DAN GUSFIELD

RECOMBINATORICS

THE ALGORITHMICS OF ANCESTRAL RECOMBINATION GRAPHS AND EXPLICIT PHYLOGENETIC NETWORKS

ReCombinatorics

Dan Gusfield

with contributions from

Charles H. Langley
Yun S. Song and
Yufeng Wu

ReCombinatorics

The Algorithmics of Ancestral Recombination Graphs and Explicit Phylogenetic Networks

The MIT Press
Cambridge, Massachusetts London, England

This book was set in Times Roman using LaTeX by TEXnology Inc. and was printed and bound in the United States of America.

Library of Congress Cataloging-in-Publication Data

Gusfield, Dan.
ReCombinatorics : the algorithmics of ancestral recombination graphs and explicit phylogenetic networks / Dan Gusfield.
p. cm
Includes bibliographical references and index.
ISBN 978-0-262-02752-6 (hardcover : alk. paper)
1. Genetic recombination. 2. Evolution (Biology) I. Title. II. Title: Re combinatorics.
QH443.G87 2014
572.8'77dc23
2013043434

10 9 8 7 6 5 4 3 2 1

I dedicate this book
to the memory of my mother
Irma Gusfield,
and to the future of my daughters
Talia and Shira

Contents

Preface

Where This Book Came From

Early history: It started with haplotyping

About fourteen years ago, Chuck Langley (population geneticist in the Department of Evolution and Ecology at UC Davis) asked me to look at a ten-year-old paper [71] on an algorithm to solve the *Haplotyping* problem (discussed in chapter 12). That seminal paper, written by Andrew Clark in 1990, was ahead of its time because the genetic data that the algorithm was designed to analyze was not yet plentiful. But by the year 2000, with the introduction of widespread, high-throughput genotyping, and later sequencing, the data flood had begun, and computational haplotyping was becoming a central issue in analyzing genotypic data for population genetics. Digging into the 1990 paper, and then developing new ideas based on it, I began a focus on computational problems in *population genetics*, and also began a continuing collaboration and conversation with Chuck. After we obtained a joint NSF grant, additional graduate students, postdocs, and visitors joined our group at UC Davis, for differing periods of time.[1]

Next came haplotype blocks and haplotyping in blocks

The announcement in 2001 of the identification of *haplotype blocks* in humans (discussed in chapter 12), motivated the next question: Finding haplotyping solutions that conform to

1 In chronological order of participation, the other people involved at UC Davis on topics in this book were: John Kececioglu (as a postdoc before this history, and later on sabbatical), R. H. Chung, Yelena Frid, Satish Eddhu, Vladimir Filkov, Zhihong Ding, Dean Hickerson, Yun S. Song, Yufeng Wu, Dan Brown (visiting from Waterloo), Fumei Lam, Simone Linz, Rob Gysel, Kristian Stevens, Balaji Venkatachalam, and Michael Coulombe. Collaborators from outside of UC Davis were V. Bafna, V. Bansal, S. Hannenhalli, G. Lancia, S. Orzack, and S. Yooseph.

the *perfect-phylogeny* **tree** model. That problem is called the *perfect-phylogeny haplotyping (PPH) problem*, and it is also discussed in chapter 12. The PPH problem is a beautiful algorithmic and combinatorial problem that at first sounds artificial but actually models biological reality in the appropriate contexts. Its solution partly uses deep combinatorial mathematics that goes back to the 1930s. Rarely does one encounter such a well-structured combinatorial optimization problem, with beautiful, efficient algorithmic solutions, where the problem captures a real application without the need for artificial assumptions. But still, the perfect-phylogeny tree model, which involves no recombination, needed to be extended for even greater genomic application. Somehow, recombination had to be added into the model.

Which led to constrained recombination and galled trees

In a fortuitous development, a seminal paper [426] had appeared a couple of years earlier, introducing an extension of the perfect-phylogeny tree model to phylogenetic *networks* with constrained recombination. That network model, later called the "galled tree" model (discussed in chapter 8), seemed appropriate for a *"modest"* level of recombination. Chuck was particularly interested in this model, explaining that we want to find regions in the genome where some amount of recombination has occurred, but not a huge amount. Recombination causes variation in genomes, which can be exploited to connect what appears in the genome (*genotypes*) to important genetic traits (*phenotypes*) in an individual. This allows one to identify genomic loci that influence genetic traits, and to better understand details (and maybe the mechanisms) of those genetic influences. So, some recombination is needed, but too much recombination obscures the signal, making the analysis too hard. Therefore, an algorithm that finds regions of *modest*, but not zero, recombination is of interest. Hence, with graduate student S. Eddhu, we dug into the network paper [426], leading to new ideas and methods on galled trees. From that introduction to networks with recombination, my focus shifted sharply to problems of phylogenetic *networks with recombination*, of the type that arise in population genetics.

The next development: Simon Myers's thesis, and Yun and Jotun's constructions

The next turning point was the publication of the paper [302] by S. Myers and R. Griffiths in 2003 (based on the Oxford University dissertation of Myers [299]), on computing *lower bounds* on the number of recombination events needed in an *ancestral recombination graph (ARG)* (formally defined in chapter 3), to create a given set of binary sequences. This paper was the first real advance on that question since 1985, and it pointed to several different research directions. The lower bounds in Myers's thesis will be discussed in chapters 5 and 10. At about the same time as Myers and Griffiths were working on lower bounds, Yun Song and Jotun Hein (also at Oxford) were developing algorithms to explicitly *build*

networks (ancestral recombination graphs) that *construct* a given set of binary sequences, minimizing the number of recombination events in the network. Those methods continued the earlier work of Jotun's a decade before. That work will be partly discussed in chapter 9.

Shortly thereafter, Yun joined our group in Davis as a postdoctoral student (supervised jointly by myself and Chuck), and Yufeng Wu and Zhihong Ding joined as Ph.D. students. This core group (along with some of the participants mentioned earlier), focused almost exclusively on problems involving networks and recombination, motivated by issues in population genomics. In our first summer together, we organized a readings seminar on coalescent theory, reading papers and the newly published book on gene genealogies and coalescent theory [168] by Hein, Schierup, and Wiuf. From that seminar, we wrote out a series of open questions that moved from the most specific to the most general, where an answer to any question would also answer the prior ones. Additional nonserial questions were also enumerated.

Middle history: The growth of phylogenetic networks

Before continuing the story of how this book came about, I digress to say something of the history of the field of phylogenetic networks. In my view, the field of phylogenetic networks has three early sources.[2] It was initially identified with the seminal work on *median networks* and *splits decomposition* and associated ideas, developed by Hans J. Bandelt, and Andreas Dress and their students, starting in the early 1990s. Those networks are now often called "data display" networks, because they represent patterns of incompatibility in data but do not try to tell an explicit story of the evolution of the data. The second source of the field actually began before the first one, but was only later considered part of the field of phylogenetic networks. That second source is the work of R. Hudson, P. Marjoram, and R. Griffiths, who defined models and networks (ancestral recombination graphs) building on coalescent theory, to explicitly represent the evolution of binary sequences through mutation and recombination. The third source of the field is the work of Jotun Hein, starting in 1990, exploring algorithms "reconstructing evolution of sequences subject to recombination." (Ancestral recombination graphs and the related networks of Jotun are now called "explicit" phylogenetic networks.)

The three sources that started in the late 1980s and early 1990s gave rise to the growth and the broadening of the field of phylogenetic networks about fifteen years later. "Who Is Who in Phylogenetic Networks" [121] shows the rapid growth of the field. The number of papers published in the years 2001 through 2005, were 10, 8, 12, 28, and 42,

2 Of course, there were other early contributors to the field, and I apologize to all those whose names I have omitted.

respectively, and 55 in 2012. More impressive, there has been a broadening of the lines of inquiry, and models, beyond the three sources. Many new problems and models have been examined (for example, *galled trees, level-k networks, cluster networks, hardwired and softwired reticulation networks, normal networks, tree-child networks*, and more). In the last decade there have also been several dissertations on phylogenetic networks, and two books published.

Additionally, there were (at least) six international meetings (that I was privileged to attend) that helped propel the field: The 2004 meeting on phylogenetic combinatorics at Uppsala University; the 2005 meeting on the mathematics of evolution and phylogeny, at the Henri Poincaré Institute in Paris; two meetings on mathematical methods in phylogenetics, sponsored by the Isaac Newton Institute at Cambridge University in 2007 and 2011; the 2009 meeting on algorithmics in human population genomics, at the Dimacs Center at Rutgers University; and most recently, a 2012 meeting sponsored by the Lorentz Institute at the University of Leiden, on the future of phylogenetic networks. These meetings brought people together from different parts of the field and from different parts of the world, and helped to define the broader field of phylogenetic networks and to create a more coherent community of phylogenetic network researchers.

Back to the story of the book

During that middle period, as the field blossomed, the work of our group at UC Davis also blossomed. In the four years after we wrote our list of questions, all but one of them were answered (by us and others). And a week before the book went to the publisher, a proposed answer to the last question was explicitly verified (see section 14.6.2).

Late history: The emerging book

With the maturation of the field, and with the graduation of Zhihong (now working at Adobe) and of Yufeng (now tenured at U Connecticut) and the departure of Yun (now tenured at UC Berkeley), it seemed that it was time to revisit the whole area, to write a book for a broad audience of computer scientists, mathematicians, and biologists, with several goals in mind.

The most concrete goal was to give a more scholarly, integrated, and unified exposition of computational issues involving ARGs; standardizing notation, adding many illustrations and examples, and simplifying, completing, correcting, and generalizing various proofs and algorithms. A book would allow a deeper treatment of various topics, and yet at a more leisurely pace than is possible in a journal publication, and it would allow new full expositions, and integration, of difficult material that had been developed by other researchers. The next goal was to tell the story of the series of results (ours and others) in which an

increasingly general understanding of combinatorics and algorithmics of ARGs was developed; but to tell it *backward*—that is, to first develop the most abstract and general results (in chapters 6 and 7), and then use that machinery to explain and prove more specific results (for example on galled trees), which chronologically had been developed before, and had led to, the more general results. The next goal was to identify, and make explicit, common ideas (mostly inspired by the coalescent theory *viewpoint* but not by any actual coalescent theory) that underlie many disparate methods to construct ARGs. This is done in the first sections of chapter 9. This goal also led to the unification of constructive and destructive methods (explained in chapter 9) and to the connection of ARG construction methods to the history lower-bound method (in chapter 10). An additional goal was to explain the use of ARGs in a variety of applications, for example in association mapping (discussed in chapters 4 and 13), and in the logic behind association mapping methods that *don't use* ARGs. Another goal was to show how ARGs fit into the larger field of phylogenetic networks, relating ARG problems and models to phylogenetic network problems and models that seem at first unrelated to ARGs.

Finally, the most general goal was to widen the biological focus beyond problems involving humans, and widen the methodological focus beyond population genomics. In particular, to explicitly (and in several ways) make the point that although the biological contexts of population genetics and phylogenetics are very different, there are mathematical and algorithmic ideas that are common to both fields, where the biological differences do not matter. In fact, there are formal ways to *transform* certain problems and results in one domain to problems and results in the other. This point is made throughout the book (perhaps ad nauseam), but most explicitly in section 3.2.3.3 and in chapter 13.

In addition to these goals, the envisioned book would illustrate the importance of recombination in solving biological problems (what I sometimes call the bio-*logical* importance of recombination), in addition to the biological importance of recombination. This is done most explicitly in chapter 4. And the book would be a vehicle to introduce and explain the utility and versatility of computational techniques (most notably, integer linear programming and dynamic programming) to a broad audience, some of whose members may not have been exposed to those techniques. Most importantly, the envisioned book would help shape the research and education agendas of the computational biology community, and enable and encourage people outside the community to enter the field.

So, having decided that there should be a book on the combinatorial structure and algorithmics of problems defined on ARGs, and how ARGs relate to other phylogenetic networks, I took a yearlong sabbatical starting July 2008, sure that everything would be finished by its end, in September 2009. Well, four years later, in October 2013 (after putting

in over 2,800 "billable hours" and typing over 1.7 million keystrokes) the book is almost done.[3]

The Title: ReCombinatorics

It is a *portmanteau* word[4] derived from the single-crossover recombination of the words "recombination" and "combinatorics":

r	e	c	o	m	b	i	n	a	t	i	o	n		
-	-	-	-	-	-	-	-	-	-					
		c	o	m	b	i	n	a	t	o	r	i	c	s
										-	-	-	-	-

Independently, other word-playful people [202] hit on "Recombinatorics" as the natural word for this field.

This Is Not a How-To Book

This book is about *ideas, models, and methods*. It is not a how-to book. I am positive that most of the general ideas exposed in this book will have productive uses or productive progeny, long after current software can no longer be compiled, let alone be executed. Moreover, I also believe that understanding the ideas and models that underly methods is helpful, if not essential, to applying the methods most effectively.

A Comment on Empirical Testing and Software

In several sections of the book, specific programs are discussed and some empirical results are mentioned, obtained from simulations or from executions on biological data, using those programs. The main purpose in mentioning programs and empirical results is to

3 Perhaps like giving birth to a child, it's good we quickly forget how hard it is or we would never do it again.

4 Lewis Carroll in Through the Looking Glass introduced this term: "Well, 'slithy' means 'lithe' and 'slimy' ... You see it's like a portmanteau – there are two meanings packed up into one word."

establish that *ideas* discussed in the book have been implemented, and that some of them work (some better than others). The empirical results provide very *crude* indications of the practicality of the methods. I am generally very skeptical and uninterested in detailed empirical results on program times, and to some extent accuracy. I use empirical results to answer in a broad-brush way, whether or not a method underlying a program is ballpark practical for some likely data of interest. So in this book, I make no effort to provide an in-depth comparison of software speed and reliability. If you have data, and there are choices for the appropriate programs, try them out. However, you can find links to the programs mentioned in this book at www.cs.ucdavis.edu/~gusfield/recsoftware.

A Note about Citations

There are expositions of established material that are new in this book, but regardless of the origin of the exposition, the book will cite the original source of any result that is not new to the book. For such results, please be sure to cite the *original* authors and the *original* publications, rather than citing only this book. If the material is original to this book, or you wish to direct a reader to this book for a new exposition of established material, I welcome that, but please cite the original papers as well. It is good scholarship to do so, and it shows proper respect for the original authors.

Acknowledgments

There are so many people to acknowledge and to thank. First are all of my research collaborators mentioned earlier. Foremost among those are Chuck Langley, Yun Song, and Yufeng Wu. In addition to their critical, central contributions in the development of many of the results that are re-exposed in this book, they watched over me during the writing of the book, providing wisdom and encouragement. I am particularly indebted to Yufeng, who provided several new ideas that are discussed in the book. Of course, any mistakes in the book are my own.

Additionally, I want to thank (in completely random order) Jane Gitscher, Jotun Hein, John Wakeley, Ladan Doroud, Laxmi Parida, Steven Kelk, Leo van Iersel, Celine Scornavacca, Chris Whidden, Charles Semple, Mike Steel, all the students in the fall 2011 UC Davis Computer Science course 224, Julia Matsieva, Tandy Warnow, Katherine St. John, Luay Nakheleh, Gabriel Valiente, Simon Myers, Bob Griffiths, Ken Burtis, Nick Shepard, Iain Mathieson, Mike Waterman, Eleazar Eskin, Eran Halperin, Dick Karp, Richard Durbin, Richard Mott, Sorin Istrail, Daniel Huson, David Morrison, Andy Clark, Ron Shamir, David Fernández-Baca, Martin Tompa, all the students in the 2012 International Winter School on Methods in Bioinformatics in Tarragona, Spain, Earl Barr, Katharina Huber, Vincent Moulton, Philippe Gambette, Sylvain Guillemot, Shawnie Briggs (who made the tree sculpture), and my neighbors Rob and Lacey Thayer (who own it).

I also want to thank the National Science Foundation for continuous support of our research on the combinatorics and algorithmics of phylogenetic networks with recombination. This was made possible through grants: SEI-BIO 0513910, CCF-0515378, IIS-0803564, and CCF-1017580. In particular, I thank the NSF program officers, Sylvia Spengler, Ding-Zhu Du, and Mitra Basu, for their trust, support, and encouragement. I thank the Simons Institute for the Theory of Computation which partially supported my efforts in the final phase of completing the book.

I thank Bob Prior at MIT Press for his trust in this project and persistence in bringing it to MIT. I thank the seven anonymous reviewers who made many helpful comments, and Virginia Crossman and Amy Hendrickson for their contributions to editing and layout.

Finally, I thank my wife, Carrie, for her love and support — although words do not suffice to thank her for understanding and tolerating the enormous amount of time I spent locked up in my upstairs office — rather than attending to house and family affairs. Thank you Carrie — I will never do it again (or maybe just a little book next time).

1 Introduction

All DNA is recombinant DNA. ... [The] natural process of recombination and mutation have acted throughout evolution. Genetic exchange works constantly to blend and rearrange chromosomes, most obviously during meiosis.
— James Watson [427]

Molecular phylogeneticists will have failed to find the 'true tree,' not because their methods are inadequate or because they have chosen the wrong genes, but because the history of life cannot properly be represented as a tree.
— Ford Doolittle [93]

Geneticists have long dreamed of determining the genetic basis of disease susceptibility by comparing variations in the human genome sequences of a large number of individuals.
— P.Y. Kwok [233]

1.1 Combinatorial Genomes and the Grand Challenge

Now that high-throughput genomic technologies are available for sequencing, resequencing, finding genomic variations of different types, finding conserved features, and screening for traits, the dream of comparing sequence variations at the population level is a reality. Moreover, population-scale sequence variations have numerous practical applications (as in association mapping) and can be used to address evolutionary and historical questions (such as the migration of populations), and to address basic questions concerning molecular genetic processes (as in the mechanics of mutation, recombination, and repair).

Nature and history, through point mutation, insertion and deletion, recombination, gene-conversion, genome rearrangement, lateral gene transfer, creation of mosaic cells and

genomes, retrotransposition, introgression, structural modification, migration and admixture of populations, random drift, selection, and many other operations, have conducted a vast number of multifactorial, combinatorial "experiments" in which DNA has been mixed and matched in different ways to create a huge variety of *combinatorial genomes* and *mosaic sequences* in the current population.[1] Some of those variants and combinations are easily *correlated* with observable organismal traits, some are very subtly correlated with traits, a minuscule proportion of the variants have been proved to be *causal* for traits, but for now, the vast majority of combinatorial variation seen in genomes only provides material for huge catalogs and databases. It may be noise, or it may be the key to curing (you fill in the blank).

The *grand challenge* is to exploit these natural, multi-factorial experiments by finding patterns in and among different genomes (i.e., **genotypes**) that have significant and biologically meaningful impact on important traits of interest (i.e., **phenotypes**). Addressing this grand challenge requires many new tools, among which are new ways to use *graph theory*, and *algorithms* to model, reason about, and compute evolutionary patterns in populations.[2]

1.2 Genealogical and Phylogenetic Networks

Genealogical and phylogenetic **networks** are graph-theoretic models of evolution that go beyond phylogenetic **trees**, the traditional representation of evolutionary history. Genealogical and phylogenetic networks incorporate *nontreelike* biological events such as *meiotic recombination* that occurs in *populations* of individuals inside a *single* species; or incorporate general *reticulation* events that occur between different species, caused

1 In their colloquial use, the phrases "mosaic genomes" and "chimeric genomes" convey the proper image of these mixed-up, combinatorial genomes. Unfortunately, the terms *chimeric genomes* and *mosaic genomes* already have well-established, narrower, technical meanings in genomics [262, 460]. So I made up the term "combinatorial genome" instead. A "mosaic genome" (as properly used in genomics) is then only one of the many kinds of combinatorial genomes that nature has created.

2 Of course, the grand challenge also requires new results in probability theory, in stochastic modeling, and statistical methods. Some will argue that these advances are more important than advances in combinatorial methods. But, there is a division of labor, and of talents, and this book is about combinatorial issues and advances. Moreover, methods for statistical *computation* often rely on combinatorial methods, for example, to efficiently enumerate, sample, and perturb structures, and to solve various optimization problems. Even in this book, we will see some interplay between combinatorial methods and statistical computation.

for example by *lateral gene transfer or hybrid speciation*. The central algorithmic problems are to reconstruct *plausible* histories, with mutations, treelike events, and nontreelike events that generate a given set of extant, observed genomic *sequences*; to determine the *minimum* number of such biological events needed to derive the sequences; to enumerate a range of plausible histories, and assess their biological fidelity; and to characterize properties of plausible or optimal histories.

This book primarily concerns combinatorial and algorithmic issues involved in reconstructing the evolutionary history of extant sequences observed in populations of diploid organisms (such as humans), where the sequences are generated by mutations and recombinations. However, many of the combinatorial and algorithmic results apply equally well at the *phylogenetic* level, namely to reticulate evolution of *species*, rather than populations, and we will point these out when they arise. Indeed, one of the goals of this book is to expose common mathematical and algorithmic structure[3] that occurs both in populations and species, despite the differences in biological origin, and differences in the biological communities that study the two areas.

