Chapter 3

A Deeper Introduction to Recombination and Networks: The Biological, Mathematical and Algorithmic Contexts

In the last chapter we discussed necessary and sufficient conditions for binary sequences to be representable by a perfect-phylogeny. When there is a perfect-phylogeny (i.e., when all sites are pairwise compatible), it serves as a hypothesis for the actual evolutionary history of the sequences. However, a perfect-phylogeny does not exist for most sets of binary sequences encountered in populations because some pairs of sites are incompatible. The principle biological reason, in the context of populations (which means over a relatively short historical time period), is that *meiotic recombination* creates new, chimeric sequences in each generation. These changes are in addition to any changes created by point mutations, resulting in sequences that have incompatible sites. Thus, the main focus of this book concerns algorithmic questions about the evolution of sequences in populations, when *both* recombination and point mutation shape the sequences.

In this chapter we introduce recombination in both a biological context and in an algorithmic/mathematical context. Building on the general introduction to recombination given in Chapter 1, we give more formal and complete definitions of many central terms and models.

3.1 The biological and physical context of recombination

Repeating the quote from Watson from Chapter 1:

> "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..."

Moreover, recombination is central to many diverse biological phenomena, at molecular, population and evolutionary levels. For example:

> "Understanding the determinants of recombination is ... crucial for the study of genome evolution" [33].

And yet:

> "Little is known about the rules that govern the distribution of recombination events, although age, sex, DNA sequence, chromatin structure, chromosomal location, and chromosome sizes have been shown to be important" [166]

Meiotic recombination is a principle force creating sequence variation in populations, and has been observed to be involved with, and often central to, many other biological phenomena. However, there are many unresolved questions about the role of recombination in those phenomena, and there are many basic questions about recombination itself.

The central thesis of this book is that genealogical networks can be constructed by efficient algorithms and programs using genome variation data in populations of individuals, and that those networks reflect true recombination history sufficiently to help resolve or clarify some of these biological issues.

3.1.1 Meiotic Crossing-Over

The best understood form of recombination is called *single-crossover recombination*, also called *crossing-over* in the biological literature, where during meiosis two equal length sequences produce a third *recombinant* sequence of the same length, consisting of a *prefix* of one of the sequences, followed (at the "crossover point" or "breakpoint") by a *suffix* of the other sequence. See Figure 3.1.

As we have discussed, meiotic crossing-over is one of the major forces shaping genetic variation within a species. It allows the mixing of genes from the two "copies" of a chromosome, creating a new chimeric chromosome that can be passed on to a child, and hence it allows the rapid creation of hybrid chromosomes even without mutations.

In addition to its role in fundamental biological questions, meiotic crossing over is central to several critical applied problems. The most important example is "association mapping" in populations, a set of techniques that are widely hoped to help find genes that influence genetic diseases and important economic

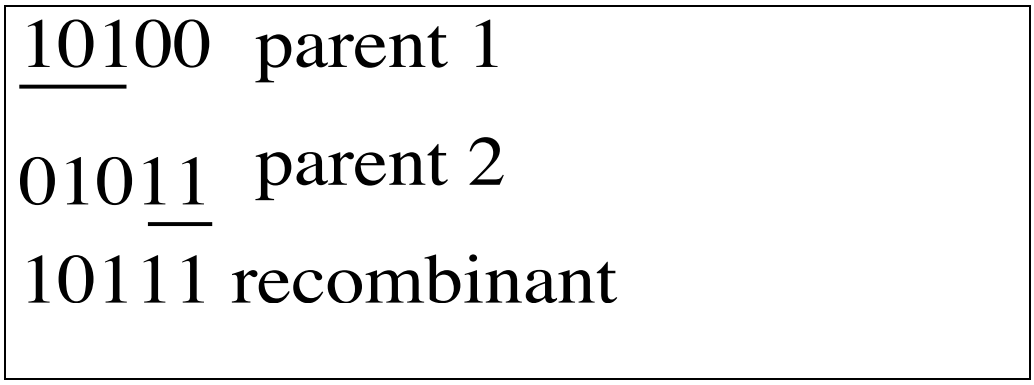

Figure 3.1: A single-crossover recombination. The prefix (underlined) contributed by Parent 1 consists of the first three characters of SNP sequence 1. The suffix (underlined) contributed by Parent 2 consists of the last two characters of SNP sequence 2.

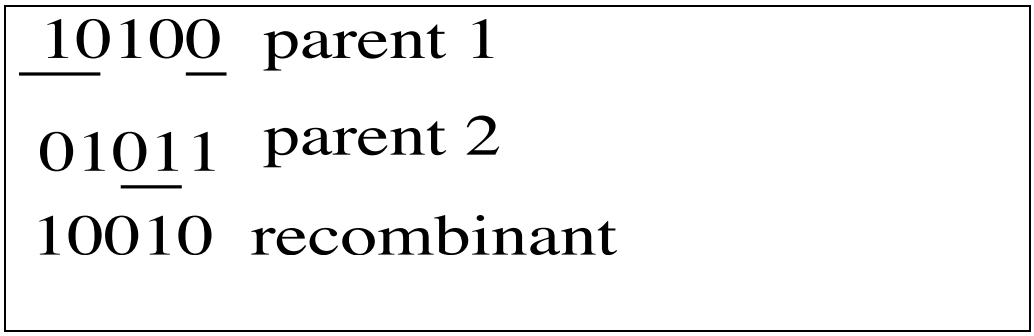

Figure 3.2: Double-crossover recombination. The prefix and suffix, underlined, contributed by Parent 1 consist of the first two characters, and the last character of sequence 1. The conversion tract, underlined, contributed by Parent 2 consists of the third and fourth characters of sequence 2.

traits in plants and animals. We will introduce the logic behind association mapping, and the crucial role of recombination in that logic, in Section 3.4.1. Association mapping will be discussed in more depth in Section **??**.

3.1.2 Double-Crossover, Gene-Conversion and Multiple-Crossover Recombination

Another form of meiotic recombination, called *double-crossover* recombination, creates a recombinant sequence from a prefix of one sequence, followed by an internal segment of a second sequence, followed by a suffix of the first sequence. Both parental sequences and the recombinant sequence are of the same length. See Figure 3.2.

A very common type of double-crossover recombination that occurs in meio-

sis is called "Gene-Conversion". In gene-conversion, the internal segment (called a "conversion track") is short, around 50 to 500 base pairs long, and it is copied from one sequence to another. Gene conversion is believed to be more common than single-crossover recombination, and is known to play a very important role in meiosis and in recombinational repair of damage to the DNA throughout development of an organism, and in all tissues. Until recently, it has been hard to study gene-conversion in populations, partly because of the lack of analytical tools and the lack of fine-scale data. For example, little is known about the distribution of track lengths. Gene conversion events that are mistaken for single-crossover recombination may also cause problems in association mapping and in other efforts to deduce information about recombination [90]. Eventually complete genomic resequencing will allow quantification of the fundamental parameters of gene conversion, and the contribution of gene conversion to the overall patterns of sequence variations in populations. A recent detailed study of gene conversion in Yeast appears in [**?**].

"Multiple-crossover" recombination, where a recombinant sequence is created from two sequences by *more than* two crossovers, occurs on a chromosomal scale, and is of biological importance in some applications. For example, the number of recombinations that occur between two sites on a chromosome is the basis for the concept of the *genetic distance* between those points. Genetic distance was the primary distance measure that could be easily obtained before DNA sequencing methods become widely available, and still remains important in some contexts. Note that when we refer to "multiple-crossover recombination", it is implied that the number of crossovers is greater than two; however there is no imposed upper bound on the number of allowed crossovers in a multiple-crossover recombination event.

