5

Evolution, Trees, and Beyond

The affinities of all the beings of the same class have sometimes been represented by a great tree. I believe this simile largely speaks the truth. The green and budding twigs may represent existing species; and those produced during each former year may represent the long succession of extinct species.

Charles Darwin

Any living cell carries with it the experience of a billion years of experimentation by its ancestors.

Max Delbruck

5.1 Introduction

It is impossible not to marvel at the richness of life on Earth: from the large mammals in the sea and on the plains, to the hardy plants of the high mountains, to the microbes living in hydrothermal vents and under the Antarctic ice. The adaptability and sheer quantity of cellular life on this planet is staggering. The challenge of classifying its diversity was recognized as far back as the fourth century BC, when Aristotle (384–322 BC) introduced one of the first systematic taxonomies of living organisms. He began his work by separating the plants from animals. Then, he split the animals into those that walked, swam or flew, and the plants into those small, medium or large in size. He further subdivided these groups based on other criteria. Beyond his taxonomy, Aristotle also proposed a hierarchy of animals, known as the “Ladder of Life,” with simple organisms on its lower rungs and humans at the top.

Modern taxonomy was founded in the eighteenth century by the Swedish scientist Carolus Linnaeus, who undertook the colossal task of classifying all known animals, plants and even minerals. In *Systema Naturae* (1735), Linnaeus proposed a hierarchical structure where similar organisms were first grouped into species, similar species would be grouped into one genus, and similar genera would be grouped into one family, and so on, generating a phylogeny built of six ranks of taxa: kingdom, class, order, family, genus, species. Thus the highest rank taxa in the Linnaean taxonomy were the three kingdoms (plants, animals, and minerals). The animal kingdom, for instance, was divided into six classes: Mammalia, Aves, Amphibia, Pisces, Insecta, and Vermes. The Linnaean system is the forerunner of most modern classifications of living and extinct organisms. Each lower rank taxon belongs to a higher rank taxon, ultimately generating a tree structure (Figure 5.1). Despite the elegance of his taxonomy, Linnaeus was troubled by the possibility that a given organism could present characteristics common to several taxa of the same rank. To deal with this contingency, he annotated some *animalia paradoxa*, or contradictory animals, that resisted hierarchical classification. Amongst these were the dragon – which looks like a snake but has wings like a bird – and the legendary Borometz or Scythian Lamb, a tree that grew lambs (Figure 5.2). Based on their failure to fit within the hierarchy, it was eventually concluded that some of these animals were mythical and did not exist beyond the medieval bestiaries and the human imagination.

In 1859, Charles Darwin published *On the Origin of Species* [133], in which he introduced his landmark theory of evolution by natural selection. Evolution arises when parent organisms reproduce, generating progeny that resemble their parents but have additional variation that allows them to adapt to different environmental pressures. Some of this variation in the progeny is inheritable and is passed on to future generations. The accumulation of inherited variation over time eventually leads to the formation of new species. *On the Origin of Species* contained a single figure, depicting the ancestry of species as a phylogenetic tree (Figure 5.3). The idea of variation and its inheritance provided a beautiful explanation of the generation of species through a branching process: different organisms in the same taxa resemble each other because they share a common ancestor. Darwin revealed that a taxonomy is fundamentally a historical document – a record of the development of life on Earth.

Since then, the tree structure has become a dominant framework for representing evolutionary processes. Prior to the advent of sequencing technologies, most comparisons between organisms were based on phenotypic traits, the set of an organism's observable characteristics. The development of technologies to decode genomic material has provided a means to track the source of inherited variation and has enabled the comparison of organisms at the most fundamental level. Inferring evolutionary trees from molecular data, the practice known as molecular phylogenetics, has become a standard process in the study of evolution.

A species tree can only be inferred from genomic data if different regions of the genome provide similar trees. In humans, however, this is not the case. We