The book is aimed broadly at computer scientists, mathematicians, and biologists. We will explain the various biological phenomena; the mathematical, population genetic, and phylogenetic models that capture the essential elements of those phenomena; the resulting combinatorial and algorithmic problems that derive from those models, and from biological questions that are formulated in terms of those models; the theoretical results (both combinatorial and algorithmic) that have been obtained; related software that has been developed; and the results of empirical testing of that software on simulated and real biological data. In addition, we will explain some needed combinatorial and algorithmic background for those readers who might not be familiar with particular existing results or techniques. We begin with some essential definitions.

Definition A *chromosome* is a single linear molecule consisting of double-stranded DNA. An individual's genes are arranged on their chromosomes.

Definition A *locus* refers to a discrete, specifiable interval of sites or positions in a chromosome. The set of *sequences* that can occur at a particular locus specifies the set of *states* or *alleles* of the locus. The plural of *locus* is *loci*.

The word "specifiable" is part of the definition to emphasize the fact that we don't always know where the locus is. It has a specifiable location, but it might not be known.

3 "Poetry is the art of giving different names to the same thing; mathematics is the art of giving the same name to different things." This misquotation is attributed to Henri Poincaré.

> An *allele* is one of a number of alternative forms of the same gene or same genetic locus. (from Wikipedia)

Definition A *diploid* organism (such as a human) has two (not necessarily identical) "copies" of each chromosome. The two copies are called *homologs* or *homologous chromosomes*, and they form a *homologous pair*.

> Homologous chromosomes are similar but not identical. Each carries the same genes in the same order, but the alleles at each site might not be the same. (from Wikipedia, again; somewhat modified)

For example, humans have 22 homologous pairs of (autosomal) chromosomes, and one pair of sex chromosomes (X, Y in males, and X, X in females).

Definition The corresponding sequences on two homologous chromosomes are called *homologous sequences*.

1.2.1 Recombination and Genealogical Networks

Meiotic recombination

The best-known biological event that creates variation in genomes is a *point mutation* where a single nucleotide changes state, say from A to one of the other three states T, C, or G. However, mutations (that don't quickly die out) are relatively rare events, and in short time periods (even thousands of years in humans) mutations are not the primary cause of variation in genomes. Instead, **meiotic recombination** during meiosis is the key biological event that creates high-variation genomes over relatively short time periods in human (and other diploid) *populations* (i.e., individuals in a single species).

Meiosis is the process in which a *gamete* (egg or sperm), containing one copy of each chromosome, is created from a cell that has a homologous pair of each chromosome. In meiosis, recombination uses a pair of homologous chromosomes to create two *recombinant* chromosomes consisting of alternating segments (usually a small number) of the two homologs (see figure 1.1). Any child of that individual then inherits *one* of the resulting recombinant chromosomes. Similarly, recombination between two homologous chromosomes in the other parent creates two recombinant chromosomes, one of which is passed down to the child. Hence, the child receives one chromosome from their mother and one from their father, and each chromosome is a recombinant chromosome created from two homologous chromosomes of one parent.

Figure 1.1 Meiotic recombination of two sequences creates a third sequence, called a *recombinant* sequence. The recombinant sequence is created from the boxed segments of the two parental sequences. This example illustrates *single-crossover* recombination. Double and multiple-crossover recombination will be introduced later.

The key observation

Because of recombination in all the prior generations, it follows that the genome that any individual inherits is a mixture and a reflection of the DNA of *all* of the individual's ancestors. In this way, meiotic recombination allows the rapid creation of *chimeric* chromosomes even without mutations. This ability to create new chromosome sequences allows species to rapidly respond to changes in the environment, and to drive out deleterious mutations. Recombination is therefore an important adaptive property that occurs (along with sexual reproduction that enables it) in almost all eukaryotic species. The existence of such rich combinatorial genomes compels the study of genomic variation in populations to discover relationships between genome content and genetically influenced traits of interest.

To a computer scientist, it is almost irresistible to view combinatorial genomes, and the associated observed traits (phenotypes), as nature's way of implementing a kind of *binary search* (or a similar *divide and conquer* method) to identify the locations of important genomic features. In that view, nature has already done the experiments, posing and answering most of the required search queries. Now, in a kind of *Genomic Jeopardy* game, we have to find the right *questions* to match and exploit nature's *answers*.

1.2.2 Why Networks?

We are interested in reconstructing plausible histories of mutations and recombinations that might have derived chromosomal sequences observed in current populations. Such histories are not in the form of trees, but rather in the form of *networks*. To explain the need for networks in describing the history of chromosome sequences, we consider first the related, but simpler, issue of family *pedigrees*.

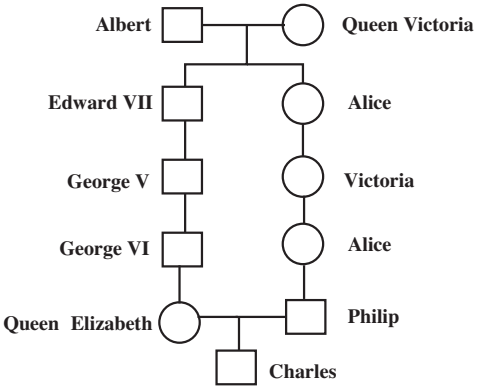


Figure 1.2 A partial pedigree of Prince Charles, who was married to Princess Diana, and is the father of Princes William and Harry (not shown). Note that the three females (Alice, Victoria, and Alice) drawn opposite to Edward and the two Georges, are not their wives. Also, we only show the ancestors of Charles who are descendants of Albert and Queen Victoria. Prince Charles had two parents (Elizabeth and Philip), who each had two parents, who each had two parents, etc. Following five generations back in time, Charles would have $2^5 = 16$ ancestors in that generation, if they were all distinct. But in fact, Charles does not have sixteen ancestors in the fifth generation back from him. In the fifth generation back from Charles, there are two *convergence* events: Edward and Alice converge at a common father, Albert, and they converge at a common mother, Victoria. That is, Edward and Alice are full siblings with the same parents. This part of the pedigree forms a *cycle* or *loop*. Note that the pedigree would also have had a cycle if Edward and Alice had only been half-siblings, i.e., if they had only shared a single parent. No parents of Albert or Queen Victoria are shown, and hence they are the *founders* of this partial pedigree.

1.2.2.1 Pedigrees

If we trace the ancestry of an individual *backward* in time, their two parental lines will expand into multiple lines (as parents expand to grandparents and great-grandparents, etc.). But some lines will eventually "converge," meaning that two distinct ancestors of the individual will have one or two common parents.[4] See figure 1.2 showing a partial pedigree of recent English royalty. It follows that the full genealogical history, or *pedigree*, of a set of individuals will contain *cycles* (often called "loops" in the genetics literature) and therefore *cannot* be represented by a *tree*; instead, the representation requires a **network**. We will develop precise definitions later in this chapter and in chapter 3.

4 A convergence event in a pedigree is sometimes called a *coalescent* event, but we will reserve that term for an event involving sequences.

1.2.2.2 Back to Sequences

Above, we considered family pedigrees in order to introduce the notions of *convergence*, *cycles*, and *networks*, and in order to distinguish pedigrees from *genealogical networks* which we will introduce shortly. But our main interest is the history of DNA *sequences*, and not the history of families or the individuals who carry those sequences. So, we now shift attention back to sequences on chromosomes. We first consider the case where there is *no* recombination and *no* mutation.

Recall that a diploid individual receives one copy of a particular chromosome (say chromosome 21) from the individual's mother and receives one copy from the individual's father. Moreover, without recombination or mutation, the copy received from the individual's mother is identical to one of the mother's homologs, and the same is true for the copy received from their father. For example, in figure 1.3, two four-character sequences are shown; these are sequences from the same location in a homologous pair of chromosomes.

Tracing the transmission history of sequences
Now consider an individual in a population and just *one* of the homologs of a homologous pair of some chromosome (say from chromosome 21). The sequence on that homolog was transmitted to the individual from just one parent, and since there is no recombination, that sequence was passed down from exactly one grandparent, etc. It follows that the transmission history of the chromosome sequence forms a *path* through ancestors in the individual's pedigree. That path begins at a founder sequence and descends to the individual. Further, since we have assumed that there is no mutation, each of the individual's ancestors on the path possesses and transmits an identical copy of the sequence.

Definition The path through sequences (overlayed on a pedigree), showing the transmission of a sequence from an ancestor of an individual, to that individual, is called a *sequence-transmission path*. Note that the elements on the path are sequences, rather than the people who possess those sequences.

For example, in figure 1.3, Charles's CCCC *must have* been transmitted from his father, Philip, even though his mother, Queen Elizabeth, also has CCCC. We deduce this because Charles also has TAAT, which he could only have gotten from his mother. Next, we deduce that Philip must have received CCCC from his mother, Alice. We deduce this because Alice must either have passed CCCC or GCTA to Philip, but he does not have sequence GCTA. Continuing with this logic, we deduce that Charles received his CCCC from Philip through a right-hand path, originating with Albert, his great-great-great-grandfather. The path tracing the transmission of CCCC from Albert to Charles is a *sequence-transmission*

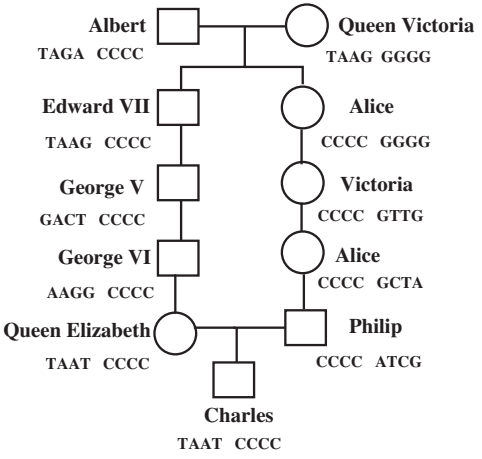


Figure 1.3 Each individual is shown with two fictitious sequences on two homologous chromosomes (say, from the two homologs of chromosome 21). For example, Charles's two sequences are TAAT and CCCC. Each individual receives one sequence from their mother and the other from their father. The order of the two sequences displayed for an individual does not indicate which came from the mother and which from the father. That information must be deduced, if possible. For example, we deduce that Elizabeth's copy of CCCC descended from a copy of CCCC possessed by Albert, via a sequence-transmission path through copies of CCCC possessed by Edward, and George V. Similarly, we deduce that Charles received CCCC from a copy possessed by Albert, via a sequence-transmission path containing sequences possessed by Alice, Victoria, Alice, and Philip. We see from this that the copy of CCCC possessed by Elizabeth is *identical by descent* with the copy of CCCC possessed by Philip. In fact, all of the copies of CCCC (other than Albert's) are identical by descent, from Albert. Looking backward in time, we say that the two *sequence-transmission* paths of the sequence CCCC possessed by Elizabeth and by Philip *coalesce* at the copy of CCCC possessed by Albert.

path. We can also deduce that Queen Elizabeth received her copy of CCCC along the left-hand path, a sequence-transmission path, also originating with Albert. Note that a sequence-transmission path traverses *sequences, not people*.

Definition If a sequence s is on a sequence-transmission path that leads (forward in time) to a sequence s' (where s might be identical to s'), then s is an *ancestral sequence* of s'.

Now, consider the *two* sequence-transmission paths of an *identical* sequence possessed by two individuals who share a common ancestor. For example, in figure 1.3, consider the sequence-transmission paths of the sequence CCCC, possessed by Elizabeth and Philip.

Definition If, traversing the sequence-transmission paths backward in time, two sequence-transmission paths intersect (at some *sequence*), then we say that the paths *coalesce* at that intersection point. This is also called a *coalescent event*.

Coalescence versus convergence

The two types of events are related, but a *coalescence event* (defined on sequences) implies a *convergence event* (defined on a pedigree), while a convergence event does not always imply a coalescence event. For example, in figure 1.4a, Edward VII and Alice (the upper one) have the same (two) parents (Albert and Queen Victoria), and that represents two convergence events in the pedigree. However, there is only one coalescence event. That event is when the two sequence-transmission paths that contain CCCC coalesce at the copy of CCCC possessed by Albert. Even though Edward and Alice have the same mother, Queen Victoria, they did not receive the same sequence from her (Edward received TAAG, and Alice received GGGG). So there is a convergence at Victoria, but not a coalescence.

Note that without mutations, all the sequences on two sequence-transmission paths that coalesce at some sequence s, must be identical to s.

Definition Without mutation, the sequences on two sequence-transmission paths that coalesce, are called *identical by descent*. See figure 1.3.

A history of sequences

To emphasize that our interest is in the history of *sequences* and sequence *transmissions*, and *not* in the history of the *people* through whom the sequences flow, in figure 1.4a, we redraw the sequence-transmission paths from figure 1.3, removing any sequence that is not on a deduced sequence-transmission path. However, we maintain the names of the people for ease of reference.

The underlying trees

When there is no recombination, the key feature of a set of sequence-transmission paths is that *no* sequence has two edges directed into it. There can be two sequences possessed by the same individual, where each sequence has a directed edge into it, but those two edges are directed to *different* sequences. For example, in figure 1.4 there are two edges directed into the sequences possessed by Edward, but one is directed to TAAG and one is directed to CCCC. If, in a set of sequence-transmission paths, we replace each sequence with a node, the resulting graph will not have any node with two edges directed into it. Hence, the graph will consist of a set of rooted *trees* that *partition* the sequences (each sequence is in one and only one tree). See figure 1.4b.

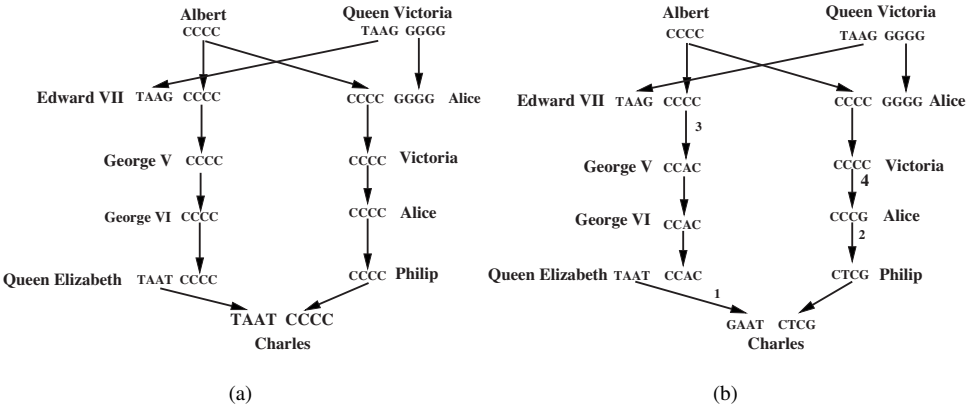


Figure 1.4 (a) The deduced sequence-transmission paths from figure 1.3. The sequence-transmission paths form four disjoint trees that partition the sequences. One tree consists of the two copies of TAAT; a second tree consists of the two copies of TAAG; a third tree consists of two copies of GGGG; and the fourth tree consists of all of the copies of CCCC.
(b) Now we allow sequences to mutate, but there is still no recombination. The number written on an edge identifies the site that mutates on that edge; the actual mutation is seen by comparing the nucleotide for that site at the head and tail of the edge. For example, on the edge with label 3, the nucleotide at site 3 changes from C to A.

Finding Adam and Eve in sequences

If we sample *sequences* in the current population that were transmitted without recombination, and we could trace their transmission paths back in time, we should eventually reach a point where all of the transmission paths coalesce (not simultaneously) at one common ancestor. That is, all the sampled sequences descend from one ancestral sequence, and the transmission paths jointly form a tree. Of course, we can't trace back in time, but we can *estimate* that tree. This was recently done for human patrilineal lineages, using the Y chromosome [340]. Only males have a Y chromosome, and only a small segment of the Y chromosome recombines (with a segment of the X chromosome). Hence, sequences from most regions on the Y chromosome have descended without recombination. The most recent common ancestor of those sequences is called the "Y-chromosome Adam." Similarly, using *mitochondrial* DNA, which is mostly (if not exclusively) inherited from one's mother, a matrilineal tree was estimated. The most recent common ancestor of the sampled mitochondrial sequences is called the "mitochondrial Eve."

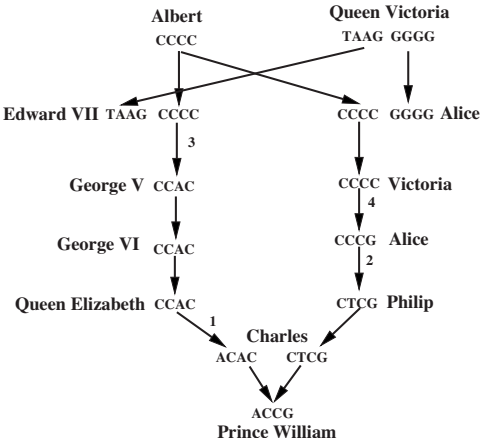


Figure 1.5 A history with recombination. The sequence, ACCG, that William receives from his father, Charles, is created by a recombination of the two sequences, ACAC and CTCG, transmitted to Charles from Elizabeth and Philip respectively. Note that sequence ACAC, transmitted to Charles from Elizabeth, is derived from Elizabeth's paternal sequence (rather than from her maternal sequence, as in figure 1.4). Also, it contains one mutated site (site 1), and hence is not identical to either of the sequences that Elizabeth possesses. The recombinant sequence ACCG contains the first two letters of ACAC and the last two letters of CTCG. William also receives a sequence from his mother, Diana, who is not shown in this partial pedigree.

1.2.2.3 Adding in Mutation and Recombination

Returning to the royals, we next add in mutation. Mutations do not change any sequence-transmission paths, but do change the sequences. So, it is no longer true that all the sequences on a sequence-transmission path are identical. For example, see figure 1.4*b*.

Finally, we add in recombination between two sequences. Then, when a sequence is transmitted to an individual from one of their parents, say their father, it might be a recombinant sequence created from the father's two sequences during meiosis. In that case, it could be different from both of the copies of the sequence that the father received. For example, in figure 1.5 we include one of Charles's sons, William, and see that the sequence William receives from Charles is a recombinant sequence, different from either of Charles's sequences.

Recombination creates cycles

In contrast to the case when there is no recombination, if a set of sequence-transmission paths contains a sequence with two incoming edges (due to recombination), then the

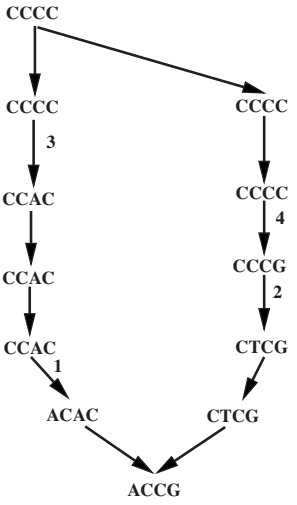


Figure 1.6 A recombination cycle. The two sequence-transmission paths in figure 1.5 form a *recombination cycle* with coalescent (or root) sequence CCCC and recombinant sequence ACCG.

transmission history is *not* a set of disjoint rooted trees. Moreover, as in figure 1.6, two sequence-transmission paths that meet at a recombinant sequence x might also coalesce at an ancestral sequence of x, say s. In that case, the two paths define a *recombination cycle* with *coalescent sequence* s (also called a *root* sequence), and *recombinant* sequence x. Further, if all the sequence-transmission paths ultimately coalesce at a *single* root sequence, then if there is any recombination, there must be a recombination cycle. For example, start at a recombinant sequence and trace back along two paths; since all paths eventually coalesce, the two traced paths must eventually coalesce, at which point the two paths specify a recombination cycle. A directed graph with a recombination cycle is not a directed tree; it is a **network**.

1.2.2.4 Transmission Paths Form (Parts of) a Genealogical Network

The network that represents the transmission history of chromosome sequences, shaped by mutation and recombination, is often called a *"phylogenetic network"* in the computer science literature, although the term *"genealogical network"* is more appropriate. So, the network shown in figure 1.4*b* is part of a genealogical network, as are the networks shown in figures 1.5 and 1.6. A first, informal definition of a genealogical network will be given

in section 1.6.2, after some concepts from graph theory have been introduced, and a more complete, formal definition will be given in chapter 3. When some additional restrictions apply (explained in chapter 3) a genealogical network is called an "*ancestral recombination graph*" (ARG). ARGs will be the networks of greatest importance in this book.