3.1.3 The physical context

Generally, the binary sequences we are concerned with are SNP sequences, where adjacent positions in the SNP sequence represent DNA sites that could be physically far apart (perhaps separated by thousands of nucleotides). We think abstractly of recombination as occurring between two adjacent SNP sites (or before or after the first/last SNP site) in a SNP sequence, but physically, the recombination crossovers in DNA can occur anywhere on a chromosome. So a crossover point will generally be in an *interval* on the chromosome between two adjacent SNP sites in the SNP sequence. The SNP sequence does not represent the DNA that lies between or around SNP sites, and so we can only represent the location of a crossover *relative* to the SNP sites. It will sometimes be important to keep the physical reality in mind, for example when discussing the accuracy of methods to find the location of mutations, given SNP data.

3.1.4 Introduction to Hybridization and General Reticulation

Not written - defer to the end to see how much hybridization is discussed in the book.

3.2 The Algorithmic and Mathematical Context of Recombination

In our treatment of recombination, we abstract away the biological detail and focus on recombination as an operation on binary sequences. In this context, the key distinction is the number of crossovers allowed at a recombination event. We distinguish the cases of a *single-crossover* event, of a *double-crossover* event, and of a *multiple-crossover* event. Some algorithmic results apply only to single-crossover recombination, while some apply to single and double-crossover recombination, and some apply to multiple-crossover recombination.

From this point on, when we speak about "a recombination" or "a recombination event", we mean a single-crossover recombination, and we will use the term "double-crossover recombination" even if the biological basis of the event is a gene-conversion.

3.2.1 Representing a history of recombinations and mutations

As we saw in Chapter 2, the evolutionary history of a set of sequences that derive from a single ancestral sequence and are modified only by successive mutations, can be represented by a directed tree where each node represents a sequence and each edge represents a mutation. Tree representations work because mutation is an operation that creates one new sequence from one existing sequence. But recombination is an operation that creates a new sequence from *two* sequences and so the historically correct derivation of a set of sequences created by both mutations and recombinations cannot be represented by a tree. Moreover, if the sequences contain a pair of sites that are incompatible, no perfect-phylogeny (even a historically incorrect one) can derive the sequences. Instead, we represent a single recombination by two directed edges entering a node (see Figure 3.3), and we represent the derivation of a set of sequences as a *directed acyclic network* or a *directed acyclic graph (DAG)* (see Figure 3.4).

Terminological Confusion Depending on the underlying biological context in which the sequences are derived, and on the research community, the DAGs that are used to represent evolution have been called *phylogenetic networks*, *reticulate networks*, *recombination networks*, *genealogical networks*, *ancestral recombination graphs (ARGs)*, *hybridization networks*, and other terms. Specialized terminology for restricted classes of networks such as *galled-trees* (to be

discussed in Chapter 7), *galled-networks* and *level-k networks* etc. have also been used. The terminology has been evolving and is confusing and sometimes contradictory.

The biggest source of confusion is that the term "phylogenetic network" has been defined differently in different literatures, and in some literatures it has often been the only term used[1]. The confusion caused by the overuse of the term "phylogenetic network" motivates us now to use terms that make more precise distinctions between different kinds of networks and the different biological contexts in which they arise.

We follow the definition in [147, 103] that a *Phylogenetic Network* is "any graph used to represent evolutionary relationships (either abstractly or explicitly) between a set of taxa that labels some of its nodes (usually the leaves)". Under that broad definition, the networks discussed in this book are phylogenetic networks. See [100, 103, 147] for a taxonomy of many different biological networks that are called "phylogenetic networks".

We focus in this book on a particular subset of phylogenetic networks that derive a set of sequences from a set of ancestral sequences (almost always a single ancestral sequence). So for greater clarity, we will use the term "Genealogical Network" to refer to a general phylogenetic network that models the derivation of sequences by both mutation and recombination events, and note that the term does not specify the number of allowed crossovers at a recombination event, nor the number of times that a site can mutate. We will use the term "Ancestral Recombination Graph", abbreviated as "ARG" for a genealogical network that obeys the additional restriction that each site mutates on at most one edge of the network, and that only single-crossover recombinations are allowed. The assumption that each site mutates only once in an ARG is again a consequence of of the infinite sites assumption used in the definition of a perfect-phylogeny. These networks will be defined more formally in the next section. We will use the term "hybridization network" to refer to a network where a node with two incoming edges models the biological event of *species hybridization* or *lateral gene transfer*. Because of its prior overuse and broad meaning, we will rarely use the term "phylogenetic network", and will use the term "reticulate network" to refer in general to any of the networks defined here, when it is not required to classify the network more precisely.

3.2.2 Formal definitions for a Genealogical Network and an Ancestral Recombination Graph

We begin with a formal definition of a *Genealogical Network*, and then specialize it to an Ancestral Recombination Graph.

[1] In fact, in much of our own research papers, we used the term "phylogenetic network" for what we now refer to as a "genealogical network" or "ancestral recombination graph".

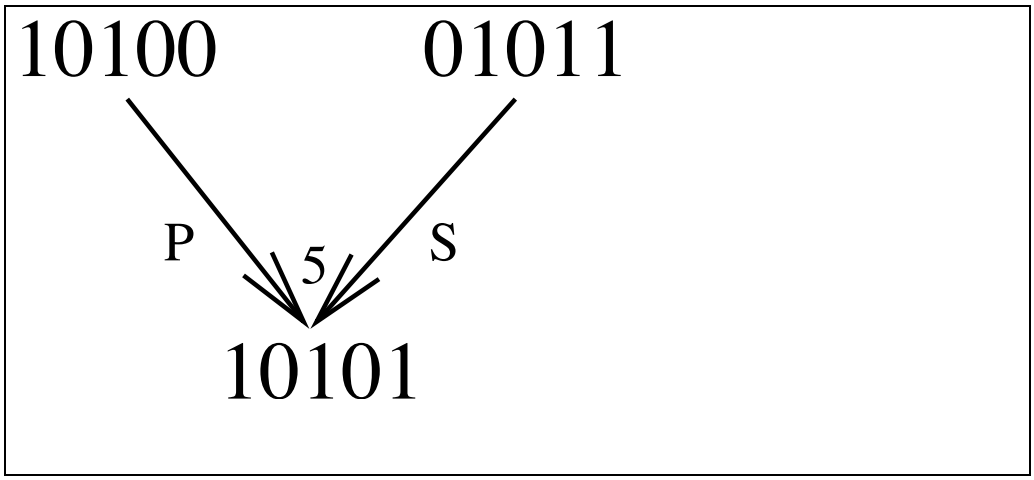

Figure 3.3: A graphical representation of a single-crossover recombination event. The contribution of the prefix is indicated by the character 'P' written on one edge, and the contribution of the suffix is indicated by the character 'S' written on the other edge. The crossover point, where the recombinant sequence begins to take characters from the suffix, is written above the recombination node.

There are four components needed to specify a genealogical network for a given set of binary sequences M.

1) (**The underlying graph**) Given a set of n binary sequences M, each of length m, a *genealogical network* $\mathcal{N}$ for M is built on a *directed acyclic graph* containing exactly one node (the root) with no incoming edges, a set of internal nodes that have both incoming and outgoing edges, and exactly n nodes, (the leaves), each with exactly one incoming edge and no outgoing edge. Each node other than the root has either one or two incoming edges. A node with a single incoming edge is called a *tree-node*; a node with two incoming edges is called a *recombination* node; a node with exactly one *incoming* edge and no out-going edges is a *leaf node*; and a node that is not the root node, nor a leaf node, is called an *internal* node. An edge into a recombination node is called a *recombination edge*. An edge whose removal disconnects the network, dividing it into two subnetworks, is called a *cut-edge*. Note that a cut-edge must be directed into a tree-node, but not all edges into tree-nodes are cut-edges.

The root node and any internal node can have any number of outgoing edges, representing the process of replication. An internal node with more than one outgoing edge is called a *branching internal* node, and an internal node with only a single outgoing edge is called a *non-branching internal* node.

