. Finally, an innovative approach using ancestral gene order reconstructions provides new perspectives on how to assess gene tree accuracy for phylogenomic analyses. By emphasising through examples the importance of using appropriate evolutionary models and assessing input data quality, these chapters alert readers to key limitations that the field as a whole strives to address.

Part 3: Resolving phylogenomic conflicts

Conflicting phylogenetic signals are commonplace and may derive from statistical or systematic bias. This part of the book addresses possible causes of conflict, discordance between gene trees and species trees and how processes that lead to such conflicts can be described by phylogenetic models. Furthermore, it provides an overview of various models and methods with examples in phylogenomics including their pros and cons. Outlined in detail is the multispecies coalescent model (MSC) and its applications in phylogenomics. An interesting aspect is that different phylogenetic signals leading to conflict are in fact a key source of information rather than a problem that can – and should – be used to point to events like introgression or hybridisation, highlighting possible future trends in this research area. Last but not least, this part of the book also addresses inferring species trees by concatenating single multiple sequence alignments (gene alignments) versus inferring the species tree based on ensembles of single gene trees pointing out advantages and disadvantages of both approaches. As an important take home message from these chapters, it is recommended to be flexible and identify the most appropriate approach for each dataset to be analysed since this may tremendously differ depending on the dataset, setting, taxa, and phylogenetic level addressed by the researcher.

Part 4: Functional evolutionary genomics

In this part of the book the focus shifts to functional considerations of phylogenomics approaches both in terms of molecular evolution and adaptation and with respect to gene expression. The utility of multi-species analysis is clearly presented in the context of annotating functional genomic elements through quantifying evolutionary constraint and protein-coding potential. An historical perspective on characterising rates of change highlights how phylogenomic datasets help to understand the modes of molecular evolution across the genome, over time, and between lineages. These are contextualised with respect to the specific aim of detecting signatures of adaptation from protein-coding DNA alignments using the example of the MutSelDP-$\omega*$ model. This is extended with the presentation of the generally rare case of adaptive sequence convergence, where consideration of appropriate models and knowledge of gene functions and phenotypic effects are needed. Constrained or relaxed, selection pressures on sequence or copy-number affect genomic elements in different ways, making the very concept of function difficult to pin down despite it being fundamental to relate the genome to the phenotype and organismal fitness. Here gene expression provides a measurable intermediate, for which the Expression Comparison tool from the Bgee suite allows exploration of expression patterns across multiple animal species taking into account anatomical homology. Overall, phylogenomics applications in functional evolutionary genomics build on a rich theoretical history from molecular analyses where integration with knowledge of gene functions is challenging but critical.

Part 5: Phylogenomic applications

Rather than attempting to review the full extent of applications linked to phylogenomics, this part of the book focuses on providing detailed specific insights into selected examples and methods concerning i) estimating divergence times, and ii) species delimitation in the era of 'omics' data. With respect to estimating divergence times, an exemplary overview is provided for fossil data recovered from geological records, either using fossil data as calibration points with an extant-species-inferred phylogeny, or using a fossilised birth-death process as a mechanistic model that accounts for lineage diversification. Included is a tutorial for a joint approach to infer phylogenies and estimate divergence times using the RevBayes software with various models implemented for different applications and datasets incorporating molecular and morphological data. An interesting excursion is outlined focusing on timescale estimates with respect to viral evolution introducing BEAGLE, a high-performance likelihood-calculation platform that can be used on multi-core systems. As a second major subject, species delimitation is addressed since currently the increasing amount of available genomic data enables extensive inferences, for instance about the degree of genetic isolation among species and ancient and recent introgression events. Describing the history of molecular species delimitation up to the current genomic era and

presenting widely used computational methods incorporating single- and multi-locus genomic data, pros and cons are addressed. Finally, a proposal for a new method for delimiting species based on empirical criteria is outlined. In the closing chapter of this part of the book, BPP (Bayesian Markov chain Monte Carlo program) for analysing multi-locus sequence data under the multispecies coalescent (MSC) model with and without introgression is introduced, including a tutorial. These examples together provide accessible details on key conceptual and methodological aspects related to the application of phylogenetics in the genomic era.

References

Scornavacca C, Delsuc F, Galtier N (2021) Phylogenetics in the Genomic Era. https://hal.inria.fr/PGE/