1.2.2.5 Genealogical Networks Relate Sequences, Not People

It is easy to confuse pedigrees and genealogical networks, so it is critical to note that genealogical networks represent relationships between *sequences*, rather than relationships between *people*, as in a pedigree. In fact, in most cases, we do not know the pedigree of the individuals who possess the sequences of interest. It is true that a genealogical network is constrained by that (unknown) pedigree, but a genealogical network displays information not contained in a pedigree, and the pedigree contains information not contained in the network. Moreover,

> Even though each individual possesses a pair of homologous chromosomes, the two homologous chromosomal sequences are represented *independently* in the genealogical network.

That is, the genealogical network (representing the history of sequences) *decouples* the two homologous sequences of each individual. In figure 1.6, the two sequences that Charles possesses, which recombine to create William's recombinant sequence, are drawn further apart than they are in figure 1.5, in order to emphasize that point.

> I may be a twin, but I am one of a kind.
> — Anonymous

Hence, the genealogical network represents a set of sequences in a history, without indicating which pairs of sequences (if any) were possessed by the same individual. The only way we can deduce that two sequences in a genealogical network were possessed by the same individual is if they recombine. Two sequences that recombine during meiosis must both be possessed by a single individual. As an example, figure 1.13 (page 31) shows a genealogical network which is unrelated to any shown pedigree. We can deduce that the sequences shown at nodes u and v must have been possessed by a single individual.

We want the historically correct genealogical network

Knowing the true, historically correct genealogical network that explicitly reveals the origin and derivation of the sequences in a current population, and shows the locations of all

the mutations and recombinations (both in the genome and in time), would tremendously facilitate the use of genomic data to address many basic biological questions, and be of use in biotechnology.

The rub

Unfortunately, we cannot directly examine the past so we cannot know (for sure) the historically correct genealogy of the extant sequences. However, a robust literature on *algorithms* that construct plausible genealogical networks, or deduce well-defined aspects of a genealogy, has developed, particularly in the last several years. Related questions about *hybridization networks*, which are similar to genealogical networks but do not generally involve explicit sequences, have also been addressed. This algorithmic research has been encouraged by a growing appreciation by biologists that many evolutionary and population genetic phenomena must be represented by networks rather than by trees.

1.3 The Central Thesis of the Book

Even though we can never know for sure that an algorithm has deduced the correct genealogical network (or features of it), we will detail in this book that applications of these algorithms have correctly answered certain biological questions, suggesting that important parts of true genealogies are captured in, or reflected by, the computations. These applications go to the heart of the book's central thesis.

> **Central Thesis:** *Explicit* genealogical networks representing a derivation of extant sequences in a population can be effectively computed, and even if those networks do not perfectly capture the true transmission history of the sequences, they can reveal parts of the history and give significant insight into basic biological phenomena, and be used to address applied problems in biotechnology.

1.4 Fundamental Definitions

The atomic objects of concern in this book are *individuals* in the context of population genetics, or *species* in the context of phylogenetics, or *molecular sequences* in the context of molecular evolution. Sometimes the particular biological context affects the mathematical models and the algorithmic problems that are defined on that model. However, many of the mathematical and algorithmic results we discuss in this book apply to all the biological models. We want to be as general as possible and so we will use a generic term for the objects of interest.

Definition We interchangeably use the terms *taxa* or *individuals* for the objects of interest, and *taxon* or *individual* for a single object.

This is intentional

There is a huge difference in the biological contexts in which the terms "taxon" and "individual" are typically used. The first term is used in phylogenetics, i.e., in the classification of *species, genus, family*, etc., and the second term refers to a single organism at the lowest level of the classification. The International Code of Zoological Nomenclature [320] defines a taxon as

> ... a population, or group of populations of organisms which are usually inferred to be phylogenetically related and which have characters in common which differentiate the unit (e.g., a geographic population, a genus, a family, an order) from other such units. A taxon encompasses all included taxa of lower rank and individual organisms.

The differences in size and time scales between taxa and individuals are huge. So why are we mixing the terms here? In particular, why use the term "taxon" at all, when the primary biological motivation for topics in this book comes from population genetics?

The answer is partly historical: the term "taxon" has become the default term in the phylogenetic networks literature for any unit of interest, while "individual" is an essential term when talking about individuals (duh) in populations. But perhaps a better reason for mixing the terms is that one of the key points we make in this book (repeatedly), is that many problems arising in vastly different biological contexts have the same (or related) mathematical and algorithmic structure.[5] That point is not well known, and barriers between the communities are very strong. So, the mixing is both due to laziness (following a crowd), and to be intentionally provocative—reflecting the effort to identify and exploit common structure. Similar justifications (or apologies) apply to other mixing of biological terms.

Definition A *character* or *trait* is a discrete property or characteristic of a taxon. By "discrete," we mean that there is a finite number of *states* (*alleles* in more biological vernacular) that a character can take on.[6]

For example, if the taxon of interest is a human, the gender (male or female) of the taxon is a *binary character* taking on one of *two* possible states. As another example, the

5 For example, in the way that multiple-crossover recombination can be used to model phylogenetic phenomena, even though recombination is not a phylogenetic phenomenon.

6 There are also continuous characters where the states are not discrete, but these are not of concern in this book.

nucleotide (A, T, C, or G) present at a particular site of a DNA sequence is a *four-state* character. The character would be binary if we only record whether the nucleotide at that site is a purine (A or G) or a pyrimidine (C or T). Another example of a binary character results from the phenomenon of DNA *inversions* where the orientation of a whole segment of DNA is inverted. Hence, the character is "segment orientation," which takes on two possible states. Similar, large-scale modifications of DNA (insertions, deletions, duplications, etc.) can give rise to additional binary characters, if those sequence modifications are independent, identifiable events [357].

Note that the meaning of the word "character" in the context of evolutionary biology is different from its colloquial meaning. In normal use, a character is a letter or a symbol in an alphabet, but in evolutionary biology a character is a trait of an individual. To confuse matters even more, a site in a DNA sequence can be considered to be a character in the sense of evolutionary biology, but the four possible states of that character are characters in the colloquial sense of the word, i.e., letters in the four-letter DNA alphabet.

1.4.1 Mutation, Infinite Sites, Perfect Characters, and Binary Sequences

Which mutations are modeled?

We introduced the concept of a *point mutation* in section 1.2.1, illustrated by a change in the state at a single DNA site (i.e., change of the nucleotide at that site). However, there are other changes that are biologically quite different from a single nucleotide change, but have characteristics that allow them to be modeled in the same way we model a nucleotide change at a single site. Hence, we want a more general definition to handle such changes.

Definition A *locus mutation* is a change of state at that locus, which is *independent* of changes at any other locus.

A point mutation is certainly a locus mutation. However, events such as DNA inversions, or DNA insertions and deletions, are locus mutations but not point mutations. Note that a change of state due to a locus mutation is distinct from a change of state due to recombination. Further, for a locus mutation to be a useful binary character, it must be short enough that it is unlikely to be split by a recombination event.

A notational conversion

In this book we are concerned with locus mutations that can be modeled as *binary characters*, that is, that have two permitted states, and are unlikely to be interrupted by recombination. As discussed before, there are several common locus mutations (e.g., point

mutations, inversions, insertions, deletions, duplications, etc.) that, under certain conditions, serve as useful binary characters. Most of the algorithmic and mathematical theory developed in this book is unaffected by which particular biological events give rise to the binary characters. Hence, we will often use the simpler term "mutation" in place of "locus mutation" or "point mutation," but will use the latter terms when the distinction is necessary. Similarly, we will generally use the term "site" instead of "locus," even though the underlying mutational event may involve an interval of sites.

A mutation model is needed

Genealogical networks represent the derivation of extant sequences which change due to mutation and recombination. Unrestricted, mutation events alone (without recombination) can derive any set of sequences, but that derivation would not likely reflect biological reality. Therefore, we need a *model* of the mutations that are permitted in sequence data.

1.4.1.1 The Infinite-Sites Model

> Essentially, all models are wrong, but some are useful.
> — George Box

The most commonly used mutation model in *population genetics* is the *infinite-sites model* where any locus (in the sample) mutates *at most once* in the entire history of the sequences. The infinite-sites model was introduced in the context of *point mutations*, but also has a clear meaning, and biological reality, for particular, more complex, locus mutations.

The infinite-sites model is justified when the probability of a mutation at any given locus is so low that the possibility of multiple mutations there can be ignored. For example, the infinite-sites model is justified when the history of a sample covers a *relatively short* time period, and mutations occur at random sites, so that a mutation at any given site is a low frequency event. Hence, the probability that a point mutation occurs twice at a site is extremely low.[7] The infinite-sites model is not appropriate for all evolutionary phenomena, but it is widely assumed in the population genetics literature, particularly for short time periods, and it is the basic mutational model for most of the results discussed in this book. See [168, 421] for more in-depth justifications of this model.

---

7 The origin of the term "infinite sites" comes from the view of a genome as having a huge (essentially infinite) number of sites so that each successive mutation, occurring at a random position in the genome, occurs at a site where no mutation has occurred before. It follows that a mutation at any given site can occur at most once, and that the character is binary.

> The number two ... is the geneticist's favorite number.
> — A. Knudson, quoted in [296]

The infinite-sites model implies that each character in any of the studied sequences can take on only one of *two* possible states: the *ancestral* and the *derived* states. Hence the sequences we observe in a population can be considered, and recoded as *binary* sequences. Note that the assumption of binary sequences does *not* come from assuming that the DNA (or other) alphabet has been reduced from four letters to two. The DNA alphabet still contains four letters, but in the set of DNA sequences found in a population, it is rare to observe more than two different ones (above a low frequency) at any given site. That is, if we fix a particular site and look at the individuals in the population to see what letters occur at that site, we rarely see more than two different letters (and generally, we only see one). A DNA site where we see *two* states in the population (above some minimum frequency) is called a *single-nucleotide polymorphism* (SNP, pronounced "snip") site.

Binary recoding
With the assumption that there will be at most two different states at any site, we can code each site as a binary character, using the alphabet $\{0, 1\}$ to represent the character states at the site, and so each sequence observed in the population becomes a binary sequence. It is common to use 0 for the most frequent of the two letters at that site (assumed to be the ancestral state), and 1 for the least frequent of the two letters (the assumed derived state). See figure 2.2 (page 38).

1.4.1.2 SNPs

The strongest empirical validation for the infinite-sites model comes from DNA sequence data in populations, and the *single-nucleotide polymorphism* (SNP) sites that are observed in the data. At a SNP site, only two of the four possible nucleotides appear in the population (with a frequency above some minimum threshold). In humans, and other well-studied organisms [116], millions of SNP sites have been found and cataloged, most prominently by the International HapMap Project [198, 199], the Human Genome Diversity Project [251], and the One Thousand Genomes Project [1, 2]. See also [62, 176] for additional empirical, population-genetic validation for the infinite-sites model.

1.4.1.3 Perfect Characters and Homoplasy

The infinite-sites model comes from population genetics. In contrast, phylogenetics concerns the history of *species*, rather than *populations*, and the term "infinite-sites" is not

used in that literature. However, phylogenetics does have a similar model. In phylogenetics a character that mutates only once in the history of the sample, has been called a *perfect character* [174, 351] or an *ideal* character [432]. Since it mutates only once, all descendants of the species where the character state changes, must also have that character state, and no other species will. Hence, at an abstract level, a perfect character in phylogenetics behaves the same as a binary character in population genetics that obeys the infinite-sites model.

> Taxonomists intuitively select character states which they postulate to define monophyletic[8] sets of species. The ideal character contains some state that both uniquely defines a set of species and has not been reversed in evolution, so that all existing species which possess this state can be said to have descended from one species in the past that evolved the state. For every such character state that can be identified, a branch in the phylogenetic tree can be added. [432]

Although the term "perfect character" originated in the phylogenetics community, we will use the term to refer to any character that obeys the infinite-sites model, whether it is a character used to study populations, or to study species.

Morphological support for perfect characters

The perfect character model in phylogenetics is supported by certain morphological data. A *morphological* character (such as having horns, or hair, or a tail) may be a trait that required many successive mutations. In that case, the probability is low that the trait will have evolved twice, independently in different species. Instead, a complex morphological trait that is common to several species is generally thought to have arisen only once, in a species that is ancestral to all of the species containing that trait. Further, once acquired, many complex traits will be retained in all descendant's species. This makes complex morphological traits ideal candidates for perfect characters in phylogenetics.

However, not all morphological traits are believed to be perfect (or near-perfect) characters, and it also believed that *convergent* or *parallel* independent evolution of highly valuable traits, such as flight or vision, has occurred. See [267] for a recently studied, exceptional case of convergent evolution. In phylogenetics, "homoplasy" is the term for violations of the perfect character model [364], due to recurrent or back mutation:

> Homoplasy is similarity that is the result not of simple ancestry, but of either reversal to an ancestral trait in a lineage or of independent evolution. [420]

---

8 A monophyletic set of species is a set that possess some character(s) in common that are not possessed by any individual outside the set. The term "clade" is a synonym.

Note that in phylogenetics, recombination is not considered a cause of homoplasy since recombination is not a phylogenetic-scale event. However, we will relax the original meaning of "homoplasy" and say that it is any violation of the perfect character model, regardless of the biological phenomena that causes it.

The phylogenetics community also uses the term *incongruence* for phylogenetic trees that don't agree with each other. Two characters are incongruent if they cannot be derived together on a single phylogeny without back or recurrent mutation.

Other perfect characters

Point mutations are the best-known and most widely studied mutations in DNA, but many other kinds of mutations also occur. The most common of these are *insertions*, *deletions*, *inversions*, or *duplications* of whole segments of DNA. When these kinds of locus mutations occur at low frequency, they generally obey the infinite-sites model, and can be considered perfect characters. For example, if a segment of DNA is inserted into some location in a chromosome of an individual, and the descendants of the individual inherit the augmented chromosome, and the same segment is never again inserted at the same location in another individual, and the segment is not changed by recombination, then the existence or non-existence of the segment at that location, is a binary character that obeys the infinite-sites assumption. Equivalently, it is a perfect character.

Specific kinds of mutations, other than point mutations, that might be used as perfect characters are the insertion of *micro-RNA* sequences [164, 201]; or insertions of mobile elements such as *SINEs* or *LINEs* [174, 314, 351]; or increases in the number of *tandem repeats*; or whole *gene duplications* in place; or duplications on a different part of the chromosome, leading to the existence of *pseudo-genes*; or inversions of segment of DNA. All of these kinds of mutations have been suggested to be perfect characters in the right biological contexts, and this list is certainly not exhaustive. Other suggestions for perfect or near-perfect characters include *ultra-conserved* elements and their flanking DNA [81]. For a general discussion of perfect characters, see [357].

> Because the probability that a SINE/LINE will be lost once it has been inserted into the genome is extremely small, and the probability that the same SINE/LINE will be inserted independently into an identical region in the genomes of two different taxa is also very small, the probability that homoplasy will obscure phylogenetic relationships is, for all practical purposes, zero. [314]

> ... microRNAs ... are highly conserved, non-coding genes that can be treated in datasets as presence/absence, like most phenotypic characters, ... with the additional advantage of being rarely lost in evolution. [201]

1.4.1.4 Perfection Is an Abstraction

> Have no fear of perfection—you'll never reach it.
> — Salvador Dali

We do not claim that every binary biological character is a perfect character, or even that the specific ones discussed here are always perfect (there is continuing debate about some of these [174, 351]). And, in fact, most characters (particularly in phylogenetics) are not perfect [364, 420]. But we claim that perfect characters are either plentiful (as in the case of SNPs) or can be found (as in micro-RNA, LINEs, SINEs, and complex morphological traits). So, there are important contexts where the infinite-sites, perfect-character models hold sufficiently well to justify their use, whether they occur in the context of population genetics, or in phylogenetics, or in other contexts such as in linguistics [306].

Further, when a site shows evidence of homoplasy, many studies will often remove that site from the data [112]. This is justified when the primary focus is on the history of the taxa, not the history of the characters, and when it is believed that the true history of the taxa, stripped of characters, is a tree with one leaf for each taxon. In that view, one should seek perfect characters (which are necessarily binary) that are sufficient to construct a history of the taxa in the form of a tree T. Characters that are not perfect (show evidence of homoplasy), are unneeded in order to define T, and can be ignored. In that view, those characters are noise imposed on top of the historically correct tree, definable with perfect characters. Similarly, in the context of SNP sites, any DNA site that has more than two variants (above some specified frequency) in a population can be removed.

> Recombination, parallel or back mutations, gene-conversion or genotypic misclassification can cause the perfect phylogeny condition to be violated. Such data may be pruned using an algorithm that deletes haplotypes or SNPs or a combination of both to give a reduced set of data consistent with a gene tree.[9] [75]

A further justification for the perfect character model is that often a clean, ideal model can be used in the *core* of a practical method that handles messier data not completely conforming to the model.[10] In that approach, the core is executed repeatedly by wrapper software that iteratively modifies the data. In the context of perfect characters, some examples of this approach are found in [105, 106, 158, 240, 363, 368, 369, 392].

---

9 "Gene tree" is a term for a perfect phylogeny in [75].

10 This is one of the foundational principles of theoretical, algorithmic computer science—one of the tribes I belong to.

Additional support for the perfect character model comes from the discussion (on the relationship of the perfect-phylogeny model to coalescent theory) in section 12.3.1.

1.4.1.5 And, We Can Often Incorporate Homoplasy

Above, we gave the argument that perfect characters (or nearly perfect characters) do exist in real applications. That is our biological justification for perfect characters. Still, from a *biological* perspective, the assumption of infinite-sites, and of perfect characters, is limiting. However, from a *mathematical, modeling* perspective, the limitations are much less severe. This is because a back or a recurrent mutation, or parallel evolution, *can be modeled* as two-crossover recombination, with no modification to the infinite-sites assumption. This will be detailed in section 3.2.3.3. So, in this book, *any* result that holds for two-crossover recombination (and of course for multiple-crossover recombination), holds for trees or genealogical networks *with homoplasy* allowed. It is only the results that depend on single-crossover recombination that do not automatically apply when homoplasy is allowed.

We will sometimes explicitly point out that a result holds with homoplasy, but we will usually leave it to the reader to realize this, since our primary expositional model is the ARG, which does not *explicitly* allow homoplasy.

1.5 The Observed Data

In this book, the data for most of the problems of interest is represented by a set of taxa and a set of binary characters, together with information on the state of each character for each taxon. This data is usually presented in the form of an n by m *matrix* M whose n rows represent taxa, and whose m columns represent characters. Each cell (f, c) of M specifies the state of character c for taxon f. For example, see figure 2.1 (page 36) showing a matrix M with five taxa and five characters. Except for discussions of empirical data, we will generally not care what biological phenomena produced M, as long as each character is a perfect character (or near-perfect character).

When talking about the matrix M we will often use the terms "taxon" and "row" interchangeably, and will often use the terms "character," "column," "locus," and "site" interchangeably, choosing whichever term is most informative for the context. Further, the ordered entries in a row f of M can be considered to form a *sequence*, and so we have the following:

Definition The *sequence s_f for taxon f*, or the *sequence for f*, is the ordered sequence formed from the entries in row f of matrix M.

Given this definition, we will also consider M to be a *set* of sequences, as well as a *matrix* representing that set of sequences. Context will often determine whether M is a set or a matrix.

1.6 A Few Graph Definitions

The principle combinatorial objects that we deal with in this book are *graphs*, and so we state a few basic definitions and facts about the kinds of graphs we will encounter.

Definition An *undirected graph* $G = (V, E)$ is a combinatorial object consisting of a set of *nodes* (also called *vertices*) V, and a set of *edges* E. Each edge in E is specified by an *unordered* pair of nodes (u, v) from V, where $u \neq v$.

Definition For an edge $e = (u, v)$ in E, nodes u and v are called the *endpoints* of edge e.

For example, the undirected graph in figure 1.7a (page 25) has node set $V = \{a, b, c, d, e\}$ and edge set $E = \{(a, b), (a, c), (a, e), (b, d), (c, d), (c, e)\}$.

The definition of an undirected graph given here does not allow an edge whose two endpoints are the same, and it does not allow multiple copies of the same edge. Alternate definitions of an undirected graph do allow such self-loops and parallel edges, but they will never appear in the graphs considered in this book.

Definition An undirected graph $G = (V, E)$ is called *bipartite* if the nodes of V can be *partitioned* into two subsets V_1, V_2, so that for every edge in E, one of the endpoints of the edge is in V_1, and the other endpoint is in V_2. See figure 1.8.

Definition A *directed graph* $G = (V, E)$ is defined by a set of nodes V, and a set of *directed edges* E, where each directed edge is specified by an *ordered* pair of nodes (u, v), where $u \neq v$. By convention, the directed edge $e = (u, v)$ is directed *from* the first node, u, in the ordered pair *to* the second node, v, in the ordered pair. The first node is called the *tail* of e and the second node is called the *head* of e, so e is directed from its tail to its head. See figure 1.7d.