2) (**The edge labels**) Each edge can be labeled with a set of integers from 1 to m, but can be unlabeled, and no labels are given to recombination edges. Note that the same integer might label different edges. The labels on an edge

represent mutations that occur in the time interval represented by the edge.

3) (**The node labels**) Each node in $\mathcal{N}$ is labeled by an m-length binary sequence, starting with the root node, which is labeled with some sequence r, called the "ancestral sequence" or the "root sequence". Since $\mathcal{N}$ is acyclic, the nodes in $\mathcal{N}$ can be (topologically) sorted into a list where every node occurs in the list only after its parent(s). Using that list, we can constructively label the non-root nodes with well-defined sequences in the order of their appearance in the list, as follows:

3a) (The tree-node labels) For a tree-node v, let e be the unique edge directed into v. The sequence labeling v is obtained from the sequence labeling v's parent by changing the state (from 0 to 1, or from 1 to 0) of site c, for every integer c that labels edge e. This corresponds to a mutation at site c occurring on edge e (i.e., during the interval of time represented by edge e).

3b) (The recombination-node labels) For a recombination node x, let Z and Z' denote the two m-length sequences labeling the two parent nodes of x. Then the "recombinant sequence" X labeling node x can be any m-length sequence provided that at every site c in X, the state (0 or 1) is equal to the state of site c in (at least) one of the sequences Z or Z'.

The creation of sequence X from Z and Z' at a recombination node is called a "recombination event", and models *multiple-crossover* recombination. To fully specify the recombination event, we must specify for every site c in X whether its parent sequence (contributing site c to X) is Z or Z'. This means specifying whether the state of c in X equals the state of c in Z or in Z'. This is forced when the states in Z and Z' of site c are different. When they are the same, a choice must be specified.

For a given recombination event at node x in $\mathcal{N}$, we say that a *crossover* or *breakpoint* occurs *between* sites c and $c+1$ if the states in X of sites c and $c+1$ come from different parents. If the crossover at x is between sites c and $c+1$, we set the *crossover index of* x, denoted b_x, to the *integer* $c+1$. When drawing network $\mathcal{N}$ (as in Figure 3.4), we display the crossover index b_x above the recombination node x to indicate that in the recombination event at node x, a change in the choice of parental sequence occurs at site $c+1$.

Sometimes we will want to determine the *minimum* number of crossovers needed to create sequence X by a recombination of specific sequences Z and Z'. That problem has an easy solution using a greedy algorithm that we will discuss later in the book.

As discussed earlier, in the case of *single-crossover* recombination, a recombinant sequence X is formed from a prefix of one of its parent sequences (Z or Z') followed by a suffix of the other parent sequence. This is consistent with the general definition of recombination given here.

4) (**The extant sequences**) The sequences labeling the leaves of $\mathcal{N}$ are the

observed, extant sequences, i.e., the sequences in M.

We say that a genealogical network $\mathcal{N}$ *derives (or explains or generates)* a set of n sequences M if and only if each sequence in M labels one of the leaves of $\mathcal{N}$.

Note that a "crossover point" or "breakpoint" refers to a physical location in a chromosome, while a "crossover index" is an integer that is part of the specification of an genealogical network[2]. If we know the true breakpoint for a recombination event, and it occurs between sites c and $c+1$ (which my be physically far apart in the chromosome), then the crossover index for the recombination event is $c+1$, even though the physical crossover could be very far from site $c+1$. When we don't know the exact (physical) breakpoint for a recombination event associated with a recombination node x, but do know that it must be between sites c and $d > c+1$ of M, we say that the crossover index b_x is in the interval $(c, d]$. Note that this interval is *open* on the left and *closed* on the right, meaning that the crossover index for node x must be specified by an integer b_x that is strictly larger than c and less than or equal to d. The asymmetry of the interval (open on the left and closed on the right) comes from the convention that the crossover index b_x specified at a recombination node indicates that a change in the choice of parental sequences contributing to the recombinant sequence, occurs at site b_x.

Since each of the n sequences in M has the same length, m, we will often consider the sequences arranged in an n by m matrix with one sequence per row, and refer to that matrix as M. In that case, we sometimes refer to a site or character as a "column".

3.2.2.1 Specializing Genealogical Networks to Ancestral Recombination Graphs

Definition Given a set of binary sequences M, an "Ancestral Recombination Graph (ARG)" for M is a genealogical network $\mathcal{N}$ that generates M, where each integer (site) from 1 to m labels *exactly* one edge in $\mathcal{N}$. See Figure 3.4.

In our treatment of ARGs, the default assumption is that every recombination event in an ARG is a *single-crossover recombination*, although often it doesn't matter which type of recombination is allowed . When it does matter, we will explicitly state whether single or multiple-crossover recombination is intended.

The assumption that each integer labels only a *single* edge is the standard "infinite sites model" in population genetics [90, 93], discussed in Chapter 1,

[2]In some treatments, the breakpoint is allowed to be a fractional number even though there are a finite number of discrete nucleotide sites on a chromosome. The practice of using fractional numbers comes from the view of a chromosome as a continuous object with essentially an infinite number of sites.

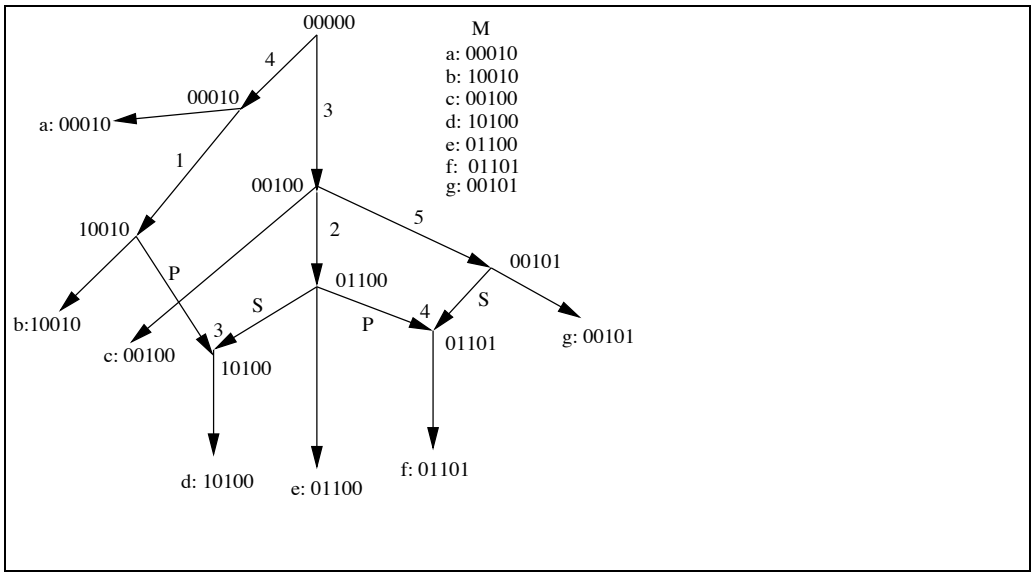

Figure 3.4: An ARG $\mathcal{N}$ with two recombination nodes. The matrix of sequences M that are derived by $\mathcal{N}$ is shown at the right. Note that the node with sequence label 01100 is sequence S for the left recombination node, and is sequence P for the right recombination node. In this example, every label of an interior node also labels a leaf, but that is not a general property of ARGs.

which is well supported in the context of evolution in a population (and hence over a short period of time).

ARGs and Coalescent Theory The term 'Ancestral Recombination Graph' arose in the the Population Genetics literature [64, 65]. More precisely an ARG is the graphical representation of the genealogical relations generated by the stochastic process called of the 'coalescent with recombination' [64]. This point is made in [135] as follows:

> ... we are using the term "ARG" to mean the data structure for representing genealogical histories. The distribution of these under the Wright-Fisher model with recombination is described by the stochastic process called the "coalescent-with-recombination" model.

