Commentary
Volume 9:9, 2021

Journal of
Phylogenetics & Evolutionary Biology

ISSN: 2329-9002 Open Access

Phylogenomics – An Overview

Roumi Ghosh*

Huxley Faculty Fellow, Department of Ecology and Evolutionary Biology, Rice University, USA

Commentary

Phylogenetics is the science of reconstructing the evolutionary history of life on Earth. Traditionally, phylogenies were constructed using morphological data only, but the introduction of Sanger sequencing and PCR in the late 1970s enabled genetic information to be incorporated into phylogenetic analyses. Early phylogenetic studies employing multilocus analyses contributed greatly to our knowledge of phylogenetic history and challenged some well-established views of the relationships among many groups of plants and animals. Since the publication of these pioneering studies, significant methodological advances in both sequencing and analytical techniques have been made, and molecular phylogenies are now broadly accepted to represent robust hypotheses of organismal relationships. Next-generation sequencing techniques, developed in the mid-2000s, revolutionized DNA sequencing and led to a dramatic reduction in sequencing cost per nucleotide and a sharp increase in data generation speed. As a result, the generation of unprecedented amounts of sequence data for both model and non-model organisms has become affordable. This development has transformed the field of molecular phylogenetic into Phylogenomics—where genome-scale data are obtained from multiple samples at once at a much reduced cost.

The phylogenomic pipeline can be very complex, presenting an overwhelming array of methodologies available for the acquisition, manipulation, analysis and interpretation of massive datasets. Researchers also have to overcome the challenges of sequencing strategy design, identification of orthologous loci, model selection and phylogeny estimation. This can be particularly daunting for researchers new to the field—both students and established scientists—who wish to delve into novel methods and data to reconstruct the evolution of their study group. Here we present an entry-level overview of the theory and tools that are central to Phylogenomics, with an emphasis on the appropriate application of techniques useful for phylogenetic analysis of genomic data. We focus on the sequencing technologies and statistical methods for phylogeny estimation, and the software implementing these methods and their application to large molecular datasets. We also discuss the tools and tradeoffs for improving the accuracy of phylogenomic analyses, including the biological and methodological sources of systematic error in phylogeny estimation. Finally, we provide a glossary of commonly encountered terms used in Phylogenomics that may be useful for those entering the field and hoping to sort through the multitude of methods, analytical tools and terminology inherent to this relatively new, but rapidly advancing field.

The word 'Phylogenomics' was first introduced in the context of prediction of gene function for genome-scale data, and soon after in the context of phylogenetic inference. The discipline of Phylogenomics owes its existence to the advances made in DNA sequencing technology over the past two decades. It comprises several areas of research at the interface between molecular and evolutionary biology and has two major goals: (i) to infer phylogenetic relationships between taxa and gain insights into the mechanisms of molecular evolution; and (ii) to use multispecies phylogenetic comparisons to infer putative functions for DNA or protein sequences.

Traditional Sanger sequencing studies include relatively few loci and are therefore limited by stochastic or sampling error. As there is a relatively small number of phylogenetic ally informative characters available in one or a few genes, this random 'noise' influences the inference of backbone nodes, potentially leading to poorly resolved or poorly supported phylogenetic trees. This problem can be addressed successfully by using much larger amounts of sequence data. Modern phylogenomic analyses, which take advantage of hundreds to thousands of loci from across the genome, are, on average, orders of magnitude larger than traditional Sanger sequencing datasets. The size of these datasets therefore significantly reduces the impact of stochastic error and data availability as a limiting factor, offering great promise for resolving historically recalcitrant nodes in the tree of life.

High-throughput sequencing technologies [also called next-generation sequencing (NGS)] have yielded genome-scale data in immense quantities. Next-generation sequencing technologies differ fundamentally from the Sanger method in that they allow for massively parallel DNA sequencing, providing extremely high throughput from multiple samples simultaneously and at a much reduced cost. Millions to billions of DNA nucleotides can be sequenced in parallel, yielding orders of magnitude more data and minimizing the need for the fragment-cloning methods that are used with Sanger sequencing. Recent progress in NGS technology and the rapid development of bioinformatics tools now allow research groups of any size to generate large amounts of genomic sequences for organisms of interest. High-throughput sequencing can be used for whole-genome sequencing, whole-Transcriptome shotgun sequencing (also called RNA sequencing, RNA-seq, or transcriptomics, whole-exome sequencing, and reduced-representation genome sequencing (also called target enrichment).

***Address for Correspondence:** Roumi Ghosh, Huxley Faculty Fellow, Department of Ecology and Evolutionary Biology, Rice University, USA, E-mail: roumighos@gmail.com*

Received 08 September 2021; **Accepted** 20 September 2021; **Published** 27 September 2021

How to cite this article: Roumi Ghosh. "Phylogenomics – An Overview." *J Phylogenetics Evol Biol* 9 (2021) 180.