Definition A *subgraph* $G' = (V', E')$ of a graph $G = (V, E)$ is a graph where $V' \subseteq V$ and $E' \subseteq E$, and for every edge $(u, v) \in E'$, both u and v are in V'. An *induced* subgraph $G' = (V', E')$ of G is a graph where $V' \subseteq V$, and E' consists of *every* edge $(u, v) \in E$ such that both u and v are in E'. See figures 1.7b and c.

Definition If G is a directed graph, or a graph with some directed and some undirected edges, the *underlying undirected graph* of G is the graph formed by ignoring the directions on the edges of G. That is, each ordered pair of nodes that defines an ordered edge in G is now considered as an unordered pair of nodes.

Definition For any node v in an undirected graph, the *degree* of v is the number of edges that touch v, that is, the number of edges where v is one of the endpoints. For a node v in a directed graph, the *in-degree* of v is the number of edges directed into v; the *out-degree* of v is the number of edges directed out of v. See figures 1.7a and d.

Definition An *undirected path* from a node v_1 to a node v_k in an undirected graph $G = (V, E)$ is specified by an ordered list of nodes $v_1, v_2, \ldots, v_k$, such that for every i from 1 to $k-1$, the node pair (v_i, v_{i+1}) is an edge in E.

Definition A *directed path* from a node v_1 to a node v_k in a directed graph $G = (V, E)$ is specified by an ordered list of nodes $v_1, v_2, \ldots, v_k$, such that for every i from 1 to $k-1$, the ordered node pair (v_i, v_{i+1}) is an edge in E, that is, an edge directed from v_i to v_{i+1}.

Definition An undirected graph G is *connected* if for ever pair of nodes u, v in G, there is a path between u and v in G. It is *biconnected* if for every pair of nodes u, v there are at least *two* paths between u and v that share no nodes other than u and v.

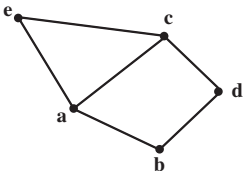
Definition A *cut edge* in a connected graph is an edge whose removal disconnects the graph.

Definition A directed graph G is connected (sometimes called "weakly connected") if the underlying undirected graph of G is connected. It is *strongly connected* if for every pair of nodes u, v, there is a directed path from u to v, and also a directed path from v to u. Those two paths may share nodes in addition to u and v.

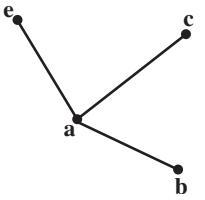
Definition An *undirected cycle* in an undirected graph G is an undirected path that starts and ends at the same node. A *directed cycle* in an directed graph G is an directed path which starts and ends at the same node.

Definition A graph G (possibly with some directed edges) is called a *network* if the underlying undirected graph of G contains an undirected cycle.

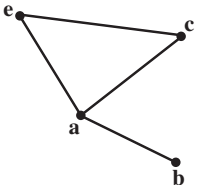
For example, when we consider the pedigree in figure 1.2 as an undirected graph, then it contains an undirected cycle. The directed graph in figure 1.6 does not contain a directed cycle, as defined before, but its underlying undirected graph does contain an undirected cycle. That cycle is what we called a *recombination cycle*.



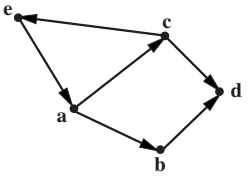
(a) An undirected graph G



(b) A subgraph G'



(c) An induced subgraph



(d) A directed graph

Figure 1.7 (a) The undirected graph G contains several cycles; one is $\{a,b,d,c\}$. The degree of node a is three. (b) A subgraph G' of G containing nodes $\{a, b, c, e\}$ and edges $\{(a, b), (a, c), (a, e)\}$. (c) The *induced* subgraph of G, induced by the node set $\{a, b, c, e\}$. (d) The directed graph contains the directed cycle $\{a,c,e\}$. It is connected but not strongly connected. Node a has in-degree one and out-degree two. Node a is the tail of edge (a, c), and the head edge (e, a).

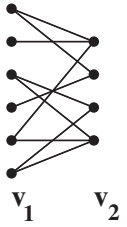
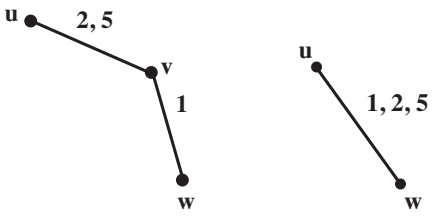


Figure 1.8 A bipartite graph. Every edge in the graph has one endpoint in V_1 and one in V_2.

Definition The *node-contraction* of a node v of degree two removes v and merges the two edges incident with v into a single edge. The new edge is labeled by the union of the characters that labeled the two merged edges. See figure 1.9.



(a) Before node-contraction

(b) After node contraction

Figure 1.9 (a) Two edges incident with node v of degree two.(b) The result of contracting node v.

Definition An *edge-contraction* of an edge $e = (u, v)$ in a graph (either directed or undirected) is the operation that superimposes nodes u and v, removing the edge e between them.

1.6.1 Trees and DAGs: The Most Central Graphs in This Book

Now we can define the most important types of graphs needed in this book, and some properties of those graphs. We have already used some of these terms, relying on the reader's general background and intuition, but now we present formal definitions.

1.6.1.1 Trees and Subtrees

Definition An undirected graph is called an *unrooted tree* or an *undirected tree* if it is connected and contains no undirected cycles. See figure 1.10a.

When it is clear by context that G is undirected, we will use "tree" in place of "unrooted tree" or "undirected tree."

Definition In an undirected tree T, a node of degree one is called a *leaf* of T, and every other node is called an *internal* or *interior* node.

The following is one of the most basic facts about trees.

Lemma 1.6.1 *Every undirected tree with n nodes has exactly $n - 1$ edges. Conversely, any connected, undirected graph with n nodes and $n - 1$ edges is a tree.*

Definition A directed graph G is called a *rooted tree* with a root node r, if the underlying undirected graph of G is a tree, and if every node in G is reachable from r via some directed path. See figure 1.10b. Equivalently, a *rooted tree* is a directed tree with one node designated as the root, where all edges are directed *away from* the root.

Definition In a rooted tree T, a node with out-degree zero is called a *leaf*; the root is the unique node with in-degree zero. Every other node is called an *internal* or *interior* node.

Definition A *subtree* T' of a tree T is a tree contained in T, where T' has fewer nodes and edges than T. Equivalently, T' is a connected subgraph of T. A subgraph of tree T that is not connected is called a *subforest* of T.

Often we will be concerned with the subtree of a special form.

Definition A subtree of a directed tree T, consisting of all the nodes and edges reachable from a particular node v in T, is called *the subtree of* T *rooted at node* v. This is usually denoted T_v. See figure 1.11.

Definition A rooted tree where every non-leaf node has out-degree two is called a *binary tree*.

The following is one of the most basic theorems concerning binary trees. It is easy to prove by induction on the number of leaves.

Theorem 1.6.1 *Any binary tree with* n *leaves has exactly* $n-1$ *non-leaf nodes, and* $2n-2$ *edges.*

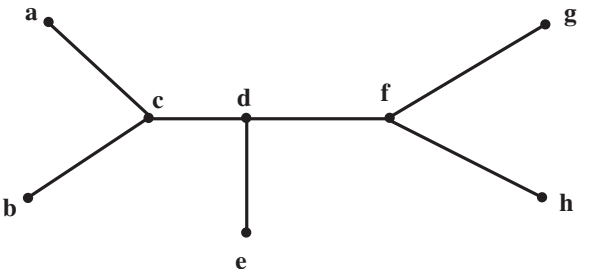
1.6.1.2 Directed Acyclic Graphs (DAGs)

Definition A directed graph that contains no *directed* cycles is called a *directed acyclic graph*, or DAG for short. See figure 1.12a.

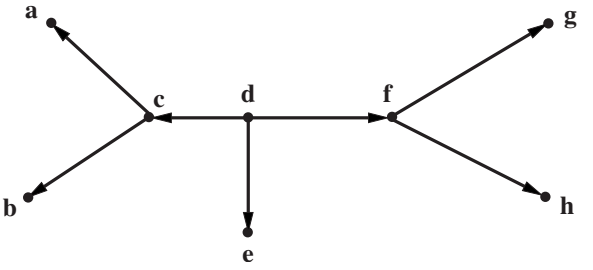
Note the underlying undirected graph of a DAG G may contain undirected cycles, even though G does not contain any directed cycles. Further, a DAG may contain a recombination cycle, since a recombination cycle is not a directed cycle.

Definition Two directed paths in a DAG that have the same start and end nodes, but otherwise share no nodes, form a *recombination cycle*. The start node of the paths is called the *coalescent* node of the cycle, and the end node of the paths is called the *recombination* node. In figure 1.5, the coalescent node is labeled Albert and the recombination node is labeled Prince William.

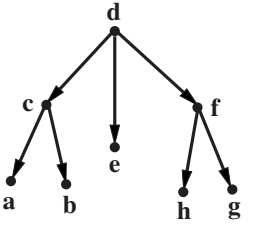
Now we establish some simple properties of DAGs.



(a) An undirected tree

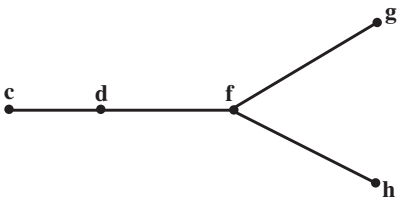


(b) A rooted tree T with root node d

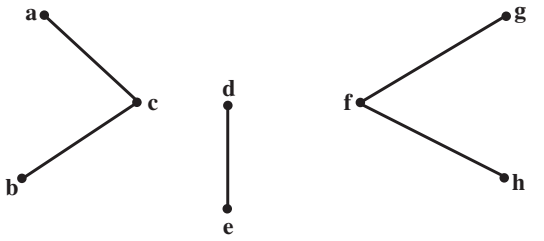


(c) Tree T drawn with
edges directed down

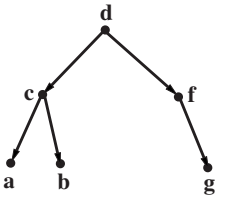
Figure 1.10 (a) An undirected tree. By definition, it has no root; its leaves are {a,b,e,g,h}.
(b) A rooted tree T, rooted at node d.
(c) Tree T drawn so that the root is at the top and all edges are directed downward. This is the standard way that directed trees are drawn in this book.



(a) A subtree



(b) A subforest



(c) A rooted subtree

(d) The subtree T_c, rooted at node c

Figure 1.11 (a) A subtree of the tree shown in figure 1.10a.
(b) A subforest of the tree in figure 1.10.
(c) A subtree of the rooted tree T in figure 1.10c.
(d) The subtree T_c, rooted at node c, of the rooted tree T in Figure 1.10c.

Lemma 1.6.2 *Every DAG G has at least one node with out-degree zero, and at least one node with in-degree zero.*

Proof: Arbitrarily pick a node u in G and find a directed path starting at u, by following any directed edge out of each successive node on the path. Since G does not contain any directed cycles, no node on this path is included twice, and since the number of nodes in G

is finite, ultimately the path must reach a node v that has out-degree zero. So the first part of the lemma is proved.

To prove that every DAG G has at least one node with in-degree zero, reverse the directions on every edge in G. This graph will also be a DAG. By what was proved above, this graph has a node u with out-degree zero, meaning that u is a node with in-degree zero in G. ∎

Note that a DAG might have more than one node of in-degree zero, or more than one node of out-degree zero.

Lemma 1.6.3 *The nodes of a DAG G can be partitioned into layers, and the layers can be ordered so that for any node v, all the nodes that can reach v via a directed path in G are in layers before the layer containing v.*

Proof: The first layer consists of all the nodes of in-degree zero. Remove all of those nodes from G. Then the second layer consists of all the nodes of in-degree zero in the resulting graph. The successive layers are obtained by repeating this process. ∎

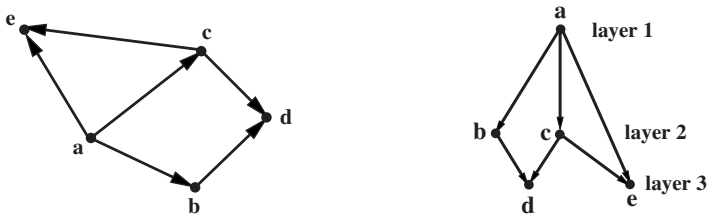
Definition A total order, Π, of the nodes of a DAG G that is consistent with the ordering of the layers defined in lemma 1.6.3 is called a *topological sort* of the nodes of G. See figure 1.12*b*.

In Π, a node u must appear before a node v, if v is reachable via a directed path in G, from node u. Lemma 1.6.3 implies that a topological sort is always possible for any DAG.

1.6.2 Genealogical Networks and ARGs: First Definitions

Genealogical networks and *ancestral recombination graphs* are the DAGs that most centrally model the biological phenomena of interest in this book. Indeed, most of the book concerns these models. Here we give somewhat *informal* definitions for these graphs. More complete, formal definitions will be given in chapter 3. Also, in chapter 14, we will introduce another DAG, called a *reticulation network*.

A *genealogical* network $\mathcal{N}$, generating a set of sequences M, each of length m, is a directed acyclic graph containing *one root* node with in-degree zero, and n leaves, each with in-degree one and out-degree zero. Every other node has in-degree one (tree nodes) or two (recombination nodes). Each site in M is assigned to a set of edges in $\mathcal{N}$, but none is assigned to any edge entering a recombination node. Each node v in $\mathcal{N}$ is labeled by an m-length binary sequence, denoted s_v. The label for any non-recombination node v is obtained from the label of v's parent $p(v)$, by changing the state of any sites labeling the



(a) A DAG G

(b) The same DAG G drawn with layers

Figure 1.12 (a) The directed graph G contains no *directed* cycles and so is a DAG. The underlying, undirected graph does contain a cycle.
(b) In this drawing of graph G, the layers defined in lemma 1.6.3 are drawn top down. Thus, the first layer consists of the single node a; the second layer consists of $\{b, c\}$; and the third consists of $\{d, e\}$. The DAG has four different topological sorts, for example (a, b, c, d, e) and (a, c, b, e, d). If we view node a as the root node, the DAG has two recombination cycles: one with coalescent node a and recombination node d; and one with coalescent node a and recombination node e.

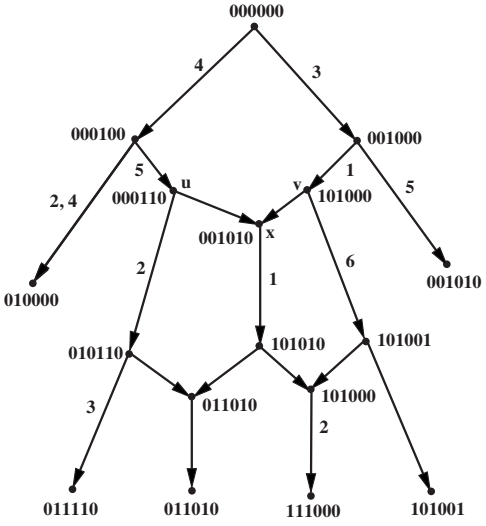


Figure 1.13 A genealogical network. In the genealogical network, sites $\{1, 2, 3, 4, 5\}$ each mutate more than once. The sequence 001010 that labels recombination node x, could have taken 00 from the parental sequence at node u, then 10 from the parental sequence at v, and then the last 10 again from the parental sequence at u. Note that the parental sequences are identical at sites 2 and 6, so those are the only sites where there is a choice for which parental sequence contributes the state. For all other sites, the choice is forced.

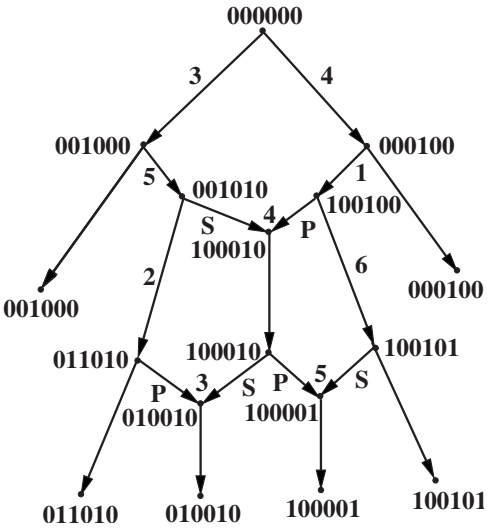


Figure 1.14 In any ARG, each site mutates exactly once, and in this ARG, each recombination event is a single-crossover recombination. The labels "P" and "S" on the edges into a recombination node indicate which parental sequence contributes the prefix and which contributes the suffix; the number over the recombination node indicates where the recombinant sequence switches from prefix to suffix contributions.

edge $(p(v), v)$. The label on a recombination node x can be any binary sequence s such that for each site c, the state of s at c equals the state at c in at least one of x's parental sequences. The sequences labeling the leaves of $\mathcal{N}$ must define the sequences in M. figure 1.13 shows a genealogical network. In the taxonomy of phylogenetic networks shown in [195] (page 69 of that book), these networks are also called *recombination networks*. A modified version of that taxonomy is shown in figure 3.5.

> An ARG is a directed graph that simultaneously describes vertical and nonvertical evolutionary events. [37]

We informally define an *ancestral recombination graph* (ARG) for M, as a genealogical network $\mathcal{N}$ for M, where each site labels *exactly one* edge in $\mathcal{N}$. Hence, the main distinction between a genealogical network and an ARG is that the mutations on an ARG obey the infinite-sites model.

Further, genealogical networks always allow multiple-crossover recombination, while ARGs can be required to only allow single-crossover recombination. When an ARG is

constrained in that way, the recombinant sequence s_x labeling a recombination node x, must be formed from a prefix of one of x's parental sequences, followed by a suffix of the other parental sequence. Figure 1.14 shows an ARG with the same DAG used for the genealogical network in figure 1.13, but deriving a different set of sequences, since each site mutates only once. Relating this definition to the quote above, mutations are *vertical* events, and recombinations are *nonvertical* events.

1.7 The Book

This book discusses algorithmic and mathematical results, mostly obtained in the last decade, concerning combinatorial structure of genealogical networks, particularly ARGs, (sometimes extending to other phylogenetic networks). The algorithms exploit the structure, and are used to deduce information (sometimes only partial) about the networks, or are used to explicitly construct networks that generate observed sequences through the biological events of mutation and recombination (and sometimes other events). The networks serve as hypotheses for the true genealogical history of the extant sequences, and help to address fundamental biological questions, or are used in practical problems such as association mapping, location of recombination hotspots, computing local recombination rates, phasing genotypic data, and identification of SNP sites (all of which will be discussed in depth). Moreover, algorithms that create explicit networks form a complement, or an alternative, to methods based on the more commonly used numerical, statistical, measures that less directly reflect the underlying genealogy.

Problems of constructing genealogical networks from sequence data, or deducing features of such networks, are significantly more complex than for the analogous problems in trees, and the field of network reconstruction is much less developed than the field of tree reconstruction. The book develops ideas and methods for reconstructing or obtaining information about genealogical networks, particularly ARGs. It concentrates on ideas and methods that we believe are *fundamental*, and on applications that illustrate the utility of these ideas and methods.

Bibliography

[1] The 1000 Genomes Project Consortium. A map of human genome variation from population-scale sequencing. *Nature*, 467:1061–1073, 2010.

[2] The 1000 Genomes Project Consortium. An integrated map of genetic variation from 1,092 genomes. *Nature*, 491:56, 2012.

[3] R. Agarwala and D. Fernandez-Baca. A polynomial-time algorithm for the perfect phylogeny problem when the number of character states is fixed. *SIAM Journal on Computing*, 23:1216–1224, 1994.

[4] R. Agarwala, D. Fernandez-Baca, and Giora Slutzki. Fast algorithms for inferring evolutionary trees. *Journal of Computational Biology*, 2:397–408, 1995.

[5] A. Aho, Y. Sagiv, T. Szymanski, and D. Ullman. Inferring a tree from lowest common ancestors with an application to the optimization of relational expressions. *SIAM Journal on Computing*, 10:405–421, 1981.

[6] R. K. Ahuja, T. L. Magnanti, and J. B. Orlin. *Network Flows: Theory, Algorithms, and Applications*. Prentice Hall, 1993.

[7] B. Albrecht, C. Scornavacca, A. Cenci, and D. Huson. Fast computation of minimum hybridization networks. *Bioinformatics*, 28:191–197, 2012.

[8] B. Allen and M. Steel. Subtree transfer operations and their induced metrics on evolutionary trees. *Annals of Combinatorics (online)*, 5, 2001.

[9] D. Altshuler, M. Daly, and E. Lander. Genetic mapping in human disease. *Science*, 322:881–888, 2008.

[10] M. Arenas, M. Patricio, D. Posada, and G. Valiente. Characterization of phylogenetic networks with NetTest. *BMC Bioinformatics*, 11:268, 2010.