See [150] or [90] or [195] for an introduction to coalescent theory, and to ARGs in the context of population genetics. With these definitions, the classic "root-unknown perfect-phylogeny", discussed in Chapter 2, is an ARG with *no* recombination nodes.

Although coalescent theory (with recombination) addresses the distribution of ARGs in the context of stochastic models, our focus in this book is on *structural, combinatorial, non-stochastic* features of ARGs. A critical distinction

between our treatment of an ARG as a digraph, and the distribution of ARGs generated by the coalescent-with-recombination, is that in the latter the edges of an ARG have lengths representing the passage of time. In our treatment of ARGs, edges have no lengths and the only temporal information in an ARG is the relative order of events implied by the reachability relation. Still, coalescent theory and the coalescent-with-recombination provides some of the insight underlying certain combinatorial methods and results discussed in this book.

Note that in the definitions of a genealogical network, there is no bound on the number of crossovers that are allowed at a recombination event (other than the number of sites minus one). Allowing an unbounded number of multiple crossovers is a convenient mathematical assumption that will allow us to model a wide variety of biological phenomena. In particular, it will be a way that we can apply some results about genealogical networks to problems concerning hybridization networks. However, as a biological reality, in *meiotic recombination* the number of crossovers is typically small, and the algorithmic/mathematical literature motivated by meiotic recombination has mostly assumed that only a single-crossover recombination is allowed at a recombination node. Single-crossover recombination is therefore the default case for our discussion of ARGs, although multiple-crossover recombination is permitted in the definition. In our discussions, we will explicitly state it when we allow multiple-crossover recombination. Multiple-crossover recombination occurs on a chromosomal scale, but in humans the number of crossovers on a single chromosome is typically under ten.

Additional helpful, but not limiting, assumptions For ease of exposition, there are several additional assumptions that we make about M and about any ARG for M. None of these assumptions limit the results obtained.

Generally, and without the need to state this each time, we assume that there is no site c where all the sequences in M have the same state. That is, every column of M contains both a 0 and a 1. This is not a limiting assumption, for if there were a site c in M where all the sequences have the same state, say one, and $\mathcal{N}$ is any ARG for all the sites in M other than c, we can insert c with state one, into the ancestral sequence for $\mathcal{N}$, obtaining an ARG for M.

We also assume that the root node of any ARG has at least two outgoing edges. To see that this is not a limiting assumption, note first that by definition, the root node of an ARG is not a leaf node, and every sequence in M must label a leaf, so if the ancestral sequence is in M, then the root must have at least two outgoing edges. So, if the root has only one outgoing edge, the ancestral sequence is not in M. Next, suppose that the root r has only one outgoing edge, and let P be the unique path from r to the first node v that has more than one outgoing edge. If any edge on P is labeled with a site c, then all sequences in M will have the same state of c, contradicting the assumption that no such site

exists. If no edges on P are labeled with a site, then all the edges of P can be contracted, making v the new root node, with the same ancestral sequence as before.

We further assume (unless otherwise stated) that every internal node has degree at least three[3]. If an internal node v has degree two, then it has one incoming edge and one outgoing edge, and these two edges can be merged without changing the set of sequences generated on the ARG. So we will assume that no such node v exists. Finally, We assume that M does not contain any duplicate rows (sequences), but note that M might contain duplicate columns.

3.2.3 There is no ARG-feasibility problem

We saw in Chapter 2 that not every set of sequences M can be derived on a perfect-phylogeny (with all-zero ancestral sequence), or on even an undirected perfect-phylogeny, where the ancestral sequence is not constrained. Therefore, the feasibility question of whether there is perfect-phylogeny for M is of interest. However, the feasibility-question for ARGs is not interesting because the answer is always "yes", as we show next.

Theorem 3.2.1 *For any set of binary sequences M, and any sequence s_r, there is an ARG $\mathcal{N}$ with ancestral sequence s_r that derives M. Further, $\mathcal{N}$ has at most $nm/2$ recombination nodes.*

Proof We prove this constructively. Using the allowed one mutation per site, create the sequence s'_r where every site c has the (binary) state that is the opposite of the state for c in s_r. The result is that in s_r and s'_r state 0 and state 1 appear at every site. Therefore, any binary sequence S can be created by recombining s_r and s'_r appropriately, using at most $m/2$ single-crossover recombinations. Therefore all the sequences in M can be generated using at most $nm/2$ recombination nodes. Since in the definition of an ARG, there is no bound on the number of allowed recombination nodes, and set of sequences M can be created in this way. ■

Therefore, unlike the perfect-phylogeny problem, the mere *existence* of an ARG $\mathcal{N}$ that generates M does not provide evidence that $\mathcal{N}$ has captured significant features of the true historical evolution of M, or even add evidence in support of the infinite sites model of mutation. To obtain biologically-plausible ARGs, we need to constrain the set of ARGs we produce, to focus on significant *properties* that such an ARG must poses. The most striking property of realistic ARGs is that they contain a relatively small number of recombination nodes, leading to the core algorithmic problem in ARG construction.

[3]Recall that the *degree* of a node v in a graph is the number of edges (in either direction) that touch v.

3.3 The Core Algorithmic Problem: Recombination Minimization

We now introduce one of the key computational problems that has been formulated in order to reconstruct plausible genealogical networks, and to study the extent and scope of historical recombination in populations. The problem is to determine or estimate the *minimum* number of recombination events needed to generate an observed set of binary sequences from an ancestral sequence (which may or may not be known), when the observed sequences were generated by both point mutations and recombinations. To make this problem precise, we must specify a model for permitted mutations. The most common mutation model is the infinite sites model, implying that any site (in the study) has mutated at most once in the entire history of the sequences. As observed earlier, this implies that each site in any of the studied sequences has taken on only two states. We now formalize the problem.

Definition Given a set of binary sequences M, we let $Rmin(M)$ denote the *minimum* number of *single-crossover* recombination events needed to generate the sequences M from any (unspecified) ancestral sequence, allowing only one mutation per site over the entire history of the sequences.

An alternative, equivalent definition is

Definition $Rmin(M)$ is the minimum number of recombination nodes that appear in any ARG $\mathcal{N}$ that derives M.

Sometimes we will explicitly emphasize that only single-crossover recombination events are allowed, and in that case we will use the notation "$R^1min(M)$" in place of "$Rmin(M)$".

To handle the case of multiple-crossover recombinations, we have the following

Definition $R^m min(M)$ is the minimum number of recombination nodes that appear in any ARG $\mathcal{N}$ that derives M, when a *multiple-crossover* recombination is allowed at any recombination node.

Sometimes the ancestral sequence is known and specified, and we need definitions that reflect that situation.

Definition $Rmin_S(M)$ is the minimum number of recombination nodes that appear in any ARG $\mathcal{N}$ that derives M, where $\mathcal{N}$ has ancestral sequence S.

Clearly, $Rmin_S(M) \geq Rmin(M)$ for any particular S.

A particularly useful case is when the ancestral sequence is the all-zero sequence.

Definition $Rmin_0(M)$ is the minimum number of recombination nodes that appear in any ARG $\mathcal{N}$ that derives M, where $\mathcal{N}$ has the all-zero ancestral sequence.

Definition The number of recombination nodes in an ARG $\mathcal{N}$ is denoted $R(\mathcal{N})$.

Definition An ARG $\mathcal{N}$ that derives a set of binary sequences M, where $R(\mathcal{N}) = Rmin(M)$, is called a *MinARG*; the problem of finding a MinARG for M is called the *MinARG Problem*.

Unable to resist the play on words, we also offer the following more formal definition relating a MinARG to an ArgMin:

$$\text{A MinARG for } M \text{ is an element of } \text{ArgMin}_{\texttt{ARG } \mathcal{N} \texttt{ for } \mathcal{M}}[R(\mathcal{N})].$$

The problem of computing $Rmin(M)$ (or computing closely related values) is NP-hard [198, 17, 18], and hence so is the problem of constructing a MinARG. Of course, $Rmin(M)$ is zero, and the MinARG is a perfect-phylogeny if and only if there are is no incompatible pair of sites in M.

