SpringerBriefs in Genetics

For further volumes:
http://www.springer.com/series/8923

Xuhua Xia

Comparative Genomics

Springer

Xuhua Xia
Department of Biology
University of Ottawa
Ottawa, ON
Canada

ISSN 2191-5563 ISSN 2191-5571 (electronic)
ISBN 978-3-642-37145-5 ISBN 978-3-642-37146-2 (eBook)
DOI 10.1007/978-3-642-37146-2
Springer Heidelberg New York Dordrecht London

Library of Congress Control Number: 2013936000

Printed on acid-free paper

Springer is part of Springer Science+Business Media (www.springer.com)

Preface

This book on comparative genomics was written for early researchers (advanced undergraduate students, postgraduates, and postdoctoral fellows). Well-established biologists should leave it alone—it is not intended to impress them.

What is comparative genomics? Before a proper definition can be put forward, we need to recognize that a genome has many primary features such as the genomic sequence, strand asymmetry, genes, gene order, regulatory motifs, and genomic structural landmarks that can be recognized or modified by cellular components with functional implications, etc. A genome also has secondary features such as the dynamic transcriptome, proteome, codon–anticodon adaptation, functional association of genes, and gene interaction networks. Comparative genomics is a branch of genomics that aims to (1) characterize the similarity and differences in genomic features and trace their gain and loss along different evolutionary lineages, (2) understand the evolutionary forces such as mutation and selection that govern the changes of these genomic features, and (3) find out how genomic evolution can help us battle diseases, restore environmental health, make money, etc.

It is better to illustrate this with an example. Suppose we have a set of bacterial genomes, with Genome A missing genes for lactose metabolism in contrast to all closely related genomes that still carry the genes. We may reasonably infer that the genes were lost in the lineage leading to Genome A. Suppose we further find that the organism carrying Genome A has inhabited an environment that is constantly lactose-free (I, as well as some of my Chinese, Finnish and German colleagues, would love to have such an environment), then we can infer that genetic alterations to the lactose-metabolizing genes are essentially neutral for the carrier of Genome A, with no functional consequence for losing the gene. Through a phylogeny-based analysis, we may find that lactose-free environment is strongly associated with the loss of lactose-metabolizing genes. If we further find that the set of genes are either strongly conserved in evolutionary lineages requiring lactose metabolism or degraded by accumulated mutations in those living in lactose-free environment, we can infer that the genes are strongly associated only for the lactose-metabolizing function. In contrast, if we find that the set of genes are still strongly conserved in lineages inhabiting lactose-free environment for a long time, then the genes may have functions other than lactose metabolism.

What basic knowledge do we need to do research in comparative genomics? The most fundamental feature of a single genome is its nucleotide sequence, and the most fundamental feature shared among a set of genomes is coancestry, or shared homology. These immediately bring into our mind the necessity of sequence-related computational tools such as sequence alignment and molecular phylogeny. For this reason, some literacy in computation and mathematics/statistics is assumed.

Much of the comparative genomics is done by genomic comparison against genomes of model organisms. Consequently, it is of tremendous value to gain a good understanding of molecular biology of some model organisms such as *Escherichia coli*, *Bacillus subtilis*, *Mycoplasma genitalium*, *Chlamydomonas reinhardtii*, *Arabidopsis thaliana*, *Saccharomyces cerevisiae*, *Caenorhabditis elegans*, *Drosophila melanogaster*, *Ciona intestinalis*, *Danio rerio*, *Takifugu rubripes*, *Xenopus laevis*, *Gallus gallus*, *Mus musculus* and, of course, *Homo sapiens*. For an evolutionary biologist, it is a great comfort to see such a diverse array of model organisms, especially for those who have lived through the bygone era dominated by the dogmatic assertion that "What is true in *E. coli* is also true in the elephant".