[11] M. Arenas. Computer programs and methodologies for the simulation of DNA sequence data with recombination. *Frontiers in Genetics*, 4, 2013.

[12] M. Arenas. The importance and application of the ancestral recombination graph. *Frontiers in Genetics*, 4, 2013.

[13] M. Arenas, G. Valiente, and D. Posada. Characterization of reticulate networks based on the coalescent with recombination. *Molecular Biology and Evolution*, 25:2517–2520, 2008.

[14] F. Ariey and D. Menard et al. A molecular marker of artemisinin-resistant *Plasmodium falciparum* malaria. *Nature*, 505:50–55, 2014.

[15] N. Arnheim, P. Calabrese, and M. Nordborg. Hot and cold spots of recombination in the human genome: The reason we should find them and how this can be achieved. *American Journal of Human Genetics*, 73:5–16, 2003.

[16] N. Arnheim, P. Calabrese, and I. Tiemann-Boege. Mammalian meiotic recombination hotspots. *Annual Review of Genetics*, 41:363–399, 2007.

[17] V. Bafna and V. Bansal. The number of recombination events in a sample history: conflict graph and lower bounds. *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, 1:78–90, 2004.

[18] V. Bafna and V. Bansal. Improved recombination lower bounds for haplotype data. In *RECOMB, The Annual International Conference on Research in Computational Molecular Biology*. LNBI 3500, Springer, 2005.

[19] V. Bafna and V. Bansal. Inference about recombination from haplotype data: Lower bounds and recombination hotspots. *Journal of Computational Biology*, 13:501–521, 2006.

[20] V. Bafna, D. Gusfield, S. Hannenhalli, and S. Yooseph. A note on efficient computation of haplotypes via perfect phylogeny. *Journal of Computational Biology*, 11(5):858–866, 2004.

[21] V. Bafna, D. Gusfield, G. Lancia, and S. Yooseph. Haplotyping as perfect phylogeny: A direct approach. *J. Computational Biology*, 10:323–340, 2003.

[22] H. J. Bandelt, P. Foster, and A. Röhl. Median-joining networks for inferring intraspecific phylogenies. *Molecular Biology and Evolution*, 16:37–48, 1999.

[23] H. J. Bandelt, P. Foster, B. Sykes, and M. Richards. Mitochondrial portaits of human populations using median networks. *Genetics*, 141:743–753, 1995.

[24] H. J. Bandelt and A. W. Dress. Split decomposition: A new and useful approach to phylogenetic analysis of distance data. *Molecular Phylogenetics and Evolution*, 1(3):242–252, 1992.

[25] E. Bapteste, Y. Boucher, and W.F. Doolittle. The fate of phylogenetics in the face of lateral gene transfer. In J.H. Schulz, editor, *Genetic Recombination Research Progress*, pages 139–162. Nova Science Press, 2008.

[26] M. Baroni, S. Grunewald, V. Moulton, and C. Semple. Bounding the number of hybridisation events for a consistent evolutionary history. *Journal of Mathematical Biology*, 51:171–182, 2005.

[27] M. Baroni, C. Semple, and M. Steel. A framework for representing reticulate evolution. *Annals of Combinatorics*, 8:391–408, 2004.

[28] T. Barzuza, J.S. Beckman, R. Shamir, and I. Pe'er. Computational problems in perfect phylogeny haplotyping: XOR genotypes and TAG SNPs. In *CPM, Symposium on Combinatorial Pattern Matching*, 2004.

[29] T. Barzuza, J.S. Beckmann, and R. Shamir. Computational problems in perfect phylogeny haplotyping: Typing without calling the allele. *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, 5(1):101–109, 2008.

[30] T. Barzuza, J.S. Beckmann, R. Shamir, and I. Pe'er. Typing without calling the allele: A strategy for inferring SNP haplotypes. *European Journal of Human Genetics*, 13:898–901, 2005.

[31] T. Bersaglier and P. C. Sabeti et al. Genetic signatures of strong recent positive selection at the lactase gene. *American Journal of Human Genetics*, 74:1111–1120, 2004.

[32] S. Besenbacher, T. Mailund, and M. Schierup. Local phylogeny mapping of quantitative traits: Higher accuracy and better ranking than single-marker association in genomewide scans. *Genetics*, 181:747–753, 2009.

[33] A. Bigham and M.D. Schriver et al. Identifying signatures of natural selection in Tibetan and Andean populations using dense genome scan data. *PLoS Genetics*, 6(9):e1001116, 09 2010.

[34] R. E. Bixby and D. K. Wagner. An almost linear-time algorithm for graph realization. *Mathematics of Operations Research*, 13:99–123, 1988.

[35] G. Blin and R. Rizzi et al. Minimum mosaic inference of a set of recombinants. *International Journal of Foundations of Computer Science*, 24:51–66, 2011.

[36] J. S. Bloom, I.M. Ehrenreich, W. T. Loo, T.L.V Lite, and L. Kruglyak. Finding the sources of missing heritability in a yeast cross. *Nature*, 494:234–237, 2013.

[37] E. Bloomquist and M.A. Suchard. Unifying vertical and nonvertical evolution: a stochastic ARG-based framework. *Systematic Biology*, 59:2741, 2010.

[38] H. Bodlaender, M. Fellows, and T. Warnow. Two strikes against perfect phylogeny. *Proceedings of the 19th International Colloquium on Automata, Languages and Programming*, pages 273–283, 1992.

[39] A. Bondy and U. S. R. Murty. *Graph Theory*. Springer, Graduate Texts in Mathematics, 2008.

[40] P. Bonizzoni. A linear-time algorithm for the perfect phylogeny haplotype problem. *Algorithmica*, 48:267–285, 2007.

[41] P. Bonizzoni, A.P. Carrieri, G. Della Vedova, R. Dondi, and T.M. Przytycka. When and how the perfect phylogeny model explains evolution. In N. Jonoska and M. Saito, editors, *Discrete and Topological Models in Molecular Biology*, Natural Computing Series, chapter 4. Springer, 2013.

[42] P. Bonizzoni, G. Della Vedova, R. Dondi, and J. Li. The haplotyping problem: Models and solutions. *Journal of Computer Science and Technology*, 18:675–688, 2003.

[43] M. Bordewich, C. McCartin, and C. Semple. A 3-approximation algorithm for the subtree distance between phylogenies. *Journal of Discrete Algorithms*, 6(3):458–471, 2008.

[44] M. Bordewich and C. Semple. On the computational complexity of the rooted subtree prune and regraft distance. *Annals of Combinatorics*, 8:409–423, 2004.

[45] M. Bordewich and C. Semple. Computing the minimum number of hybridization events for a consistent evolutionary history. *Discrete Applied Math*, 155:914–928, 2007.

[46] D. Botstein, R. L. White, M. Skolnick, and R. W. Davis. Construction of a genetic linkage map in man using restriction fragment length polymorphisms. *American Journal of Human Genetics*, 32:314–331, 1980.

[47] U. Brandes and S. Cornelsen. Phylogenetic graph models beyond trees. *Discrete Applied Math*, 157:2361–2369, 2009.

[48] D. Brown and I. Harrower. A new integer programming formulation for the pure parsimony problem in haplotype analysis. In *WABI, Workshop on Algorithms in Bioinformatics*, volume 3240, pages 254–265. LNCS, Springer, 2004.

[49] D. Brown and I. Harrower. Toward an algebraic understanding of haplotype inference by pure parsimony. In *Computational Systems Bioinformatics Conference*. LNCS, Springer-Verlag, 2006.

[50] D. Brown and I.M. Harrower. Integer Programming Approaches to Haplotype Inference by Pure Parsimony. *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, 3(2):141–154, 2006.

[51] S.R. Browning and B.L. Browning. Rapid and accurate haplotype phasing and missing-data inference for whole-genome association studies by use of localized haplotype clustering. *American Journal of Human Genetics*, 81:1084–1097, 2007.

[52] S.R. Browning and B.L. Browning. Haplotype phasing: Existing methods and new developments. *Nature Reviews Genetics*, 12:703–714, 2011.

[53] D. Bryant and V. Moulton. Neighbornet: An agglomerative method for the construction of planar phylogenetic networks. In R. Guigo and D. Gusfield, editors, *WABI, Workshop on Algorithms in Bioinformatics*, volume 2452 of *Lecture Notes in Computer Science*, pages 375–391. Springer, 2002.

[54] P. Buneman. The recovery of trees from measures of dissimilarity. In D.G. Kendall and P. Tautu, editors, *Mathematics in the archaeological and historical sciences*, pages 387–385. Edinburgh University Press, 1971.

[55] P. Buneman. A characterization of rigid circuit graphs. *Discrete Math*, 9:205–212, 1974.

[56] C. Campbell, Z. Wang, and Y. Qian. Ancestral recombination histories for error detection in genome sequencing. Technical report, University of Oxford, Statistics Department, 2010.

[57] G. Cardona, M. Llabres, F. Rossello, and G. Valiente. A distance metric for a class of tree-sibling phylogenetic networks. *Bioinformatics*, 24(13):1481–1488, 2008.

[58] G. Cardona, M. Llabres, F. Rossello, and G. Valiente. Metrics for phylogenetic networks i: Generalizations of the Robinson-Foulds metric. *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, 6(1):46–61, 2009.

[59] G. Cardona, M. Llabres, F. Rossello, and G. Valiente. Comparison of galled trees. *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, 8(2):410–427, 2011.

[60] G. Cardona, F. Rossello, and G. Valiente. Comparison of tree-child phylogenetic networks. *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, 6:552–569, 2009.

[61] N. Casali and F. Drobniewski et al. Evolution and transmission of drug-resistant tuberculosis in a russian population. *Nature Genetics*, 46:279–286, 2014.

[62] A. Chakravarti. It's raining SNP's, hallelujah? *Nature Genetics*, 19:216–217, 1998.

[63] P. Charbit, M. Habib, W. Limouzy, F. de Montgolfier, M. Raffinot, and M. Rao. A note on computing set overlap classes. *Information Processing Letters*, 108:186–191, 2008.

[64] B. Charlesworth. Effective population size and patterns of molecular evolution and variation. *Nature Reviews: Genetics*, 10:195–205, 2009.

[65] N. Charlton, I. Carbone, S. Tavantzis, and M. Cubeta. Phylogenetic relatedness of the M2 double-stranded RNA in *Rhizoctonia* fungi. *Mycologia*, 100:555–64, 2008.

[66] Z. Z. Chen and L. Wang. Algorithms for reticulate networks of multiple phylogenetic trees. *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, 9(2):372–384, 2012.

[67] Z. Z. Chen and L. Wang. An ultrafast algorithm for reticulate networks. *Journal of Computational Biology*, 20:38–41, 2013.

[68] C. Chewapreecha and S. Bentley et al. Dense genomic sampling identifies highways of pneumococcal recombination. *Nature Genetics*, 46:305–309, 2014.

[69] R.H. Chung and D. Gusfield. Empirical exploration of perfect phylogeny haplotyping and haplotypers. In *Proceedings of the 9th International Conference on Computing and Combinatorics COCOON03*, volume 2697 of *LNCS*, pages 5–19, 2003.

[70] A. Clark. Association testing with phased haplotypes vs. unphased diplotypes. Talk given at USC RECOMB workshop on Computational Methods for SNPs and Haplotypes, January 27, 2007.

[71] A. Clark. Inference of haplotypes from PCR-amplified samples of diploid populations. *Molecular Biology and Evolution*, 7:111–122, 1990.

[72] A. Clark. Finding genes underlying risk of complex disease by linkage disequilibrium mapping. *Current Opinion in Genetics & Development*, 13:296–302, 2003.

[73] A. Clark, X. Wang, and T. Matise. Contrasting methods of quantifying fine structure of human recombination. *Annual Review of Genomics and Human Genetics*, 11:4564, 2010.

[74] A. Clark, K. Weiss, and D. Nickerson et al. Haplotype structure and population genetic inferences from nucleotide-sequence variation in human lipoprotein lipase. *American Journal of Human Genetics*, 63:595–612, 1998.

[75] T. G. Clark, M. De Iorio, and R. C. Griffiths. Bayesian logistic regression using a perfect phylogeny. *Biostatistics*, 8:32–52, 2007.

[76] F. Cohan. What are bacterial species? *Annual Review of Microbiology*, 56:45787, 2002.

[77] F. Cole, S. Keeney, and M. Jasin. Preaching about the converted: How meiotic gene conversion influences genomic diversity. *Annals of the New York Academy of Sciences*, 1267(1):95–102, 2012.

[78] J. Comeron, R. Ratnappan, and S. Bailin. The many landscapes of recombination in *drosophila melanogaster*. *PLoS Genetics*, 8:e1002905, 10 2012.

[79] G. Coop and M. Przeworski. An evolutionary view of human recombination. *Nature Reviews Genetics*, 8:23–34, 2007.

[80] T. Cormen, C. Leiserson, R. Rivest, and C. Stein. *Introduction to Algorithms, 3rd edition*. MIT Press, 2009.

[81] N. C. Crawford and T. C. Glenn et al. More than 1000 ultraconserved elements provide evidence that turtles are the sister group of archosaurs. *Biology Letters*, 8:783 – 786, 2012.

[82] T. Dagan and W. Martin. Getting a better picture of microbial evolution en route to a network of genomes. *Philosophical Transactions of the Royal Society B*, 364(1527):2187–2196, 2009.

[83] E. Dahlhaus. Parallel algorithms for hierarchial clustering and applications to split decomposition and parity graph recognition. *Journal of Algorithms*, 36:205–240, 2000.

[84] M. Daly, J. Rioux, S. Schaffner, T. Hudson, and E. Lander. Fine-structure haplotype map of 5q31: Implications for gene-based studies and genomic LD mapping. Abstract of talk presented at the American Associate of Human Genetics National meeting, October 14, 2001.

[85] M. Daly, J. Rioux, S. Schaffner, T. Hudson, and E. Lander. High-resolution haplotype structure in the human genome. *Nature Genetics*, 29:229–232, 2001.

[86] S. Dasgupta, C. H. Papadimitriou, and U. V. Vazirani. *Algorithms*. McGraw Hill, 2008.

[87] W. H. Day and D. Sankoff. Computational complexity of inferring phylogenies by compatibility. *Systematic Zoology*, 35:224–229, 1986.

[88] A. de Vries and G. te Meerman. A haplotype sharing method for determining the relative age of SNP alleles. *Human Heredity*, 69:52–59, 2010.

[89] Z. Ding, V. Filkov, and D. Gusfield. A linear-time algorithm for the perfect phylogeny haplotyping problem. In *RECOMB, The Annual International Conference on Research in Computational Molecular Biology*, pages 585–600. LNCI 3500, Springer, 2005.

[90] Z. Ding, V. Filkov, and D. Gusfield. A linear-time algorithm for the perfect phylogeny haplotyping problem. *Journal of Computational Biology*, 13(2):522–553, 2006.

[91] Z. Ding, T. Mailund, and Y. S. Song. Efficient whole-genome association mapping using local phylogenies for unphased genotype data. *Bioinformatics*, 24(19):2215–2221, 2008.

[92] L. Dollo. Le lois de l'évolution. *Bulletin de la Societé Belge de Géologie de Paléontologie et d'Hydrologie*, 7:164–167, 1893.

[93] W. F. Doolittle. Phylogenetic classification and the universal tree. *Science*, 284:2124–2129, 1999.

[94] W. F. Doolittle. Uprooting the tree of life. *Scientific American*, 282:90–95, 2000.

[95] J. Dopazo, A. W. M. Dress, and A. von Haeseler. Split decomposition: a new technique to analyse viral evolution. *Proceedings of the National Academy of Sciences (USA)*, 90:10320–10324, 1993.

[96] A. Dress, K. T. Huber, J. Koolen, V. Moulton, and A. Spillner. *Basic Phylogenetic Combinatorics*. Cambridge University Press, 2012.

[97] A. Dress and M. Steel. Convex tree realizations of partitions. *Applied Math Letters*, 5:3–6, 1993.

[98] A. W. M. Dress and D. Huson. Constructing splits graphs. *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, 1(3):109–115, 2004.

[99] C. Drysdale and S. Ligget et al. Complex promoter and coding region β2-adrenergic receptor haplotypes alter receptor expression and predict *in vivo* responsiveness. *Proceedings of the National Academy of Sciences (USA)*, 97:10483–10488, 2000.

[100] A. Efros and E. Halperin. Haplotype reconstruction using perfect phylogeny and sequence data. *BMC Bioinformatics*, 13, 2012.

[101] I. M. Ehrenreich and M. D. Purugganan et al. Candidate gene association mapping of arabidopsis flowering time. *Genetics*, 183:325–335, 2009.

[102] N. El-Mabrouk. Deriving haplotypes through recombination and gene conversion. *Journal of Bioinformatics and Computational Biology*, 2:241–256, 2004.

[103] N. El-Mabrouk and D. Labuda. Haplotypes histories as pathways of recombinations. *Bioinformatics*, 20:1836–1841, 2004.

[104] J.A. Endler. *Natural Selection in the Wild*. Princeton University Press, 1986.

[105] E. Eskin, E. Halperin, and R. M. Karp. Large scale reconstruction of haplotypes from genotype data. In *RECOMB, The Annual International Conference on Research in Computational Molecular Biology*, pages 104–113, 2003.

[106] E. Eskin, E. Halperin, and R.M. Karp. Efficient reconstruction of haplotype structure via perfect phylogeny. *Journal of Bioinformatics and Computational Biology*, 1:1–20, 2003.

[107] G. Estabrook, C. Johnson, and F. McMorris. An idealized concept of the true cladistic character. *Mathematical Bioscience*, 23:263–272, 1975.

[108] G. Estabrook, C. Johnson, and F. McMorris. An algebraic analysis of cladistic characters. *Discrete Math*, 16:141–147, 1976.

[109] G. Estabrook, C. Johnson, and F. McMorris. A mathematical foundation for the analysis of cladistic character compatibility. *Mathematical Bioscience*, 29:181–187, 1976.

[110] P. Fearnhead and P. Donnelly. Estimating recombination rates from population genetic data. *Genetics*, 159:1299–1318, 2001.

[111] P. Fearnhead, R.M. Harding, J.A. Schneider, S. Myers, and P. Donnelly. Application of coalescent methods to reveal fine scale rate variation and recombination hotspots. *Genetics*, 167:2067–2081, 2004.

[112] J. Felsenstein. *Inferring Phylogenies*. Sinauer, 2004.

[113] J. Felsenstein. Trees of genes in populations. In O. Gascuel and M. Steel, editors, *Reconstructing Evolution: New Mathematical and Computational Advances*, pages 3–25. Oxford University Press, 2007.

[114] D. Fernandez-Baca. The perfect phylogeny problem. In D.Z. Du and X. Cheng, editors, *Steiner Trees in Industries*. Kluwer Academic Publishers, 2000.

[115] W. Fitch. Towards finding the tree of maximum parsimony. In G. F. Estabrook, editor, *Proceedings of the Eighth International Conference on Numerical Taxonomy*, pages 189–230. W. H. Freeman, 1975.

[116] K. Frazer, E. Eskin, and D. R. Cox et al. A sequence-based variation map of 8.27 million SNPs in inbred mouse strains. *Nature*, 448:831–843, 2007.

[117] R. H. French-Constant. The molecular genetics of insecticide resistance. *Genetics*, 194:807–815, 2013.

[118] L. Friss, R. Hudson, A. Bartoszewicz, J. Wall, T. Donfalk, and A. Di Rienzo. Gene conversion and differential population histories may explain the contrast between polymorphism and linkage disequilibrium levels. *American Journal of Human Genetics*, 69:831–843, 2001.

[119] W. Fu and J.M. Akey et al. Analysis of 6,515 exomes reveals the recent origin of most human protein-coding variants. *Nature*, 493:216–220, 2013.

[120] P. Gambette. http://www.lirmm.fr/~gambette/RePhylogeneticNetworks.php.

[121] P. Gambette. Who is who in phylogenetic networks: Articles, authors and programs. http://www.atgc-montpellier.fr/phylnet.

[122] P. Gambette and K. Huber. On encodings of phylogenetic networks of bounded level. *Mathematical Biology*, 65:157–180, 2012.

[123] M. Garey and D. Johnson. *Computers and intractability*. W.H. Freeman, 1979.

[124] F. Gavril and R. Tamari. An algorithm for constructing edge-trees from hypergraphs. *Networks*, 13:377–388, 1983.

[125] G. Gibson. Rare and common variants: Twenty arguments. *Nature Review Genetics*, 13:135–145, 2012.

[126] P. Gogarten and J. P. Townsand. Horizontal gene transfer, genome innovation and evolution. *Nature Reviews, Microbiology*, 9:679–687, 2005.

[127] A. Graça, I. Lynce, J. Marques-Silva, and A. L. Oliveira. Haplotype inference by pure parsimony: A survey. *Journal of Computational Biology*, 17:969–992, 2010.