Note that knowing a MinARG for M reveals $Rmin(M)$, but it is conceivable that we can determine $Rmin(M)$ without knowing any MinARG for M.

3.3.1 Why do we care about $Rmin(M)$ and MinARGs?

When the true ARG that generated a set of sequences is known (through simulations of ARGs and the sequences they derived), studies have empirically observed that $Rmin(M)$ is typically much lower than the true number of recombination events that occur, and somewhat lower than the number of *observable* recombination events that occur [98]. A recombination event in an ARG is "observable" if that recombination has a traceable effect on the extant sequences. For example, recombination between two identical sequences produces another identical copy of those sequences, and so that recombination has no effect on the set of extant sequences. More precisely, we could omit that recombination event (by removing one of the recombinant edges into the recombination node) and the resulting ARG would still generate the same set of sequences. Recombinations can be unobservable by other scenarios as well (we will be more precise about this later in the book). Hence the number of observable recombinations in an ARG is generally much less than the number of true recombinations that occurred in the generation of a set of sequences.

Despite the fact that $Rmin(M)$ can be lower than the number of observable recombinations that occurred in the true derivation of M, in this book, and in the research that has lead up to it, considerable attention is given to problems concerning the computation of $Rmin$ or the computation of information about $Rmin$, and to problems of finding a MinARG or a non-optimal but observably "near-optimal" ARG. The MinARG problem, and the problem of computing $Rmin$, are motivated by the general utility of *parsimony* in biological problems,

and because most evolutionary histories are thought to contain a small number of observable recombinations. Moreover, of all the statistics that we would like to determine concerning the history of recombinations, *Rmin* is one that can be concretely defined and, in principle, computed. Finally, even *lower bounds* on *Rmin* can be used to answer questions about recombination, such as finding potential recombination hotspots in genomes [54, 6, 207] and in estimating the recombination rate in observed haplotypes [196]. Similarly, explicitly computing a MinARG or a near-Min ARG has been useful in addressing biological problems, such as gene finding via association mapping [135, 204, 189] or finding haplotypes underlying genotype data [206, 207], or distinguishing the role of gene-conversion from single-crossover recombination [179, 180, 137].

It is easy to show that for every binary matrix M, there is an ARG that derives M using $O(nm)$ recombination nodes, but that is not of great interest because in most evolutionary histories the number of observable recombinations is thought to be relatively small, much smaller than nm. So in order to obtain biologically-informative results concerning recombination in populations, and to have the best chance at constructing an ARG that captures some (or all) of the correct historical features, we concentrate on the problems of computing $Rmin(M)$ and of constructing MinARGs.

The focus on minimizing the number of recombination nodes is further motivated by the fact that for any ARG $\mathcal{N}$ deriving n sequences, if $\mathcal{N}$ contains R recombination nodes and I internal, non-recombination nodes, then $I \leq n+R-2$, and the total number of nodes and edges are at most $2n+2R-1$ and $2n+3R-2$ respectively. That fact requires (as assumed) that every internal node has degree at least three. In fact, if every internal node has degree *exactly* three and the root has degree two (which are biologically sensible assumptions under the coalescent-with-recombination model), then $I = n + R - 2$, and the total number of nodes and edges are exactly $2n + 2R - 1$ and $2n + 3R - 2$ respectively. In that case, the size of $\mathcal{N}$ is captured by the single parameter R, the number of recombination nodes, so that the goal of minimizing the size of $\mathcal{N}$ is fully reflected in the goal of minimizing the number of recombination nodes in $\mathcal{N}$.

3.3.2 A robust literature

Although we cannot know for sure the history of mutations and recombinations that has created a given set of extant sequences, a robust literature has developed on *algorithms* to construct *plausible, biologically informative* genealogical networks, MinARGs, and near-MinARGs; or to study the history of recombinations; or to deduce well-defined aspects of a genealogy. This literature has grown particularly in the last twenty years, and includes [98, 88, 89, 109, 181, 182, 198, 142, 143, 78, 77, 202, 112, 45, 46, 127, 5, 183, 73, 74, 184, 126, 179, 180, 4, 6, 185, 169, 209, 210, 81, 207, 208, 75, 204, 138, 212, 135, 137, 66, 189, 211, 190, 153, 100, 99, 103, 101, 102, 135, 115, 205, 147, 41, 121, 122]. Related

questions about hybridization networks have also been addressed [145, 11, 102, 136, 146, 99, 101, 169, 10, 18, 147].

The need for networks instead of simple trees has long been understood by population geneticists, formalized by the *coalescent with recombination*, or *Ancestral Recombination Graph.* This understanding has been more recently matched by a a wider range of evolutionary biologists who have presented the view that many evolutionary phenomena must be represented by networks rather than by trees [42, 43, 157, 158, 139, 140, 63, 129, **?**].

In addition to helping to recreate and understand history, explicit genealogical networks, even if they do not completely capture the true history, can also allow better solution of many biological problems than do methods based on the more commonly used *numerical* measures that only indirectly reflect the underlying history [4].

3.4 Two examples of current problems where recombination is central

To further motivate the importance of understanding patterns of recombination in populations, we discuss here two current, high-visibility, problems and approaches to their solution. These two illustrations are highly simplified, with the intent of showing the role of recombination in the *logic* of the solutions, particularly for readers who may not have had any exposure to these problems or solutions. The first solution also illustrates the utility of computationally reconstructing explicit plausible genealogical networks.

3.4.1 An idealized introduction to association mapping: one use for the true genealogical network

Perhaps the most important practical application of methods to deduce information about recombination is in the search for *causal genes* using population-based *association mapping*, a method that has long been hoped to be able to efficiently locate genes that contribute to genetic diseases or to important agri-

[4] For example, most of what has been inferred about recombination rates in humans comes from the use of "statistical methods ... based on patterns of linkage disequilibrium (LD)" [159], essentially measures of correlation between the states (alleles) of pairs of sites in the genome [196]. The logic is that if the states of two sites are highly correlated (in a set of sampled sequences from a population) then one may conclude that there has been little past recombination between those two sites. But this indirect reflection of past recombination has several well-known problems. One is that in the most common measure of LD, called r^2, high levels of recombination in a region leads to low r^2, but low r^2 can also occur in regions of low recombination, depending on the location of the mutations on the true ARG deriving the sequences [132, 90]. That problem would be avoided if we could examine the true ARG (or important features of it) that generated the sequences.

cultural/commercial traits [28]. The method has been successful in verifying that specific *candidate genes* are associated with certain genetic traits (particularly simple Mendelian traits), and recently, several major successes in *genome-wide* association mapping [128] (where one does not have any initial conjecture of where the causal mutations are located) have been reported and verified[5] [32, 173, 31].

In this section we discuss a very simplified, idealized example of association mapping that illustrates the utility of knowing the true genealogical network that derived the extant SNP sequences. In the example, we consider a *pure-Mendelian* genetic disease. A disease is pure-Mendelian if it is caused by a mutation at a *single* fixed site in the genome, and everyone who has a specific allele at that site has the disease, and no one else has the disease[6]. So there is a single ("causal") site c^* in the genome, and a single causal state ("allele") i for c^*, such that any individual in the population will have the disease if and only if they have state i at site c^*. We assume that the state of site c^* mutated to i only once in the history of the population (reflecting the infinite sites model). We also assume that we can correctly identify the individuals (called the "cases") who have the disease, and hence identify the individuals (called the "controls") who do not. These assumptions are idealized, but allow a simple example that explains the biological basis of association mapping and illustrates the role of recombination.

We are given a set of SNP sequences M for the cases and the controls, and it is assumed that the causal site c^* is located somewhere in the genome spanned by the SNP sites in M; however c^* need not be (and generally will not be) one of the SNP sites in M. The (association mapping) problem is to use this population data to bracket the location of c^* in the genome as precisely as possible.