What about viruses? Can one do research in comparative genomics of viral genomes? The main difficulty with viral genomes is that viral lineages are often so diverse that they do not share any detectable homology. So comparative genomics is typically limited to closely related lineages such as among different subtypes of influenza viruses or among HIV/SIV viruses. However, lack of homology does not preclude one extremely important aspect of evolutionary studies, i.e., the study of convergent evolution. Diverse bacteriophage lineages can parasitize the same host and serve as a fertile ground for studying convergent evolution in response to the same intracellular environment of the host. However, it is the demonstration of functional equivalence, instead of homology, of the genes that is at the center of lime light in the study of convergent evolution in comparative viral genomics.

Comparative genomic research should be guided by the conceptual framework of evolutionary biology, so readers are assumed to have read something Darwinian. There are two most fundamental problems in evolutionary biology. The first is the origin and maintenance of new features and new species. There is no better way to address this question than comparative genomics, where the gain and loss of functional genes, as well as modification of a gene to gain a new function, can often be unequivocally identified from a set of related genomes. Many bacterial species are competent in pick up environmental DNA segments and integrate them into their genomes. Some of these DNA segments contain functional genes, leading to inheritance of the newly "acquired characters" and changes in subsequent evolutionary trajectories.

The second fundamental problem in evolutionary biology is the establishment of the links among genotype, phenotype, and environment. The greatest stumbling block to this line of enquiry has been the characterization of the genotype. This block is essentially non-existent when we have all the genomes and can characterize various aspects of the genotype, e.g., the presence/absence of a set of genes. We can then use phylogeny-based methods to systematically characterize the

association between this matrix of genotypes and the matrix of phenotypes or the matrix of environmental factors.

The diverse genomes we see today did not originate independently, but represent products of descent with modification. This has fundamental implications on the methodology in comparative genomics. A good phylogeny is typically required for any comparative genomic study involving more than two genomes. The reader is therefore assumed to have gained basic understanding of phylogenetics.

Many examples of comparative genomic research are illustrated throughout the book. The first chapter includes many small-scale research examples, while the second chapter is heavy with large-scale studies and their associated statistical methods, in particular the comparative methods involving both continuous and discrete variables. The effort to develop phylogeny-based comparative methods was initiated by Joe Felsenstein and subsequently further developed and promoted by Paul Harvey and Mark Pagel. I numerically illustrated these methods in such a way that researchers with basic statistical and programing skills can include these methods in their programs. It should also facilitate further development of the methods by people well-versed in stochastic processes. The third chapter presents frequently used methods for detecting viral recombination.

The comparative approach has gone way beyond biology. For example, social scientists have characterized "phenotypes" of different forms of government and how much of the "phenotypic" differences can be attributed to historical inertia and environmental and cultural determinants. From a social biogeographic point of view, there are two possibilities for why Government Form A (GF_A) is found in Area X but GF_B is found in Area Y. First, GF_A is "good" for people in Area X and "bad" for people in Area Y. Likewise, GF_B is "good" for people in Area Y but "bad" for people in Area X. In this case, we should leave these people alone. Second, GF_A is "better" than GF_B in both areas but has never got a chance to be practised by people in Area Y. In this case, we might try to persuade people in Area Y to practise GF_A. Phylogeny-based methods can help us discriminate between the two possibilities, although some politicians and religious leaders have long settled for the second possibility, i.e., one particular GF or religion is better than all alternatives and should be promoted and practised everywhere in the world.

This book is not on democracy or religion, and is not good for everyone. In fact, book authors universally acknowledge the truth that a book is never good for everyone. For this reason, many authors are profusely apologetic in the preface, although there are also a few courageous ones who simply stated "Please read the book". I do not want to be apologetic and obviously do not want to draw reader's attention to problems in my book, but feel that I have to list a few things below just to conform to the convention.

First, this book does not cover all aspects of comparative genomics. In particular, it does not cover any aspect of genome rearrangement, for three reasons. First, many books entitled "Comparative Genomics" include extensive coverage of genome rearrangement. Second, most genes in eukaryotes and operons in prokaryotes appear to function well without being constrained by their location in

the genome. Third, I myself do not work on genome rearrangement, which is my strongest justification for the omission. I do not think that anyone wants to read a professional book, or even part of it, written by a layperson.