[128] J. Gramm, T. Nierhoff, R. Sharan, and T. Tantau. On the complexity of haplotyping via perfect phylogeny. In *Second RECOMB Satellite Workshop on Computational Methods for SNPs and Haplotypes*, pages 20–21. LNBI, Springer, 2004.

[129] J. Gramm, T. Nierhoff, and T. Tantau. Perfect path phylogeny haplotyping with missing data is fixed-parameter tractable. In *First International Workshop on Parametrized and Exact Computation (IWPEC 2004)*. LNCS, Springer, 2004.

[130] R. C. Griffiths and P. Marjoram. Ancestral inference from samples of DNA sequences with recombination. *Journal of Computational Biology*, 3:479–502, 1996.

[131] R. C. Griffiths and P. Marjoram. An ancestral recombination graph. In P. Donnelly and S. Tavare, editors, *Progress in Population Genetics and Human Evolution*, pages 257–270. IMA Volumes in Mathematics and Its Applications, vol. 87, 1997.

[132] S. R. Grossman and P. C. Sabeti et al. Identifying recent adaptations in large-scale genomic data. *Cell*, 152(4):703–713, 2013.

[133] A. Gupta, J. Manuch, L. Stacho, and X. Zhao. Haplotype inferring via galled-tree networks is NP-complete. In *Computing and Combinatorics*, volume 5092, pages 287–298. LNCS, Springer, 2008.

[134] A. Gupta, J. Manuch, L. Stacho, and X. Zhao. Algorithm for haplotype inference via galled-tree networks with simple galls. *Journal of Computational Biology*, 19:439–454, 2012.

[135] A. Gupta, J. Manuch, X. Zhao, and L. Stacho. Characterization of the existence of galled-tree networks. *Journal of Bioinformatics and Computational Biology*, 4:1309–1328, 2006.

[136] D. Gusfield. Efficient algorithms for inferring evolutionary history. *Networks*, 21:19–28, 1991.

[137] D. Gusfield. *Algorithms on Strings, Trees and Sequences: Computer Science and Computational Biology*. Cambridge University Press, 1997.

[138] D. Gusfield. A practical algorithm for deducing haplotypes in diploid populations. In *Proceedings of 8th International Conference on Intelligent Systems in Molecular Biology*, pages 183–189. AAAI Press, 2000.

[139] D. Gusfield. Haplotyping as perfect phylogeny: Conceptual framework and efficient solutions. In *RECOMB, The Annual International Conference on Research in Computational Molecular Biology*, pages 166–175. ACM Press, 2002.

[140] D. Gusfield. Haplotype inference by pure parsimony. In R. Baeza-Yates, E. Chavez, and M. Chrochemore, editors, *Proceedings of the Annual Symposium on Combinatorial Pattern Matching*, volume 2676, pages 144–155. LNCS, Springer, 2003.

[141] D. Gusfield. On the decomposition optimality conjecture for phylogenetic networks. Technical report, UC Davis, Department of Computer Science, 2005.

[142] D. Gusfield. Optimal, efficient reconstruction of root-unknown phylogenetic networks with constrained and structured recombination. *Journal of Computer and System Sciences*, 70:381–398, 2005.

[143] D. Gusfield. The multi-state perfect phylogeny problem with missing and removable data: Solutions via integer-programming and chordal graph theory. *Journal of Computational Biology*, 17:383–399, 2010.

[144] D. Gusfield and V. Bansal. A fundamental decomposition theory for phylogenetic networks and incompatible characters. In *RECOMB, The Annual International Conference on Research in Computational Molecular Biology*, pages 217–232. LNBI 3500, Springer, 2005.

[145] D. Gusfield, V. Bansal, V. Bafna, and Y. S. Song. A decomposition theory for phylogenetic networks and incompatible characters. *Journal of Computational Biology*, 14:1247–1272, 2007.

[146] D. Gusfield, S. Eddhu, and C. Langley. The fine structure of galls in phylogenetic networks. *INFORMS Journal on Computing, Special Issue on Computational Biology*, 16:459–469, 2004.

[147] D. Gusfield, S. Eddhu, and C. Langley. Optimal, efficient reconstruction of phylogenetic networks with constrained recombination. *Journal of Bioinformatics and Computational Biology*, 2(1):173–213, 2004.

[148] D. Gusfield, Y. Frid, and D. Brown. Integer programming formulations and computations solving phylogenetic and population genetic problems with missing or genotypic data. In *Proceedings of 13th Annual International Conference on Combinatorics and Computing*, pages 51–64. LNCS 4598, Springer, 2007.

[149] D. Gusfield and D. Hickerson. A new lower bound on the number of needed recombination nodes in both unrooted and rooted phylogenetic networks. Report UCD-ECS-2004-06. Technical report, University of California, Davis, 2004.

[150] D. Gusfield, D. Hickerson, and S. Eddhu. An efficiently-computed lower bound on the number of recombinations in phylogenetic networks: Theory and empirical study. *Discrete Applied Math, Special Issue on Computational Biology, 2007*, 155:806–830, 2007.

[151] D. Gusfield and S. Orzack. Haplotype inference. In S. Aluru, editor, *Handbook of Computational Molecular Biology*, pages 18/1–18/25. Chapman and Hall/CRC, 2005.

[152] D. Gusfield and Y. Wu. The three-state perfect phylogeny problem reduces to 2-SAT. *Communication and Information Sciences*, 9:195–201, 2009.

[153] R. Gysel and D. Gusfield. Extensions and improvements to the chordal graph approach to the multistate perfect phylogeny problem. *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, 8(4):912–917, 2011.

[154] R. Gysel, F. Lam, and D. Gusfield. Constructing perfect phylogenies and proper triangulations for three-state characters. *Algorithms in Molecular Biology*, 7:26, 2012.

[155] M. Habib and T-H To. On a conjecture about compatibility of multi-states characters. In T. Przytycka and M.F. Sagot, editors, *Algorithms in Bioinformatics*, volume 6833 of *Lecture Notes in Computer Science*, pages 116–127. Springer, 2011.

[156] B. Halldorsson, V. Bafna, N. Edwards, R. Lipert, S. Yooseph, and S. Istrail. Combinatorial problems arising in SNP and haplotype analysis. In C. Calude, M. Dinneen, and V. Vajnovski, editors, *Discrete Mathematics and Theoretical Computer Science. Proceedings of DMTCS 2003*, volume 2731 of *Lecture Notes in Computer Science*, pages 26–47. Springer, 2003.

[157] B. Halldorsson, V. Bafna, N. Edwards, R. Lipert, S. Yooseph, and S. Istrail. A survey of computational methods for determining haplotypes. In *Proceedings of the First RECOMB Satellite on Computational Methods for SNPs and Haplotype Inference*, volume 2983 of *Lecture Notes in Bioinformatics, LNBI*, pages 26–47. Springer, 2004.

[158] E. Halperin and E. Eskin. Haplotype reconstruction from genotype data using imperfect phylogeny. *Bioinformatics*, 20:1842–1849, 2004.

[159] E. Halperin and R. M. Karp. Perfect phylogeny and haplotype assignment. In *RECOMB, The Annual International Conference on Research in Computational Molecular Biology*, pages 10–19. ACM Press, 2004.

[160] W. Hao, V.G. Allen, F.B. Jamieson, D.E. Low, and D.C. Alexander. Phylogenetic incongruence in E. coli O104: Understanding the evolutionary relationships of emerging pathogens in the face of homologous recombination. *PLoS ONE*, 7:e33971, 2012.

[161] D. Hartl. *A Primer of Population Genetics, 2nd Edition*. Sinauer, 1988.

[162] J. He and A. Zelikovsky. Linear reduction for haplotype inference. In *WABI, Workshop on Algorithms in Bioinformatics*, volume 3240, pages 242–253. LNCS, Springer, 2004.

[163] Y. He, C. Li, C. Amos, M. Xiong, and H. Ling L. Jin. Accelerating haplotype-based genome-wide association study using perfect phylogeny and phase-known reference data. *PLoS One*, 6:e22097, 2011.

[164] A.M. Heimberg and R. Cowper-Sal lari et al. MicroRNAs reveal the interrelationships of hagfish, lampreys and gnathostomes and the nature of the ancestral vertebrate. *Proceedings of the National Academy of Sciences (USA)*, 107:19379–19383, 2010.

[165] J. Hein. Reconstructing evolution of sequences subject to recombination using parsimony. *Mathematical Bioscience*, 98:185–200, 1990.

[166] J. Hein. A heuristic method to reconstruct the history of sequences subject to recombination. *Journal of Molecular Evolution*, 36:396–405, 1993.

[167] J. Hein, T. Jiang, L. Wang, and K. Zhang. On the complexity of comparing evolutionary trees. *Discrete Applied Math*, 71:153–169, 1996.

[168] J. Hein, M. Schierup, and C. Wiuf. *Gene Genealogies, Variation and Evolution: A Primer in Coalescent Theory*. Oxford University Press, 2005.

[169] G. Hellenthal, G. Busby, G. Band, J. Wilson, C. Capelli, D. Falush, and S. Myers. A genetic atlas of human admixture history. *Science*, 343:747–751, 2014.

[170] L. Helmuth. Genome research: Map of the human genome 3.0. *Science*, 293(5530):583–585, 2001.

[171] T. Hill and H. Schiöth et al. SPRIT: Identifying horizontal gene transfer in rooted phylogenetic trees. *BMC Evolutionary Biology (online)*, 10(42), 2010.

[172] D. Hillis and C. Moritz (eds). *Molecular Systematics*. Sinauer Associates, 1990.

[173] D. Hillis, C. Moritz, and B. Mable (eds). *Molecular Systematics, 2nd edition*. Sinauer Associates, 1996.

[174] D. M. Hillis. SINEs of the perfect character. *Proceedings of the National Academy of Sciences (USA)*, 96:9979–9981, 1999.

[175] A. Hinch, D. Reich, and S. Myers et al. The landscape of recombination in African Americans. *Nature*, 476:170–175, 2011.

[176] D. Hinds, L. Stuve, G. Nilsen, E. Halperin, E. Eskin, D. Gallinger, K. Frazer, and D. Cox. Whole-genome patterns of common DNA variation in three human populations. *Science*, 307:1072–1079, 2005.

[177] A. Van't Holt, N. Edmonds, M. Kalikova, F. Marec, and I. Saccheri. Industrial melanism in British peppered moths has a singular and recent mutational origin. *Science*, 332:958–960, 2011.

[178] E. Hubbel. Personal communication, 2000.

[179] K. T. Huber, L. van Iersel, S. Kelk, and R. Suchecki. A practical algorithm for reconstructing level-1 phylogenetic networks. *TCBB*, 8(3):607–620, 2011.

[180] R. Hudson. Properties of a neutral allele model with intragenic recombination. *Theoretical Population Biology*, 23:183201, 1983.

[181] R. Hudson. Gene genealogies and the coalescent process. *Oxford Survey of Evolutionary Biology*, 7:1–44, 1990.

[182] R. Hudson. Generating samples under the Wright-Fisher neutral model of genetic variation. *Bioinformatics*, 18(2):337–338, 2002.

[183] R. Hudson and N. Kaplan. Statistical properties of the number of recombination events in the history of a sample of DNA sequences. *Genetics*, 111:147–164, 1985.

[184] R. Hudson, N. Kaplan, and C. H. Langley. The hitchhiking effect revisited. *Genetics*, 123:887–899, 1989.

[185] R. Hudson, A. G. Saez, and F. J. Ayala. DNA variation at the Sod locus of *Drosophila melanogaster*: An unfolding story of natural selection. *Proceedings of the National Academy of Sciences (USA)*, 94:7725–7729, 1997.

[186] P. J. Humphries, S. Linz, and C. Semple. On the complexity of computing the temporal hybridization number for two phylogenies. *Discrete Applied Mathematics*, 161:871–880, 2013.

[187] D. J. Hunter and P. Kraft. Drinking from the fire hose — Statistical issues in genomewide association studies. *New England Journal of Medicine*, 357:436–439, 2007.

[188] M. Hurles. How homologous recombination generates a mutable genome. *Human Genomics*, 2:1016–1017, 2005.

[189] D. Huson. Split networks and reticulate networks. In O. Gascuel and M. Steel, editors, *Reconstructing Evolution: New Mathematical and Computational Advances*, pages 247–276. Oxford University Press, 2007.

[190] D. Huson and D. Bryant. Application of phylogenetic networks in evolutionary studies. *Molelular Biology and Evolution*, 23:254–267, 2006.

[191] D. Huson and H. Kloepper. Computing recombination networks from binary sequences. *Bioinformatics, supplement 2*, 21:ii159–ii165, 2005.

[192] D. Huson and T. Klopper. Beyond galled trees — Decomposition and computation of galled networks. In *RECOMB, The Annual International Conference on Research in Computational Molecular Biology*, pages 211–225. LNBI 4453, Springer, 2007.

[193] D. Huson, T. Klopper, P. Lockhart, and M. Steel. Reconstruction of reticulate networks from gene trees. In *RECOMB, The Annual International Conference on Research in Computational Molecular Biology*, pages 233–249. LNBI 3500, Springer, 2005.

[194] D. Huson, R. Rupp, V. Berry, P. Gambette, and C. Paul. Computing galled networks from real data. *Bioinformatics*, 25(12):i85–i93, 2009.

[195] D. Huson, R. Rupp, and C. Scornavacca. *Phylogenetic Networks*. Cambridge University Press, 2010.

[196] D. Huson and C. Scornavacca. A survey of combinatorial methods for phylogenetic networks. *Genome Biology and Evolution*, 3:23–35, 2010.

[197] J. R. Huyghe and K. L. Mohlke et al. Exome array analysis identifies new loci and low-frequency variants influencing insulin processing and secretion. *Nature Genetics*, 45:197–201, 2013.

[198] International HapMap Consortium. The HapMap project. *Nature*, 426:789–796, 2003.

[199] International HapMap Consortium. A haplotype map of the human genome. *Nature*, 437: 1299 1320, 2005.

[200] International HapMap 3 Consortium. Integrating common and rare genetic variation in diverse human populations. *Nature*, 467:52–58, 2010.

[201] P. Janvier. MicroRNAs revive old views about jawless vertebrate divergence and evolution. *Proceedings of the National Academy of Sciences (USA)*, 107:19137–19138, 2010.

[202] A. Javed and L. Parida. *Recombinomics*: Population genomics from a recombination perspective. In *Proceedings of the Third C* Conference on Computer Science and Software Engineering*, pages 129–137, 2010.

[203] A. Javed, M. Pybus, M. Melé, F. Utro, J. Bertranpetit, F. Calafell, and L. Parida. IRiS: Construction of ARG networks at genomic scales. *Bioinformatics*, 27(17):2448–2450, 2011.

[204] A. Jeffreys and R. Neumann. The rise and fall of a human recombination hot spot. *Nature Genetics*, 41:625–629, 2009.

[205] A. J. Jeffreys and C. A. May. Intense and highly localized gene conversion activity in human meiotic crossover hot spots. *Nature Genetics*, 36:151–156, 2004.

[206] A. J. Jeffreys and R. Neumann. Reciprocal crossover asymmetry and meiotic drive in a human recombination hot spot. *Nature Genetics*, 31:267–271, 2002.

[207] A.J. Jeffreys, L. Kauppi, and R. Neumann. Intensely punctated meiotic recombination in the class ii region of the major histocompatibility complex. *Nature Genetics*, 29:217–222, 2001.

[208] A.J. Jeffreys, R. Neumann, M. Panayi, S. Myers, and P. Donnelly. Human recombination hot spots hidden in regions of strong marker association. *Nature Genetics*, 37:601–606, 2005.

[209] P.A. Jenkins, Y.S. Song, and R.B. Brem. Genealogy-based methods for inference of historical recombination and gene flow and their application in *Saccharomyces cerevisiae*. *PLoS ONE*, 7:e46947, 2012.

[210] Z. Jiang, P. Pevzner, and E. Eichler et al. Ancestral reconstruction of segmental duplications reveals punctuated cores of human genome evolution. *Nature Genetics*, 39:1361–1368, 2007.

[211] L. Jin, P. Underhill, V. Doctor, R. Davis, P. Shen, L. Luca Cavalli-Sforza, and P. Oefner. Distribution of haplotypes from a chromosome 21 region distinguishes multiple prehistoric human migrations. *Proceedings of the National Academy of Sciences (USA)*, 96:3796–3800, 1999.

[212] H.R. Johnston and D. J. Cutler. Population demographic history can cause the appearance of recombination hotspots. *American Journal of Human Genetics*, 90:774–783, 2012.

[213] J. Kaiser. Genetic influences on diseases remain rare. *Science*, 338:1016–1017, 2012.

[214] K. Kalpakis and P. Namjoshi. Haplotype phasing using semidefinite programming. In *Proceedings of IEEE Conference on Bioinformatics and Bioengineering*, pages 145–152, 2005.

[215] S. Kannan and T. Warnow. Inferring evolutionary history from DNA sequences. *SIAM Journal on Computing*, 23:713–737, 1994.

[216] S. Kannan and T. Warnow. A fast algorithm for the computation and enumeration of perfect phylogenies when the number of character states is fixed. *SIAM Journal on Computing*, 26:1749–1763, 1997.

[217] E. K. Karlsson and K. Lindblad-Toh et al. Efficient mapping of Mendelian traits in dogs through genome-wide association. *Nature Genetics*, 39:1321–1328, 2007.

[218] J. D. Kececioglu and D. Gusfield. Reconstructing a history of recombinations from a set of sequences. *Discrete Applied Mathematics*, 88:239–260, 1998.

[219] S. Kelk, C. Scornavacca, and L. van Iersel. On the elusiveness of clusters. *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, 9:517–534, 2012.

[220] S. Kelk, C. Scornavacca, L. van Iersel, and C. Whidden. Personal communication, 2012.

[221] H. L. Kim, L. Hie, and S. Schuster. Poor man's 1000 genome project: Recent human population expansion confounds the detection of disease alleles in 7,098 complete mitochondrial genomes. *Frontiers in Genetics*, 4(13), 2013.

[222] G. Kimmel, R. Karp, M. Jordan, and E. Halperin. Association mapping and significance estimation via the coalescent. *American Journal of Human Genetics*, 83:675–683, 2008.

[223] G. Kimmel and R. Shamir. GERBIL: Genotype resolution and block identification using likelihood. *Proceedings of the National Academy of Sciences (USA)*, 102:158–162, 2005.

[224] G. Kimmel and R. Shamir. The incomplete perfect phylogeny haplotype problem. *Journal of Bioinformatics and Computational Biology*, 3:359–384, 2005.

[225] J. Kingman. On the genealogy of large populations. *Journal of Applied Probability*, 19:2743, 1982.

[226] B. Kirkpatrick. Haplotypes versus genotypes on pedigrees. In *WABI, Workshop on Algorithms in Bioinformatics*, volume 6293, pages 136–147. LNCS, Springer, 2010.

[227] C. Klein, K. Lohmann, and A. Ziegler. The promise and limitations of genome-wide association studies. *The Journal of the American Medical Association*, 308(18):1867–1868, 2012.

[228] J. Kleinberg and E. Tardos. *Algorithm Design*. Addison-Wesley Longman, 2005.

[229] A. Kong and K. Stefansson et al. A high-resolution recombination map of the human genome. *Nature Genetics*, 31:241–247, 2002.

[230] M. Kreitman. Nucleotide polymorphism at the alcohol dehydrogenase locus of *Drosophila melanogaster*. *Nature*, 304:412–417, 1983.

[231] M.K. Kuhner and J. Felsenstein. Sampling among haplotype resolutions in a coalescent-based genealogy sampler. *Genetic Epidemiology*, 19:S15–S21, 2000.

[232] V. Kunin, L. Goldovsky, N. Darzentas, and C. A. Ouzounis. The net of life: Reconstructing the microbial phylogenetic network. *Genome Research*, 15:954–959, 2005.

[233] P.Y. Kwok. Genomics: Genetic association by whole-genome analysis? *Science*, 294:1669–1670, 2001.

[234] L.A. Hindorff LA, J. MacArthur, J. Morales, H.A. Junkins HA, P.N. Hall, A.K. Klemm, and T.A. Manolio. A catalog of published genome-wide association studies. Available at: www.genome.gov/gwastudies.

[235] M. Lajoie and N. El-Mabrouk. Recovering haplotype structure through recombination and gene conversion. *Bioinformatics*, 21:1173–1179, 2005.

[236] F. Lam. Personal communication, 2010.

[237] F. Lam. Personal communication, 2012.

[238] F. Lam, D. Gusfield, and S. Sridhar. Generalizing the four gamete condition and splits equivalence theorem: Perfect phylogeny on three state characters. In S.L. Salzberg and T. Warnow, editors, *WABI, Workshop on Algorithms in Bioinformatics*, volume 5724 of *Lecture Notes in Computer Science*, pages 206–219. Springer, 2009.