To show the critical role of recombination in solving this problem, suppose we also know the true genealogical network $\mathcal{N}$ that derived the SNP sequences in M. Consider the ARG shown in Figure 3.5, which is similar, but not identical to the ARG in Figure 3.4. Individuals e and f have the disease, but none of the other individuals have it.

Recall that ARG $\mathcal{N}$ represents the evolution of the entire DNA molecules

[5]However, many of the mutations found only explain a small percentage of the genetic influence on the disease. A major assumption underlying the hope that association mapping would be more productive is the 'common disease, common gene" assumption. That assumption was that the influence of genes on common diseases (such as hypertension) would be due to a small number of genes. What now appears to be the case is that many common diseases are affected by many different genes and different mutations, each accounting for a small percentage of the genetic influence. It is however expected that genome-wide association will be more successful when based on full DNA sequence data rather than SNPs, and when tens of thousands of individuals are sampled.

[6]There are pure-Mendelian diseases, but they are rare. More often, even if a disease is caused by a single specific allele at a single fixed site, not everyone who has that allele will get the disease, and there may also be other causes of the disease.

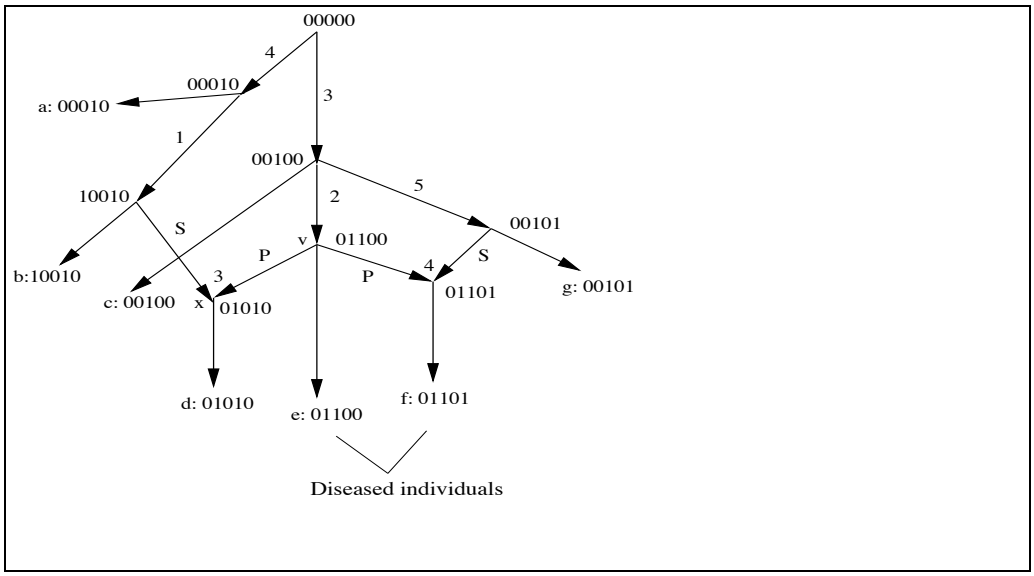

Figure 3.5: The "true" ARG displaying the derivation of the SNP sequences for seven individuals. We assume that individuals e and f have a pure-Mendelian disease, caused by a mutation at a single unknown causal site, c^*, in their genome. Note that the choice of P and S edges into recombination node x reverses the usual layout.

that the seven extant individuals receive, even though the only data we have is for particular SNP sites in those molecules. ARG $\mathcal{N}$ can be used to deduce some of the evolutionary history of the unknown site c^*. It is useful to think of a physical DNA molecule originating at the root of the ARG and then, by replications (represented by two or more out-edges leaving a node), mutations and recombinations, descending and evolving through the ARG, finally arriving at the leaves, delivering the DNA molecules to the individuals represented there.

The first thing we can deduce is the edge (and the interval of time that it represents) where site c^* must have mutated to the causal state. The mutation must have occurred on an edge that is ancestral to (leads to) leaves e and f, so it must either be on the edge where sites 2 mutated or the edge where site 3 mutated. However, if it was on the edge where site 3 mutated, then individual c (and g) would have the disease, since that edge is ancestral to leaf c via a path that does not contain a recombination node. We conclude that the mutation must have occurred on the edge labeled 2, and that the DNA molecule that arrives at node v in Figure 3.5 must have the causal state of site c^*. Knowing the edge where site c^* mutates brackets the time of the mutation but not the location of c^* in the genome.

Next, we note that individual d does not have the disease, even though the edge labeled 2 is ancestral to d, and d receives part of its DNA via node v. Since

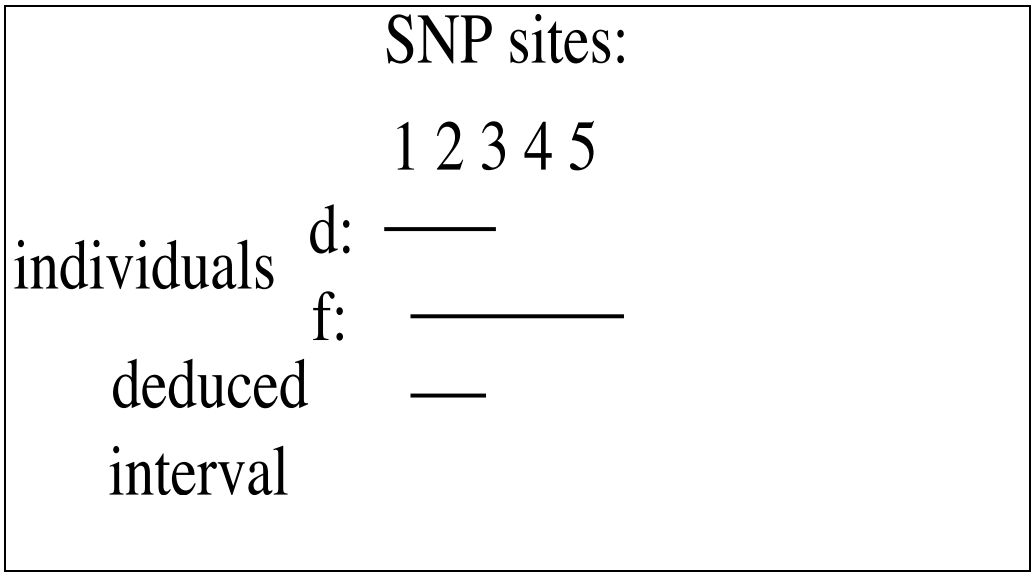

Figure 3.6: The intervals (deduced from individuals d and f) where c^* might be located, relative to the five SNP sites.

the disease is pure-Mendelian, the reason d does not have the disease must be due to the recombination event at node x, allowing d to receive the DNA at site c^* from its suffix-contributing parent (who does not have the causal state of c^*), rather than from its prefix-contributing parent (at node v), who has the causal state of c^*. Thus we can bracket the location of c^* in the genome by asking what segment(s) of DNA individuals d, e and f obtain via node v? Site c^* must be located in a segment that individual f receives via node v, and must not be in the segment that individual d receives via node v. Individual e obtains all of its DNA via node v and so information about e does not help bracket the location of c^*.

Individual d obtains DNA from v that starts at the left end of the chromosome and ends at some crossover point between SNP sites 2 and 3 (recall that the physical location of a crossover can be between two adjacent SNP sites). Therefore, c^* must be to the right of SNP site 2. Individual f obtains DNA from v that starts at the left end of the chromosome and ends at some crossover point between SNP sites 3 and 4. So c^* must be to the left of SNP site 4 (see Figure 3.6). Therefore, we can conclude that c^* must be in the interval of DNA between SNP sites 2 and 4. No finer localization of c^* is deducible from this data.