Second, do not be infuriated when you find your important works not cited in the book because this book has a mandate to be brief. If you keep up your good work, readers of the book will discover you sooner or later. You would be a modern Mendel if you get rediscovered by three separate investigators, which perhaps is not a bad thing after all.

Third, I am a Chinese, and English is not my mother tongue. If you come across a grammatical error, please do not immediately shred the book or angrily demand refund. Let me see if I can squeeze a smile out of you by sharing a little story of me. The textbook of English during my undergraduate years in China typically had a list of new English words/phrases and their Chinese equivalents side by side. "Should" and "to be supposed to" happened to have the same Chinese equivalent that means "should", and I had since considered "should" and "to be supposed to" as synonymous. Then there came a time when I was doing my graduate research in a field station with a group of other Canadian students. I typically would wash dishes because others did the cooking which took much more time and energy. Once my fellow students suggested that I should share the dishwashing with others, and I wanted to say "I should wash the dishes" because others did the cooking. But then I thought that "to be supposed to" seemed much more grandiose than the plain "should". So I replied that "I am supposed to wash the dishes", privately thinking that they would be really impressed by my command of English. The resulting behaviour of my Canadian fellow students puzzled me for a whole field season, and I wrote home that "culture shock" was so real and that Canadians could truly be weird and unpredictable.

I hope that this book will not create many "weird and unpredictable" readers.

Acknowledgments

An experienced publisher once pointed me to a few examples of “effective use” of acknowledgment, each with an impressive list of well-known scientists, tactfully acknowledged to boost the reputation of the book author. The practice reminded me of some recent scientific conferences each with a list of 8–11 Nobel laureates as session chairs or keynote speakers. A journal would not have legitimacy if it does not have a list of silverbacks in the editorial board, even though some of the silverbacks are never involved in the manuscript-screening process. A person’s worth is often evaluated by the number of “like” in social networks. We are entering a world in which a masterpiece in art is no longer evaluated on its own merit, but on whether it features gold-plated frame or displayed in a prominent location in a museum or gallery!

Should I mould a few famous names into a gold-plated frame for my limited painting of comparative genomics? I did have the good fortune of being associated with a number of silverbacks. Some helped me to switch to molecular evolution and phylogenetics when I was forced to switch fields because of severe allergies toward rodents that I used to study. Some offered me their books as gifts that inspired me and cultivated in my mind a strong desire to produce something similar. Some donated their previous field data or bacterial strains that led to results included in this book. Some have commented much of the book and corrected errors in the second chapter of this book. However, there are also little known people, but much greater in number, who have helped me and supported me in various ways during the writing process. If the “effective use” of Acknowledgement implies the exclusion of little known names, then let me engrave all these names in my heart without mentioning any here. I think that they would all like it this way.

But some explicit acknowledgments are absolutely essential—there would be serious repercussions if I did not. Scientists, just as religious monks, need patronage to carry out their daily routines and rituals. Without generous patronage, there would be neither religious freedom nor academic freedom. So here goes my acknowledgment to funding agencies: NSERC (Discovery Grant) and CAS/SAFEA (International Partnership Program for Creative Research Teams). While the money has never been sufficient for research, it is perhaps worth as much as a gold-plated frame for decorating the book.

I should also thank Evelyn Best who encouraged me to write this book as an expansion of a previous book chapter. I was initially reluctant because the word "expansion" reminds me of software bloating. To paraphrase Joe Armstrong (creator of Erlang), when a reader asks for only a banana, should I give him a gorilla holding a banana or even an entire jungle? However, I soon realized that the banana alone does not make a healthy meal. Hence this book, with some additional berries, but no gorilla or jungle in it.

My limited command of the English language becomes particularly acute when I come to express my appreciation for my wife (Zheng) and my children. They are a miracle to me. The arrow of time has brought so much wonderful transformation to our little ones and created so many memorable moments. By just looking at them, I am convinced that the world after me will be much nicer, gentler, and smarter. May they grow up and enjoy reading this book!

Contents