[239] F. Lam, D. Gusfield, and S. Sridhar. Generalizing the four gamete condition and splits equivalence theorem: Perfect phylogeny on three state characters. *SIAM Journal on Discrete Math*, 25:1144–1175, 2011.

[240] F. Lam, C. H. Langley, and Y. S. Song. On the genealogy of asexual diploids. *Journal of Computational Biology*, 18:415–428, 2011.

[241] F. Lam, R. Tarpine, and S. Istrail. The imperfect ancestral recombination graph reconstruction problem. Upper bounds for recombination and homoplasy. *Journal of Computational Biology*, 17:767–781, 2010.

[242] J. Lam, K. Roeder, and B. Devlin. Haplotype fine mapping by evolutionary trees. *American Journal of Human Genetics*, 66:659–673, 2000.

[243] G. Lancia, C. Pinotti, and R. Rizzi. Haplotyping populations by pure parsimony: Complexity, exact and approximation algorithms. *INFORMS Journal on Computing, Special Issue on Computational Biology*, 16:348–359, 2004.

[244] E. Lander. Mapping heredity. In E. Lander and M. S. Waterman, editors, *Calculating the Secrets of Life*. National Academy Press, 1995.

[245] F. Larribe, S. Lessard, and N. J. Schork. Gene mapping via the ancestral recombination graph. *Theoretical Population Biology*, 62:215–229, 2002.

[246] S. Q. Le and R. Durbin. SNP detection and genotyping from low-coverage sequencing data on multiple diploid samples. *Genome Research*, 21:952–960, 2011.

[247] J. F. Lefebvre and D. Labuda. Fraction of informative recombinations: A heuristic approach to analyze recombination rates. *Genetics*, 178:20692079, 2008.

[248] W. J. LeQuesne. A method of selection of characters in numerical taxonomy. *Systematic Zoology*, 18:201–205, 1969.

[249] S. Levi, G. Sutton, and J. C. Venter et al. The diploid genome sequence of an individual human. *PLoS Biology*, 5, 2007.

[250] N. Lewis-Rogers, K.A. Crandall, and D. Posada. Evolutionary analysis of genetic recombination. In V. Parisi, V. de Fonzo, and F. Aluffi-Pentini, editors, *Dynamical Genetics*, pages 49–78. Research Signpost, 2004.

[251] J. Li and L. Cavalli-Sforza et al. Worldwide human relationships inferred from genome-wide patterns of variation. *Science*, 319:1100–1104, 2008.

[252] N. Li and M. Stephens. Modelling linkage disequilibrium, and identifying recombination hotspots using SNP data. *Genetics*, 165:2213–2233, 2003.

[253] J.Z. Lin, A. Brown, and M. T. Clegg. Heterogeneous geographic patterns of nucleotide sequence diversity between two alcohol dehydrogenase genes in wild barley (*Hordeum vulgare* subspecies *spontaneum*). *Proceedings of the National Academy of Sciences (USA)*, 98:531–536, 2001.

[254] S. Lin, D. Cutler, M. Zwick, and A. Chakravarti. Haplotype inference in random population samples. *American Journal of Human Genetics*, 71:1129–1137, 2002.

[255] S. Linz. *Reticulation in Evolution*. PhD thesis, University of Düsseldorf, 2008.

[256] S. Linz and C. Semple. Hybridization in nonbinary trees. *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, 6:30–45, January 2009.

[257] S. Linz, C. Semple, and T. Stadler. Analyzing and reconstructing reticulation networks under timing constraints. *Journal of Mathematical Biology*, 61:715–737, 2010.

[258] W. Lipski and F. Preparata. Efficient algorithms for finding maximum matchings in convex bipartite graphs and related problems. *Acta Informatica*, 15:329–346, 1988.

[259] X. Liu and Y. X. Fu. Algorithms to estimate the lower bounds of recombination with or without recurrent mutations. *BMC Genomics*, 9:S24, 2008.

[260] Y. Liu and C. Q. Zhang. A linear solution for haplotype perfect phylogeny problem (extended abstract). In *Advances in Bioinformatics and Its Applications*, pages 173–184. World Scientific, 2004.

[261] L. Löfgren. Irredundant and redundant Boolean branch networks. *IRE Transactions on Circuit Theory*, CT-6:158–175, 1959.

[262] J. R. Lupski. Genome mosaicism — One human, multiple genomes. *Science*, 341:358–359, 2013.

[263] R. Lyngsø. Specifying scoring schemes in Kwarg. http://www.stats.ox.ac.uk/~lyngsoe/section26/kwarg_scoring.pdf. [Online; accessed December 30, 2010].

[264] R. Lyngsø. Tools for recombination analysis in the coalescent. http://www.stats.ox.ac.uk/~lyngsoe/section26/. [Online; accessed December 30, 2010].

[265] R. Lyngsø, Y.S. Song, and J. Hein. Minimum recombination histories by branch and bound. In *WABI, Workshop on Algorithms in Bioinformatics*, volume 3692, pages 239–250. LNCS, Springer, 2005.

[266] B. Maher. The case of the missing heritability. *Nature*, 451:18–21, 2008.

[267] L. Mahler, T. Ingram, J. Revell, and J. Losos. Exceptional convergence on the macroevolutionary landscape in island lizard radiations. *Science*, 341:292–295, 2013.

[268] T. Mailund, S. Besenbacher, and M. Schierup. Whole genome association mapping by incompatibilities and local perfect phylogenies. *BMC Bioinformatics*, 7:454, 2006.

[269] E. Mancera, R. Bourgon, A. Brozzi, W. Huber, and L. Steinmetz. High-resolution mapping of meiotic crossovers and non-crossovers in yeast. *Nature*, 454:479–485, 2008.

[270] T. A. Manolio, L. D. Brooks, and F. S. Collins. A HapMap harvest of insights into the genetics of common disease. *Journal of Clinical Investgation 2*, 118(5), 2008.

[271] T. A. Manolio and F. S. Collins. The HapMap and genome-wide association studies in diagnosis and therapy. *Annual Review of Medicine*, 60:443–456, 2009.

[272] J. Marchini, D. Cutler, N. Patterson, M. Stephens, E. Eskin, E. Halperin, S. Lin, Z.S. Qin, H.M. Munro, G.R. Abecasis, and P. Donnelly. A comparison of phasing algorithms for trios and unrelated individuals. *American Journal of Human Genetics*, 78:437–450, 2006.

[273] J. Marchini, P. Donnelly, and L. Cardon. Genome-wide strategies for detecting multiple loci that influence complex diseases. *Nature Genetics*, 37:413–417, 2005.

[274] P. Marjoram and R. Joyce. Practical implications of coalescent theory. In L.S. Heath and N. Ramakrishnan, editors, *Problem Solving Handbook in Computational Biology and Bioinformatics*, pages 63–84. Springer, 2011.

[275] P. Marjoram and J. D. Wall. Fast "coalescent simulation". *BMC Genetics*, 7, 2006.

[276] B. Martin. Endosymbiosis and lateral gene transfer: Biologists know that the tree of life is not a tree, but what are mathematicians doing about it? Lecture Abstract for Current Challenges and Problems in Phylogenetics, Isaac Newton Institute, Cambridge University, September 2–7, 2007.

[277] D. P. Martin, P. Lemey, and D. Posada. Analysing recombination in nucleotide sequences. *Molecular Ecology Resources*, 11:943–955, 2011.

[278] E.R. Martin, J.R. Gilbert, and E.H. Lai et al. Analysis of association at single nucleotide polymorphisms in the APOE region. *Genomics*, 63, 2000.

[279] J. Matsieva. A static formulation of the history bound. Master's Thesis, University of California, Davis, Computer Science, 2014.

[280] J. Maynard-Smith and J. Haigh. The hitch-hiking effect of a favourable gene. *Genetical Research*, 23:23–35, 2 1974.

[281] T. A. McKee and F.R. McMorris. *Topics in Intersection Graph Theory*. SIAM Monographs on Discrete Mathematics, 1999.

[282] F. McMorris. On the compatability of binary qualitative taxonomic haracters. *Bulletin of Mathematical Biology*, 39:133–138, 1977.

[283] G. McVean and N. J. Cardin. Approximating the coalescent with recombination. *Philosophical Transactions of the Royal Society B*, 360:1387–1393, 2005.

[284] G. McVicker, D. Gordon, C. Davis, and P. Green. Widespread genomic signatures of natural selection in Hominid evolution. *PLoS Genetics*, 5:1–16, 2009.

[285] C. A. Meacham. Theoretical and computational considerations of the compatibility of qualitative taxonomic characters. In J. Felsenstein, editor, *Numerical Taxonomy*, pages 304–314. Springer-Verlag Nato ASI series Vol. G1, 1983.

[286] M. Minichiello and R. Durbin. Mapping trait loci using inferred ancestral recombination graphs. *American Journal of Human Genetics*, 79:910–922, 2006.

[287] T. Mitchell-Olds, J. H. Willis, and D. B. Goldstein. Which evolutionary processes influence natural genetic variation for phynotypic traits? *Nature Reviews: Genetics*, 8(11):845–856, 2007.

[288] B. Moret, L. Nakhleh, T. Warnow, C.R. Linder, A. Tholse, A. Padolina, J. Sun, and R. Timme. Phylogenetic networks: Modeling, reconstructibility, and accuracy. *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, pages 13–23, 2004.

[289] P.L. Morrell, D.M. Toleno, K.E. Lundy, and M.T. Clegg. Estimating the contribution of mutation, recombination and gene conversion in the generation of haplotypic diversity. *Genetics*, 173, 2006.

[290] A. P. Morris, J. C. Whittaker, and D. J. Balding. Fine-scale mapping of disease loci via shattered coalescent modeling of genealogies. *American Journal of Human Genetics*, 70:686–707, 2002.

[291] A.P. Morris, J.C. Whittaker, and D.J. Balding. Little loss of information due to unknown phase for fine-scale linkage-disequilibrium mapping with single-nucleotide-polymorphism genotype data. *American Journal of Human Genetics*, 74:945–953, 2004.

[292] D. Morrison. Networks in phylogenetic analysis: new tools for population biology. *International Journal for Parasitology*, 35:567–582, 2005.

[293] D. Morrison. Phylogenetic networks in systematic biology (and elsewhere). In *Research Advances in Systematic Biology*, pages 1–48, Trivandrum, India, 2009. Global Research Network.

[294] D. Morrison. *Introduction to Phylogenetic Networks*. RJR Productions, 2011.

[295] J. Mu, R. Myers, and X. Su et al. *Plasmodium falciparum* genome-wide scans for positive selection, recombination hot spots and resistance to antimalarial drugs. *Nature Genetics*, 42:268–271, 2010.

[296] S. Mukherjee. *The Emperor of All Maladies*. Simon and Schuster, 2010.

[297] M. Mutsuddi, D. Morriss, S. Waggoner, M. Daly, E. Scolnick, and P. Sklar. Analysis of high-resolution HapMap of DTNBP1 (Dysbindin) suggests no consistency between reported common variant associations and schizophrenia. *American Journal of Human Genetics*, 79:903–909, 2006.

[298] S. Myers. Personal communication, 2004.

[299] S. Myers. *The Detection of Recombination Events Using DNA Sequence Data*. PhD thesis, University of Oxford, Department of Statistics, 2003.

[300] S. Myers, L. Bottolo, C. Freeman, G. McVean, and P. Donnelly. A fine-scale map of recombination rates and hotspots across the human genome. *Science*, 310:321–324, 2005.

[301] S. Myers, R. Bowden, A. Tumian, R. Bontrop, C. Freeman, T. MacFie, G. McVean, and P. Donnelly. Drive against hotspot motifs in primates implicates the prdm9 gene in meiotic recombination. *Science*, 327(5967):876–879, 2010.

[302] S. Myers and R. C. Griffiths. Bounds on the minimum number of recombination events in a sample history. *Genetics*, 163:375–394, 2003.

[303] A. Nahajan and A. Morris et al. Genome-wide trans-ancestry meta-analysis provides insight into the genetic architecture of type 2 diabetes susceptibility. *Nature Genetics*, 46:234–244, 2014.

[304] L. Nakhleh. A metric on the space of reduced phylogenetic networks. *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, 7(2), 2010.

[305] L. Nakhleh. Evolutionary phylogenetic networks: Models and issues. In L. S. Heath and N. Ramakrishnan, editors, *Problem Solving Handbook in Computational Biology and Bioinformatics*, pages 125–158. Springer, 2011.

[306] L. Nakhleh, D. Ringe, and T. Warnow. Perfect phylogenetic networks: A new methodology for reconstructing the evolutionary history of natural languages. *Language*, 81:382–420, 2005.

[307] L. Nakhleh, J. Sun, T. Warnow, C.R. Linder, B.M.E. Moret, and A. Tholse. Towards the development of computational tools for evaluating phylogenetic network reconstruction methods. In *Proceedings of the Pacific Symposium on Biocomputing (PSB)*, pages 315–326, 2003.

[308] L. Nakhleh, T. Warnow, C.R. Linder, and K. St. John. Reconstructing reticulate evolution in species — Theory and practice. *Journal of Computational Biology*, 12:796–811, 2005.

[309] M. Nelson and V. Mooser et al. An abundance of rare functional variants in 202 drug target genes sequenced in 14,002 people. *Science*, 337:100–104, 2013.

[310] C. T. Nguyen, N. B. Nguyen, W. K. Sung, and L. Zhang. Reconstructing recombination network from sequence data: The small parsimony problem. *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, 4(3):394–402, 2007.

[311] D. Nickerson, S. Taylor, K. Weiss, and A. Clark et al. DNA sequence diversity in a 9.7-kb region of the human lipoprotein lipase gene. *Nature Genetics*, 19:233–240, 1998.

[312] R. Nielsen, I. Hellman, M Hubisz, C. Bustamante, and A. Clark. Recent and ongoing selection in the human genome. *Nature Reviews: Genetics*, 8:857–868, 2007.

[313] NIH. National Library of Medicine, Online Mendelian inheritance in man, http://www.ncbi.nlm.nih.gov/omim.

[314] M. Nikaido, A. P. Rooney, and N. Okada. Phylogenetic relationships among cetartiodactyls based on insertions of short and long interpersed elements: Hippopotamuses are the closest extant relatives of whales. *Proceedings of the National Academy of Sciences (USA)*, 96:10261–10266, 1999.

[315] M. Nordborg. Coalescent theory. In D. Balding, M. Bishop, and C. Cannings, editors, *Handbook of Statistical Genetics*, pages 179–212. Wiley, 2001.

[316] M. Nordborg and S. Tavare. Linkage disequilibrium: What history has to tell us. *Trends in Genetics*, 18:83–90, 2002.

[317] J. Novembre, J. K. Pritchard, and G. Coop. Adaptive drool in the gene pool. *Nature Genetics*, 39:1188–1190, 2007.

[318] K. O'Donnell, H. Krister, B. Tacke, and H. Casper. Gene fenealogies reveal global phylogeographic structure and reproductive isolation among lineages of *Fusarium graminearum*, the fungus causing wheat scab. *Proceedings of the National Academy of Sciences (USA)*, 97:7905–7910, 2000.

[319] Y. Okada and R. M. Plenge et al. Genetics of rheumatoid arthritis contributes to biology and drug discovery. *Nature*, 506:376–381, 2014.

[320] International Commission on Zoological Nomenclature. *INTERNATIONAL CODE OF ZOOLOGICAL NOMENCLATURE, Fourth Edition*. The International Trust for Zoological Nomenclature, 1999.

[321] S.H. Orzack, D. Gusfield, J. Olson, S. Nesbitt, L. Subrahmanyan, and Jr. V. P. Stanton. Analysis and exploration of the use of rule-based algorithms and consensus methods for the inferral of haplotypes. *Genetics*, 165:915–928, 2003.

[322] B. Padhukasahasram and B. Rannala. Bayesian population genomic inference of crossing over and gene conversion. *Genetics*, 189:607–619, 2011.

[323] F. Pan, L. McMillan, F. Pardo-Manuel De Villena, D. Threadgill, and W. Wang. TreeQA: Quantitative genome wide association mapping using local perfect phylogeny trees. In *Proceedings of the Pacific Symposium on Biocomputing*, pages 415–426. World Scientific Press, 2009.

[324] L. Parida. Ancestral recombinations graph: A reconstructability perspective using random-graphs framework. *Journal of Computational Biology*, 17:1345–1370, 2010.

[325] L. Parida. Combinatorics in recombinational population genomics. In *ISBRA, International Symposium on Bioinformatics Research and Applications*, pages 126–127. LNCS 6053, Springer, 2010.

[326] L. Parida. Non-reduncant representation of ancestral recombination graphs. In M. Anisimova, editor, *Evolutionary Genomics: Statistical and Computational Methods, Volume 2*, volume 856 of *Methods in Molecular Biology*, chapter 13, pages 315–332. Springer, 2012.

[327] L. Parida, A. Javed, M. Melé, F. Calafell, J. Bertranpetit, and the Genographic Consortium. Minimizing recombinations in consensus networks for phylogeographic studies. *BMC Bioinformatics*, 10(S-1), 2009.

[328] L. Parida, M. Melé, F. Calafell, J. Bertranpetit, and the Genographic Consortium. Estimating the ancestral recombinations graph (ARG) as compatible networks of SNP patterns. *Journal of Computational Biology*, 15:1133–1153, 2008.

[329] L. Parida, P. F. Palamara, and A. Javed. A minimal descriptor of an ancestral recombinations graph. *BMC Bioinformatics*, 12 (Suppl 1):S6, 2011.

[330] J. Paul and Y. S. Song. A principled approach to deriving approximate conditional sampling distributions in population genetics models with recombination. *Genetics*, 186:321–338, 2010.

[331] J. Paul, M. Steinrücken, and Y. S. Song. An accurate sequentially markov conditional sampling distribution for the coalescent with recombination. *Genetics*, 187:1115–1128, 2011.

[332] I. Pe'er, T. Pupko, R. Shamir, and R. Sharan. Incomplete directed perfect phylogeny. *SIAM Journal on Computing*, 33:590–607, 2004.

[333] E. Pennisi. Human evolution: More genomes from Denisova cave show mixing of early human groups. *Science*, 340:799, 2013.

[334] G. Perry and N. J. Dominy et al. Diet and evolution of human amylase gene copy number variation. *Nature Genetics*, 39:1256–1260, 2007.

[335] J. Pickrell, G. Coop, J. Novembre, and J. Pritchard et al. Signals of recent positive selection in a worldwide sample of human populations. *Genome Research*, 19:826–837, 2009.

[336] T. Piovessan and S.M. Kelk. A simple fixed parameter tractable algorithm for computing the hybridization number of two (not necessarilly binary) trees. *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, 10:18–25, 2013.

[337] C. Plowe. Malaria: Resistance nailed. *Nature*, 505:30–31, 2014.

[338] D. Posada and K. Crandall. Intraspecific gene genealogies: Trees grafting into networks. *Trends in Ecology and Evolution*, 16:37–45, 2001.

[339] D. Posada, K. A. Crandall, and E. Holmes. Recombination in evolutionary genomics. *Annual Review of Genetics*, 36:75–97, 2002.

[340] G.D. Poznik and C. D. Bustamante et al. Sequencing Y chromosomes resolves discrepancy in time to common ancestor of males versus females. *Science*, 341:562–565, 2013.

[341] J. K. Pritchard, J. K. Pickrell, and G. Coop. The genetics of human adaptation: Hard sweeps, soft sweeps, and polygenic adaptation. *Current Biology*, 20:R208–R215, 2010.

[342] S. Proulx, D. Promislow, and P. Phillips. Network thinking in ecology and evolution. *Trends in Ecology and Evolution*, 20, 2005.

[343] K. Prüfer and S. Pääbo et al. The complete sequence of a Neanderthal from the Altai mountains. *Nature*, 505:43 – 49, 2014.

[344] M. Przeworski. Motivating hotspots. *Science*, 310:247–248, 2005.

[345] T. Przytycka, G. Davis, N. Song, and D. Durand. Graph theoretical insights into evolution of multidomain proteins. *Journal of Computational Biology*, 13:351–363, 2006.

[346] P. Puigbo, Y. Wolf, and E. V. Koonin. The tree and net components of prokaryote evolution. *Genome Biology and Evolution*, 2:745–756, 2010.

[347] J. Qi, A. Wijeratne, L. Tomsho, Y. Hu, S Schuster, and H. Ma. Characterization of meiotic crossovers and gene conversion by whole-genome sequencing in *Saccharomyces cerevisiae*. *BMC Genomics (online)*, 10:475, 2009.

[348] M. Rasmussen, M. Hubisz, I. Gronau, and A. Siepel. Genome-wide inference of ancestral recombination graphs. ArXiv 1306.5110v3 [q-bio.PE] December 3, 2013.