This idealized example shows the biological basis for association mapping, and the logic of identifying and examining segments of DNA that are common to the cases but not to the controls. Recombination reduces the size of contiguous intervals in the genome that descend intact to the an individual from any

ancestor of that individual. This shortening over time is one of the key elements that makes association mapping possible.

What to do when the ARG is not known Clearly, the true genealogical network of the extant SNP sequences, showing the locations (both in the genome and in time) of all mutations and recombinations, explicitly reveals segments shared by cases but not controls, at least up to the resolution of the SNP sites. Therefore, having an explicit genealogical network would greatly facilitate association mapping. However, we don't generally know the true genealogical network of the SNP sequences, and most current association mapping methods do not attempt to deduce a full genealogy, or even partially deduce features of the genealogy. Instead, they use features of the data that can be explained by reference to ARGs but can be determined without knowing an ARG.

To explain this, we again use a simple scenario. As before, we suppose that the causal mutation happened once in the history of the population, so that the cases must have inherited a segment of DNA that contains the causal site from a single common ancestor (called the "founder"), while controls did not inherit their DNA segment from the founder. Thus the cases should have a highly similar segment of DNA around the causal site. In contrast, that segment of DNA will be different and more diverse in the controls, since it did not originate from the founder, and may have originated from many different individuals.

In principle then, to bracket the location of the causal site, one looks for an interval in the SNP sequences (or even a single SNP site) where some (unknown) pattern occurs more frequently in cases than in controls. Because of mutations, the "pattern" might not be identical in all the cases, but it should be highly similar. Over time, recombination shortens the length of shared, intact segments that bracket the causal site, making it plausible that the casual site can be finely located. An even cruder reflection of a "pattern" is that the states of two SNPs that are physically close to the causal site should be more highly correlated among the cases than among the controls. Therefore, to help locate the causal site, one looks for pairs of SNPs whose states are highly correlated among the cases but not the controls. This kind of correlation is the basis for the notion of "linkage disequilibrium (LD)", and for measures of LD that are used in most association mapping methods [196].

Measures of LD reflect, but do not require or expose explicit genealogical networks, or necessarily correlate well with levels of recombination in different regions[7]. It is well understood that a full genealogical network contains more information than does any of the more indirect numerical reflection of it: "The

[7] "It is customary in genomics for researchers to debate which measure of linkage disequilibrium to use to characterize the joint distribution of variation at linked sites. The correct answer is 'none of them' ... one needs a full coalescent calculation" [56], i.e., a calculation that considers the space of all ARGs that could have generated the data.

best information that we could possibly get about association is to know the full coalescent genealogy ... " [212]. Further "If the true ARG were known, it would provide the optimal amount of information for mapping – no extra information would be available from the genotypes. Not only would disease-associated regions be identified, but the ARG would give the ages of the causative mutations, would specify the haplotypic background of those mutations and so forth. It would also be possible to optimally impute missing data." [135]

3.4.2 A brief discussion on locating signatures of recent positive selection

With the recent and increasing availability of data on genomic variation in humans and other species, population geneticists and evolutionary biologists have developed several methods to detect signatures of positive selection [149, 163, 164, 165, 133, 194, 156]. In this section, we introduce the logic of one of the newer, more sophisticated methods, the "Extended Haplotype Homozygosity" method [163], which we will call the "long-haplotype" method. This method, and variants of it, has been used to identify chromosomal regions containing likely causal mutations for recent positively selected traits, or to suggest that specific traits of interest were recently positively selected, or to provide evidence that known causal mutations for known traits are recent and that the trait were positively selected. Most notably, it has been used to detect positive selection in humans that occurred within the last 30,000 years. Recombination is at the heart of the logic of the long-haplotype method.

We will later give a more biologically informed definition of a *haplotype*, but for now the term can be read as: a sequence observed in a population, where the sequence is from a segment of one of the two "copies" of some particular chromosome. See Figures 3.7 and 3.8.

3.4.2.1 Positive Selection

Positive Section refers to a process through which a beneficial genetic trait becomes more frequent in a population. When a heritable genetic trait first appears in a population, for example through a mutation in a germ-line cell (egg or sperm), it appears in a single individual called a "founder". Through purely stochastic effects of uneven numbers of offspring (i.e, random drift), the trait can go extinct in the population, or with a much lower probability, the frequency of the trait can increase over time. But even if random drift causes the frequency of the trait to increase, the increase will be very slow and over short time periods it should occur in very low frequency in the population. However, if the (heritable) trait is beneficial, i.e., contributes to an individual having more viable offspring than do individuals without the trait, the frequency of individuals who have the trait will likely increase in successive generations, and the rate of increase

	1	2	3	4	5	6
person 1: haplotype 1	A	G	G	C	C	A
person 1: haplotype 2	A	A	G	C	C	T
person 2: haplotype 1	A	G	G	C	C	T
person 2: haplotype 2	A	G	G	C	C	A
person 3: haplotype 1	A	G	A	T	T	A
person 3: haplotype 2	A	G	A	T	T	A
person 4: haplotype 1	G	G	A	T	C	A
person 4: haplotype 2	A	G	A	T	T	A

Figure 3.7: Four hypothetical pairs of SNP haplotypes in four individuals. These haplotypes, but not the pairings, are the haplotypes in the Human Dysbindin gene on chromosome six reported in [141]. Each individual has two "copies" of chromosome six, and so has two haplotypes in this region. Notice that there are five distinct haplotypes among the eight haplotypes possessed by the four individuals. These are shown in Figure 3.8. The reported phylogenetic history of these haplotypes is shown in Figure 2.2 (page 24).

	1	2	3	4	5	6
42%	A	G	G	C	C	A
6%	A	A	G	C	C	T
33%	A	G	G	C	C	T
8%	A	G	A	T	T	A
11%	G	G	A	T	C	A

Figure 3.8: The five distinct reported SNP haplotpyes in the Human Dysbindin gene, and their reported frequencies in the sampled population [141].

can be very rapid. Such a trait is said to be "under positive selection" or just "under selection" and to "sweep" the population (or sweep away the variation in the population); ultimately, if (nearly) everyone in the population comes to have that trait, the trait is said to be "fixed" in the population.

The speed by which a beneficial trait under positive selection can sweep the population, compared to random drift, can be very dramatic. It is known [87] that *if* a trait becomes fixed in a population due to random drift alone, the expected number of generations until fixation is proportional to the size of the "effective population" (which crudely can be thought of as the size of a virtual subpopulation in which there is random mating). In humans, the effective population size is believed to be somewhere between three and ten thousand. So, even if a human trait becomes fixed due to random drift (and there is a much higher probability that it will go extinct), the expected time until fixation is proportional to thousands of generations. In contrast, there are beneficial traits such as changes in the color, size and patterns of spots that act as protective camouflage in fish that have been *observed* to become fixed in a population in a handful of generations [47]. Antibiotic Resistance in certain bacteria, which is a beneficial trait to the bacteria if not to us, is another well-known trait that has swept populations of bacteria in a matter of decades.

It is of great interest to identify traits that have swept a population due to positive selection. The issue discussed in this section is how such traits can be identified when we cannot observe the trait frequencies from the past, but only observe genomes in the current population.

3.4.2.2 Identity by descent without recombination

To understand the main idea to come, it is helpful to consider first the situation where no recombination occurs, and to assume that there is only a single causal mutation for a beneficial trait. That mutation initially appears in a single individual, the founder, and in a single chromosome of the founder, called the "enclosing chromosome". Moreover, if at the time of the mutation, individuals in the population are not identical, then the founder's enclosing chromosome will be distinguishable from the enclosing chromosome possessed by other members of the population. Without recombination, any descendant of the founder who inherits the trait will also inherit a "copy" of the founder's enclosing chromosome. Over time, additional mutations will occur, but overwhelmingly (again without recombination) the enclosing chromosome of an individual with the trait will be highly similar to that of the founder, and distinguishable from the enclosing chromosomes of individuals without the trait. Therefore, in the current population, the enclosing chromosomes of the set of individuals who have the trait will be highly similar to each other, and will be distinguishable from the enclosing chromosomes of individuals who lack the trait. Individuals with the trait will have enclosing chromosomes that are said to be (nearly) "identical by

descent".