[349] P. Rastas, M. Koivisto, H. Mannila, and E. Ukkonen. A hidden Markov technique for haplotype reconstruction. In *WABI, Workshop on Algorithms in Bioinformatics*, volume 3692, pages 140–151. LNCS, Springer, 2005.

[350] P. Rastas and E. Ukkonen. Haplotype inference via hierarchical genotype parsing. In *WABI, Workshop on Algorithms in Bioinformatics*, volume 4645, pages 85–97. LNCS, Springer, 2007.

[351] D. A. Ray, J. Xing, A-H. Salem, and M. A. Batzer. SINEs of the *nearly* perfect character. *Systematic Biology*, 55:928–935, 2006.

[352] R. Redon and M. Hurles et al. Global variation in copy number in the human genome. *Nature*, 444:444–454, 2006.

[353] N. Risch and K. Merikangas. The future of genetic studies of complex human diseases. *Science*, 275:1516–1517, 1996.

[354] R. J. Robbins. Introduction to the republication of the 1913 paper by A. H. Sturtevant: The linear arrangement of six sex-linked factors in *Drosophila*, as shown by their mode of association. Republication by Electronic Scholarly Publishing, www.esp.org, 1998.

[355] D.F. Robinson and L. R. Foulds. Comparison of phylogenetic trees. *Mathematical Bioscience*, 53:131–147, 1981.

[356] I. B. Rogozin, Y. I. Wolf, V. N. Babenko, and E. V. Koonin. Dollo parsimony and the reconstruction of genome evolution. In V. A. Albert, editor, *Parsimony, Phylogeny, and Genomics*. Oxford University Press, 2006.

[357] A. Rokas and P. Holland. Rare genomic changes as a tool for phylogenetics. *Trends in Evolution and Ecology*, 15:454–459, 2000.

[358] F. Rossell and G. Valiente. All that glisters is not galled. *Mathematical Biosciences*, 221:54–59, 2009.

[359] E. Ruark and M. Rahman et al. Mosaic PPM1D mutations are associated with predisposition to breast and ovarian cancer. *Nature*, 493:406–410, 2013.

[360] P. Sabeti, D. Reich, and E. Lander et al. Detecting recent positive selection in the human genome from haplotype structure. *Nature*, 419:832–837, 2002.

[361] P. Sabeti, S. Schaffner, and E. Lander et al. Positive natural selection in the human lineage. *Science*, 312:1614–1620, 2006.

[362] P. Sabeti, P. Varilly, and P. Fry et al. Genome-wide detection and characterization of positive selection in human populations. *Nature*, 449:913–918, 2007.

[363] R. Salari and S. Batzoglou et al. Inference of tumor phylogenies with improved somatic mutation discovery. In *RECOMB, The Annual International Conference on Research in Computational Molecular Biology*, pages 246–263. LNBI 7821, Springer, 2013.

[364] M. Sanderson and L. Hufford. *Homoplasy: The Recurrence of Similarity in Evolution*. Academic Press, 1996.

[365] I. Sandovici, S Kassovska-Bratinova, J. Vaughn, R. Stewart, M. Leppert, and C. Sapienza. Human imprinted chromosomal regions are historical hot-spots of recombination. *PLoS Genetics*, 2:944–954, 2006.

[366] R. V. Satya and A. Mukherjee. An optimal algorithm for perfect phylogeny haplotyping. In *Proceedings of the CSB Bioinformatics Conference*. IEEE Press, 2005.

[367] R. V. Satya and A. Mukherjee. An optimal algorithm for perfect phylogeny haplotyping. *Journal of Computational Biology*, 13(4):897–928, 2006.

[368] R. V. Satya and A. Mukherjee. The undirected incomplete perfect phylogeny problem. *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, 5:618–629, 2008.

[369] R.V. Satya, A. Mukherjee, G. Alexe, L. Parida, and G. Bhanot. Constructing near-perfect phylogenies with multiple homoplasy events. *Bioinformatics*, 22:e514–i522, 2006. Bioinformatics Supplement, Proceedings of ISMB.

[370] P. Scheet and M. Stephens. A fast and flexible statistical model for large-scale population genotype data: applications to inferring missing genotypes and haplotypic phase. *American Journal of Human Genetics*, 78:629–644, 2006.

[371] M. H. Schierup and J. Hein. Consequences of recombination on traditional phylogenetic analysis. *Genetics*, 156:879–891, 2000.

[372] C. Semple. Hybridization networks. In O. Gascuel and M. Steel, editors, *Reconstructing Evolution: New Mathematical and Computational Advances*, pages 277–309. Oxford University Press, 2007.

[373] C. Semple and M. Steel. *Phylogenetics*. Oxford University Press, 2003.

[374] P. Sevon, H. Toivonen, and V. Ollikainen. TreeDT: Tree pattern mining for gene mapping. *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, 3:174–185, 2006.

[375] A. J. Sharp and E. E. Eichler et al. Segmental duplications and copy-number variation in the human genome. *Amercian Journal of Human Genetics*, 77:78–88, 2005.

[376] C. J. Shaw and J. R. Lupski. Implications of human genome architecture for rearrangement-based disorders: The genomic basis of disease. *Human Molecular Genetics*, 13:R57–R64, 2004.

[377] S. Sheehan, K. Harris, and Y. S. Song. Estimating variable effective population sizes from multiple genomes: A sequentially markov conditional sampling distribution approach. *Genetics*, 194:647–66, 2013.

[378] B. Shutters and D. Fernandez-Baca. A simple characterization of the minimal obstruction sets for three-state perfect phylogenies. *Applied Math Letters*, 25:1226–1229, 2012.

[379] B. Shutters, S. Vakati, and D. Fernandez-Baca. Improved lower bounds on the compatibility of quartets, triplets, and multi-state characters. In B. Raphael and J. Tang, editors, *Algorithms in Bioinformatics*, volume 7534 of *Lecture Notes in Computer Science*, pages 190–200. Springer, Berlin, 2012.

[380] A. Siepel. Phylogenomics of primates and their ancestral populations. *Genome Research*, 19:1929–1941, 2009.

[381] R. Sladek and P. Froguel et al. A genome-wide association study identifies novel risk loci for type 2 diabetes. *Nature*, 445:881–885, 2007.

[382] F. Smagulova and G. Petukhova et al. Genome-wide analysis reveals novel molecular features of mouse recombination hotspots. *Nature*, 472:375–378, 2011.

[383] P. Sneath, M. J. Sackin, and R. P. Ambler. Detecting evolutionary incompatibilities from protein sequences. *Systematic Zoology*, 24:311–332, 1975.

[384] J. Soares and M. Stefanes. Algorithms for maximum independent set in convex bipartite graphs. *Algorithmica*, 53:35–49, 2009.

[385] Y. S. Song. On the combinatorics of rooted binary phylogenetic trees. *Annals of Combinatorics*, 7:365–379, 2003.

[386] Y. S. Song. Properties of subtree-prune-and-regraft operations on totally-ordered phylogenetic trees. *Annals of Combinatorics*, 10:129–146, 2006.

[387] Y. S. Song, Z. Ding, D. Gusfield, C. H. Langley, and Y. Wu. Algorithms to distinguish the role of gene-conversion from single-crossover recombination in the derivation of SNP sequences in populations. *Journal of Computational Biology*, 14:1273–1286, 2007.

[388] Y. S. Song and J. Hein. Parsimonious reconstruction of sequence evolution and haplotype blocks: Finding the minimum number of recombination events. In *WABI, Workshop on Algorithms in Bioinformatics*, volume 2812, pages 287–302. LNCS, Springer, 2003.

[389] Y. S. Song and J. Hein. On the minimum number of recombination events in the evolutionary history of DNA sequences. *Journal of Mathematical Biology*, 48:160–186, 2004.

[390] Y. S. Song, Y. Wu, and D. Gusfield. Efficient computation of close lower and upper bounds on the minimum number of needed recombinations in the evolution of biological sequences. *Bioinformatics*, 21:i413–i422, 2005. Bioinformatics Suppl. 1, Proceedings of ISMB 2005.

[391] Y. S. Song, Y. Wu, and D. Gusfield. Efficient computation of close lower and upper bounds on the minimum number of needed recombinations in the evolution of biological sequences. *Bioinformatics*, 21:i413–i422, 2005. Bioinformatics Suppl. 1, Proceedings of ISMB 2005.

[392] Y. S. Song, Y. Wu, and D. Gusfield. Haplotyping with one homoplasy or recombination event. In *WABI, Workshop on Algorithms in Bioinformatics*, volume 3692, pages 152–164. LNCS, Springer, 2005.

[393] Y. S. Song. A concise necessary and sufficient condition for the existence of a galled-tree. *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, 3:186–191, 2006.

[394] Y. S. Song and J. Hein. Constructing minimal ancestral recombination graphs. *Journal of Computational Biology*, 12:159–178, 2005.

[395] M. Steel. The complexity of reconstructing trees from qualitative characters and subtrees. *Journal of Classification*, 9:91–116, 1992.

[396] J. C. Stephens and J. F. Vovis et al. Haplotype variation and linkage disequilibrium in 313 human genes. *Science*, 293:489–493, 2001.

[397] J.C. Stephens. On the frequency of undetectable recombination events. *Genetics*, 112:923–926, 1986.

[398] M. Stephens, N. Smith, and P. Donnelly. A new statistical method for haplotype reconstruction from population data. *American Journal of Human Genetics*, 68:978–989, 2001.

[399] K. Stevens and D. Gusfield. Reducing multi-state to binary perfect phylogeny with applications to missing, removable, inserted, and deleted data. In V. Moulton and M. Singh, editors, *WABI, Workshop on Algorithms in Bioinformatics*, volume 6293 of *Lecture Notes in Computer Science*, pages 274–287. Springer, 2010.

[400] B. E. Stranger, A. C. Nica, and E. T. Dermitzakis. Populations genomics of human gene expression. *Nature Genetics*, 39:1217–1224, 2007.

[401] A. H. Sturtevant. The linear arrangement of six sex-linked factors in *Drosophila*, as shown by their mode of association. *Journal of Experimental Zoology*, 14:43–59, 1913.

[402] Y. Sun and J. Ambrose et al. Deep genome-wide measurement of meiotic gene conversion using tetrad analysis in *Arabidopsis thaliana*. *PLoS Genetics*, 8:e1002968, 10 2012.

[403] N. Sutter, C. D. Bustamante, and E. Ostrander et al. A single IGF1 allele is a major determinant of small size in dogs. *Science*, 316:112–115, 2007.

[404] K. Swenson, P. Guertin, H. Deschênes, and A. Bergeron. Reconstructing the modular recombination history of staphylococcus aureus phages. *BMC Bioinformatics*, 14(Suppl 15:S17), 2013.

[405] I. Tachmazidou, C. Verzilli, and M. De Iorio. Genetic association mapping via evolution-based clustering of haplotypes. *PLoS Genetics*, 3:e111, 07 2007.

[406] K. Tang, K. Thornton, and M. Stoneking. A new approach for using genome scans to detect recent positive selection in the human genome. *PLoS Biology*, 5:1587–1602, 2007.

[407] M. Tennesen and J. Akey et al. Evolution and functional impact of rare coding variation from deep sequencing of human exomes. *Science*, 337:64–69, 2013.

[408] K.M. Teshima, G. Coop, and M. Przeworski. How reliable are empirical genome scans for selective sweeps? *Genome Research*, 16:702–712, 2006.

[409] S. Tishkoff and F. Reed et al. Convergent adaptation of human lactase persistence in Africa and Europe. *Nature Genetics*, 39:31–40, 2007.

[410] W.T. Tutte. An algorithm for determining whether a given binary matroid is graphic. *Proceedings of the American Mathematical Society*, 11:905–917, 1960.

[411] E. Ukkonen. Finding founder sequences from a set of recombinants. In *WABI, Workshop on Algorithms in Bioinformatics*, volume 2452, pages 277–286. LNCS, Springer, 2002.

[412] F. Utro, O. Cornejo, D. Livingstone, J.C. Motamayor, and L. Parida. ARG-based genome-wide analysis of cacao cultivars. *BMC Bioinformatics*, 13: Suppl 19, 2012.

[413] L. van Iersel. *Algorithms, Haplotypes and Phylogenetic Networks*. PhD thesis, Technische Universiteit Eindhoven, the Netherlands, 2009.

[414] L. van Iersel. Different topological restrictions of rooted phylogenetic networks. Which make biological sense?, March 3, 2013. Available at: phylonetworks.blogspot.com.

[415] L. van Iersel, J. Keijsper, S. Kelk, L. Stougie, F. Hagen, and T. Boekhout. Constructing level-2 phylogenetic networks from triplets. *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, 6(4):667–681, 2009.

[416] L. van Iersel and S. Kelk. When two trees go to war. *Journal of Theoretical Biology*, 269:245–255, 2011.

[417] P. M. Visscher, M. A. Brown, M. I. McCarthy, and J. Yang. Five years of GWAS discovery. *American Journal of Human Genetics*, 90:7–24, 2012.

[418] B.F. Voight, S. Kudaravalli, X. Wen, and J.K. Pritchard. A map of recent positive selection in the human genome. *PLoS Biology*, 4, 2006.

[419] C. Wade and M. Daly et al. The mosaic structure of variation in the laboratory mouse genome. *Nature*, 420:574–578, 2002.

[420] D.B. Wake, M.H. Wake, and C.D. Specht. Homoplasy: From detecting pattern to determining process and mechanism of evolution. *Science*, 331:1032–1035, 2011.

[421] J. Wakeley. *Coalescent Theory*. Roberts and Co., 2009.

[422] J. D. Wall. A comparison of estimators of the population recombination rate. *Molecular Biology and Evolution*, 17:156–163, 2000.

[423] J. D. Wall. Close look at gene conversion hot spots. *Nature Genetics*, 36:114–115, 2004.

[424] J. D. Wall and J. K. Pritchard. Haplotype blocks and linkage disequilibrium in the human genome. *Nature Reviews — Genetics*, 4:587–597, 2003.

[425] A. Walsh, R. Kortschak, M. Gardner, T. Bertozzi, and D. Adelson. Widespread horizontal transfer of retrotransposons. *Proceedings of the National Academy of Sciences (USA)*, 110(3):1012–1016, 2013.

[426] L. Wang, K. Zhang, and L. Zhang. Perfect phylogenetic networks with recombination. *Journal of Computational Biology*, 8:69–78, 2001.

[427] J. Watson, N. Hopkins, J. Roberts, J. Steitz, and A. Weiner. *Molecular Biology of the Gene (4th edition)*. Benjamin Cummings, 1987.

[428] Wellcome Trust Case Control Consortium. Genome-wide association study of 14,000 cases of seven common diseases and 3,000 shared controls. *Nature*, 447:661–678, 2007.

[429] C. Whidden, R. Beiko, and N. Zeh. Fixed-parameter algorithms for maximum agreement forests. *SIAM Journal on Computing*, 42:1431–1466, 2013.

[430] H. Whitney. Congruent graphs and the connectivity of graphs. *American Journal of Mathematics*, 54:150–168, 1932.

[431] H. Whitney. 2-isomorphic graphs. *American Journal of Mathematics*, 55:245–254, 1933.

[432] E. O. Wilson. A consistency test for phylogenies based on contemporaneous species. *Systematic Zoology*, 14:214–220, 1965.

[433] W. Winckler, S. Myers, and D. Altschuler et al. Comparison of fine-scale recombination rates in humans and chimpanzees. *Science*, 308:107–111, 2005.

[434] C. Wiuf. Inference of recombination and block structure using unphased data. *Genetics*, 166:537–545, 2004.

[435] C. Wiuf, T. Christensen, and J. Hein. A simulation study of the reliability of recombination detection methods. *Molecular Biology and Evolution*, 18:1929–1939, 2001.

[436] C. Wiuf and J. Hein. Recombination as a point process along sequences. *Theoretical Population Biology*, 55:1217–1228, 1999.

[437] K. Wong and R. deLeeuw et al. A comprehensive analysis of common copy-number variations in the human genome. *American Journal of Human Genetics*, 80:91–104, 2007.

[438] Y. Wu. Personal communication, 2011.

[439] Y. Wu. Association mapping of complex diseases with ancestral recombination graphs: Models and efficient algorithms. In T. Speed and H. Huang, editors, *RECOMB, The Annual International Conference on Research in Computational Molecular Biology*, volume 4453, pages 488–502. LNBI, Springer, 2007.

[440] Y. Wu. Association mapping of complex diseases with ancestral recombination graphs: Models and efficient algorithms. *Journal of Computational Biology*, 15:667–684, 2008.

[441] Y. Wu. An analytical upper bound on the minimum number of recombinations in the history of SNP sequences in populations. *Information Processing Letters*, 109:427–431, 2009.

[442] Y. Wu. A practical method for exact computation of subtree prune and regraft distance. *Bioinformatics*, 25(2):190–196, 2009.

[443] Y. Wu. Bounds on the minimum mosaic of population sequences under recombination. In *Proceedings of the Annual Symposium on Combinatorial Pattern Matching*, volume 6129, pages 152–163. LNCS, Springer, 2010.

[444] Y. Wu. Close lower and upper bounds for the minimum reticulate network of multiple phylogenetic trees. *Bioinformatics*, 26:140–148, 2010.

[445] Y. Wu. New methods for inference of local tree topologies with recombinant SNP sequences in populations. *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, 8:182–193, 2011.

[446] Y. Wu. An algorithm for constructing parsimonious hybridization networks with multiple phylogenetic trees. In *RECOMB, The Annual International Conference on Research in Computational Molecular Biology*, volume 7821, pages 291–303. LNBI, Springer, 2013.

[447] Y. Wu and D. Gusfield. Efficient computation of minimum recombination with genotypes (not haplotypes). *Journal of Bioinformatics and Computational Biology*, pages 181–200, 2007.

[448] Y. Wu and D. Gusfield. Improved algorithms for inferring the minimum mosaic of a set of recombinants. In *Proceedings of the Annual Symposium on Combinatorial Pattern Matching*, volume 4580, pages 150–161. LNCS, Springer, 2007.

[449] Y. Wu and D. Gusfield. A new recombination lower bound and the minimum perfect phylogenetic forest problem. *Journal of Combinatorial Optimization*, 16:229–247, 2008.

[450] Y. Wu and J. Wang. Fast computation of the exact hybridization number of two phylogenetic trees. In M. Borodovsky, J.P. Gogarten, T.M. Przytycka, and S. Rajasekaran, editors, *ISBRA, International Symposium on Bioinformatics Research and Applications*, volume 6053, pages 203–214. Springer, 2010.

[451] B. Yalcin, J. Flint, and R. Mott et al. Genetic dissection of a behavioral quantitative trait locus shows that Rgs2 modulates anxiety in mice. *Nature Genetics*, 36:1197–1202, 2004.

[452] J. Yang and M. Visscher. Common SNPs explain a large proportion of the heritability for human height. *Nature Genetics*, 42:565–569, 2010.

[453] T. Yang, H-W Deng, and T. Niu. Critical assessment of coalescent simulators in modeling recombination hotspots in genomic sequences. *BMC Bioinformatics*, 15:3, 2014.

[454] M. Yeager and N. Orr et al. Genome-wide association study of prostate cancer identifies a second risk locus at 8q24. *Nature Genetics*, 39:645–649, 2007.

[455] J. Yin, M. Jordan, and Y.S. Song. Joint estimation of gene conversion rates and mean conversion tract lengths from population SNP data. *Bioinformatics*, 25:Pp. i231–i239, 2009.

[456] J. Zhang, W. Rowe, A. Clark, and K. Buetow. Genomewide distribution of high-frequency, completely mismatching SNP haplotype pairs observed to be common across human populations. *American Journal of Human Genetics*, 73:1073–1081, 2003.

[457] Q. Zhang, L. McMillan, and D. Threadgill et al. Genotype sequence segmentation: Handling constraints and noise. In *WABI, Workshop on Algorithms in Bioinformatics*, volume 5251, pages 271–283. LNCS, Springer, 2008.

[458] Q. Zhang, L. McMillan, and D. Threadgill et al. Inferring genome-wide mosaic structure. In *Proceedings of Pacific Symposium on Biocomputing*, pages 150–161. World Scientific Press, 2009.

[459] Z. Zhang, X. Zhang, and W. Wang. HTreeQA: Using semi-perfect phylogeny trees in quantitative trait loci study on genotype data. *G3*, 2:175–189, 2012.

[460] C. Zimmer. DNA doubletake. *New York Times*, September 16, 2013.

[461] S. Zöllner and J.K. Pritchard. Coalescent-based association mapping and fine mapping of complex trait loci. *Genetics*, 169:1071–1092, 2005.

[462] O. Zuk, E. Hechter, S.R. Sunyaev, and E. S. Lander. The mystery of missing heritability: Genetic interactions create phantom heritability. *Proceedings of the National Academy of Sciences (USA)*, 109:1193–1198, 2012.