Combining the concept of identity by descent with the discussion of positive selection, it follows that without recombination, the frequency in a population of the copy of the enclosing chromosome of a positively selected trait will increase rapidly in successive generations. And, unless the trait becomes fixed in the population, there will be an identifiable subpopulation, the individuals with the trait, whose enclosing chromosomes are highly similar to each other, and are distinguishable from the enclosing chromosome of individuals who lack the trait. But these facts alone don't allow us to recognize that the trait was positively selected because we don't know how rapidly the frequency of the trait increased. To make that recognition, we need to introduce the effect of recombination.

3.4.2.3 Recombination and haplotype length

Now we consider how recombination changes the story, leading to the second component in the long-haplotype method. Recall that for clarity of the exposition, we have assumed that the positively selected trait is caused by a single mutation. Without recombination, the frequency of the founder's enclosing chromosome will increase rapidly, but since we can't observe the past we don't know how rapidly the frequency increased. However, recombination creates chimeric chromosomes and *reduces* the length of the segment around the site of the causal mutation. Therefore, the length of the chromosome segment that is highly similar in the individuals with the trait, and distinguishable from the segments contained by individuals without the trait, *decreases* as recombinations occur, and hence decreases over time. See Figure **??**.

Due to recombination, it is no longer true that the entire enclosing chromosomes of individuals with the trait will be highly similar, or that the entire enclosing chromosome will be distinguishable from the entire enclosing chromosomes of individuals who lack the trait. However, if the time since the causal mutation is not too great, so that recombination has not reduced the length around the causal mutation to something too small to identify, there will still be some identifiable interval around the location of the causal mutation where the chromosomes of the individuals with the trait will be highly similar, and distinguishable from the interval in individuals who lack the trait. See Figure **??**. Finally, this leads to a more valid definition of a haplotype.

Definition A *haplotype* is a segment of one copy of a chromosome that is observed in a population of individuals, and is identical by descent, i.e., where the contiguous loci in the segment are transmitted together to a set of individuals in the population from some common ancestor of those individuals.

The above definition of a haplotype does not make it easy to determine if a segment is a haplotype because we don't know the transmition history of the chromosomes. But more constructive, operational definitions derive from

this one. For example, a haplotype can be defined as a maximal segment of one copy of a chromosome where high LD (above some statitistically significant threshold) is observed, i.e., high correlation (calculated in a set of individuals in a population) between the states observed at pairs of sites in the segment.

Similarly, a haplotype can be defined as a segment of one copy of a chromosome that is "highly similar" in a significantly large subset of individuals in a population, and dissimilar from the segments possessed by the individuals outside that subset. See Figure **??**.

Note that a key property of all of these "definitions" of a haplotype is that it is a sequence derived from only one of the two "copies" of a chromosome possesed by individuals in a population. So in any segment of the genome, each individual has two haplotypes.

Having a more biologicall grounded definition of a haplotype, we can now observe that the effect of recombination over time is to decrease the lengths of observed haplotypes in a population. Moreover, in general (with some exceptions) the rate of recombination is not influenced by the rate of selection of a trait. The segment around the site for a positively selected trait will experience recombinations at the same rate as if the trait was a neutral trait (not positively or negatively selected). Therefore, through recombination, the length of the haplotype around the causal site for a trait provides a clock measuring the time since the causal mutation occurred. Recalling that an "allele" is a technical term for a variant, the following quote summarizes this discussion:

> Positive selection is expected to more rapidly increase the frequency of an allele, and hence, the length of the haplotype (extent of DNA segment) associated with the selected allele, relative to those that are not under selection. [191]

We can now introduce the main idea behind the long-haplotype method for detecting recent positive selection:

> When we observe a haplotype that is highly frequent in the population relative to its length (meaning that it is much more frequent than is typical for a haplotype of that length), we can suspect that it contains a causal mutation for a positively selected trait.

Note that we haven't assumed that we know what the trait is, nor have we assumed that we initially knew anything about the location of the causal site for the trait.

The simplest implementation of the long-haplotype method is to empirically examine the SNP sequences of individuals in a population to identify haplotypes, and to measure their lengths and frequencies in the regions examined. The data can then be used to identify any haplotypes that are unusually long compared to

their frequencies, and well-known methods can determine the statistical significance of that deviation. Other techniques are based on theoretical derivations of the expected frequency of a haplotype as a function of time, in the absence of positive selection, and derivations of the expected length of a haplotype as a function of time.

Adding Case-Controls The power of the long-haplotype method is increased when a specific trait of interest has been identified and sampled individuals in the population have been segragated into those that have the trait of interest (the *cases*) and those that don't (the *controls*). Then, we look for a long haplotype (relative to its frequency) in the cases, that does not occur in the controls. The method is even more powerful when the location of the causal mutation for the trait is known. In that case, we look for a long haplotype (relative to its frequency) that encloses the causal site. The most powerful variant is when the causal location is known and cases and controls have been identified.

Different variants of the long-haplotype method have been used to identify hundreds of traits, and/or putative causal mutations, that are believed to have been positively selected in the recent past. In humans, some of the more notable recent traits (sweeping the population within the last several thousand years) identified in this way include the ability of adults in nothern European populations to metabalize lactose [**?**], and the ability of Tibetans to more effectively utilize the limited oxygen at hhigh elevations [**?**]. Selection for Lactase persistence is described in [149] as follows:

> The striking pattern of genomic variability that is observed in this locus involves a long, high frequency haplotype that contains an allele associated with lactase persistence. The haplotypes that carry the allele are almost identical in regions close to the location of the causative SNP, whereas haplotypes that do not carry the allele show a normal level of variability. This is exactly the pattern we would expect to observe if the allele has recently increased in frequency as the result of positive selection.

One other example of historical interest is the case of the rapid spred of the black form (appropriately called *carbonaria*) of the peppered moth in 19th-century Britian. The wild-type of this moth is lightly colored, and the spred of the black form was corrolated with the growth of dark air pollutants in industrial Britian. This change of color has been widely used as a textbook example of observed evolution, hypothesized to be a response to environmental change. However, the genetic and molecular basis of the change of color was only recently determined. In [94], the chromosomal sites responsible for the change of color were mapped to a 200-kilobase region. The black colored moths were shown to

contain a single haplotype in that region that differed from the haplotypes of the light-colored moths:

> We have genetically mapped the *carbonaria* morph to a 200-kilobase region ... and shown that there is only one core sequence variant associated with the carbonaria morph, carrying a signature of recent strong selection. [94]

Of course, the actual implementation of the long-haplotype method involves several technical issues. The most important ones are how to recognize haplotypes in practice and how to statistically define significant length and frequency associations. Another important technical issue that recombination rates can be highly variable across the genome (recombination hotspots and deserts are known), although the recombination rate is not generally related to whether the haplotype contains a positively selected trait. Since recombination breaks down the length of the haplotpye, recombination rates affect the rate at which haplotype lengths change. Well done studies must take account of what is known about the varying rates of recombination. The more that is known about the parameters of recombination in different parts of a genome, the more these variations can be normalized, reducing their effect on the long-haplotype method.

This highly simplified introduction to the long-haplotype method is intended to illustrate the central role that recombination plays in the important question of identifying traits that were positively selected, and also to motivate the general issue of understanding the patterns of recombination in a population. It is not yet clear how helpful explicit genealogical networks will be for approaches like the long-haplotype method, but it is certainly plausible that explicit deduced histories, showing temporal locations of mutations and both spacial and temporal locations of deduced recombination events, would be of value in efforts to locate recent causative mutations for positively selected traits.