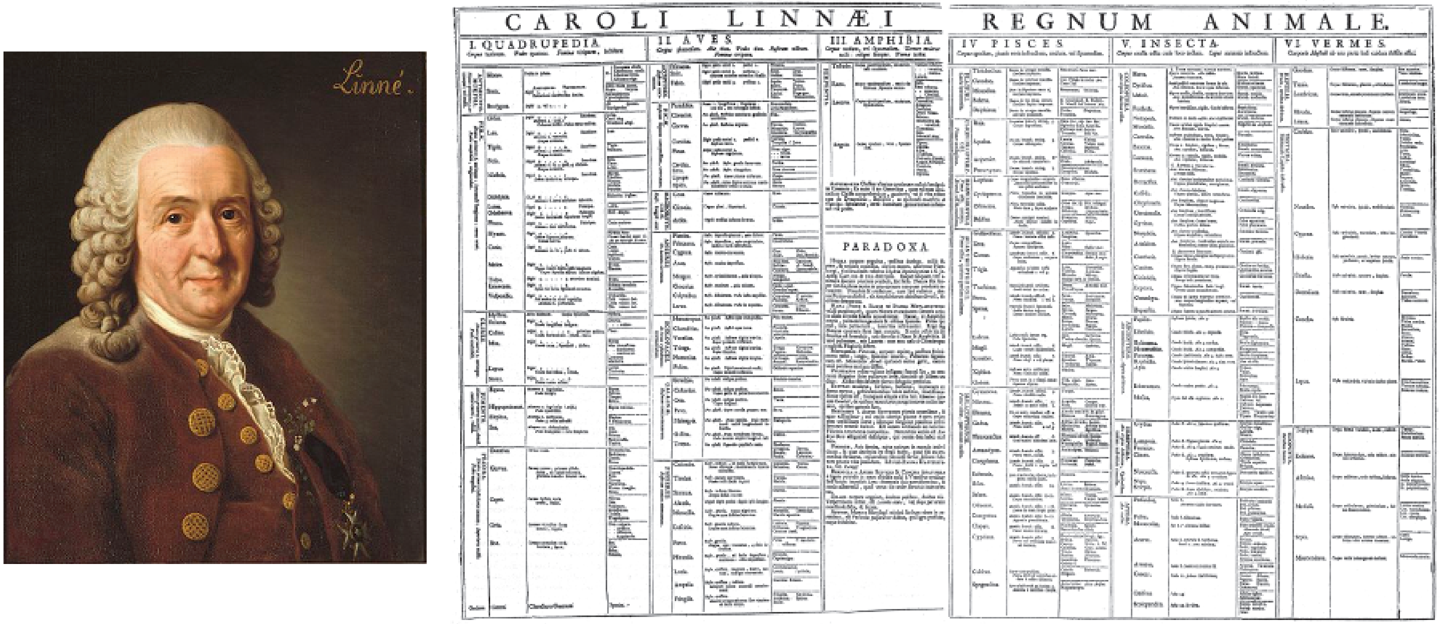


Figure 5.1 In 1735, the Swedish scientist Carolus Linnaeus published the *Systema Naturae*, a hierarchical classification of all known animals, plants, and minerals. These three groups formed his kingdoms, each of which was further divided into classes. For instance, except for a few exceptions (*animalia paradoxa*), all known animals were segregated as mammals, birds, amphibians, fish, insects or worms. Interestingly, the *animalia paradoxa* presented features from several classes and could not be unequivocally classified. Source: (1) Portrait of Carl Linnaeus, 1707–1778, Painted by Alexander Roslin in 1775, NMGrh 1053, Nationalmuseum, Stockholm, public domain. (2) Table of the Animal Kingdom (Regnum Animale) from Carolus Linnaeus's first edition (1735) of *Systema Naturae*.

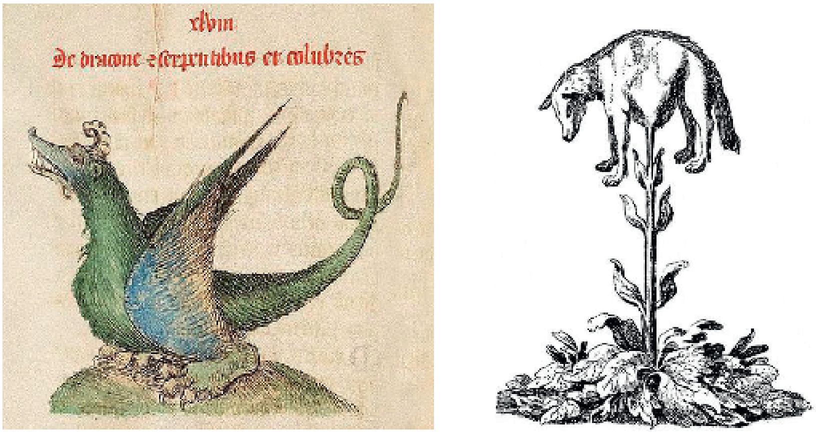


Figure 5.2 The "animalia paradoxa" were animals that challenged Linnaean taxonomy because they possessed similarities with organisms belonging to at least two different higher taxa. The dragon, for instance, had a body similar to that of a reptile but also wings like birds (illustration from the *Liber Floridus*, or *Book of Flowers*, circa 1100AD, public domain). The Borometz, or Scythian Lamb, was a plant that grew lambs. Source: Lee, H. 1887. The Vegetable Lamb of Tartary: a Curious Fable of the Cotton Plant, to Which Is Added a Sketch of the History of Cotton and the Cotton Trade. S. Low, Marston, Searle & Rivington, London.

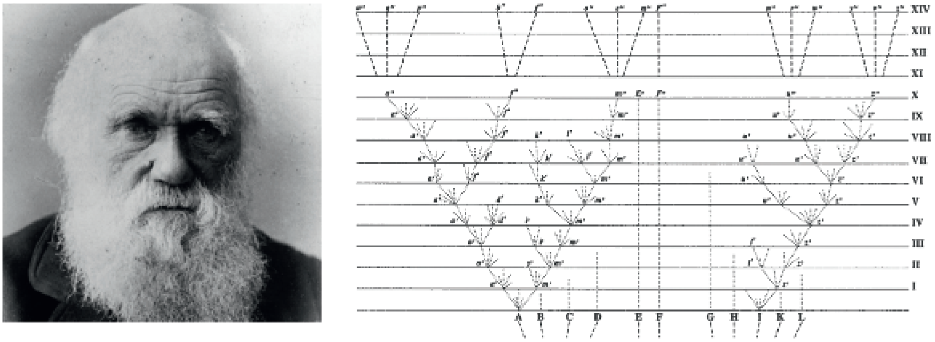


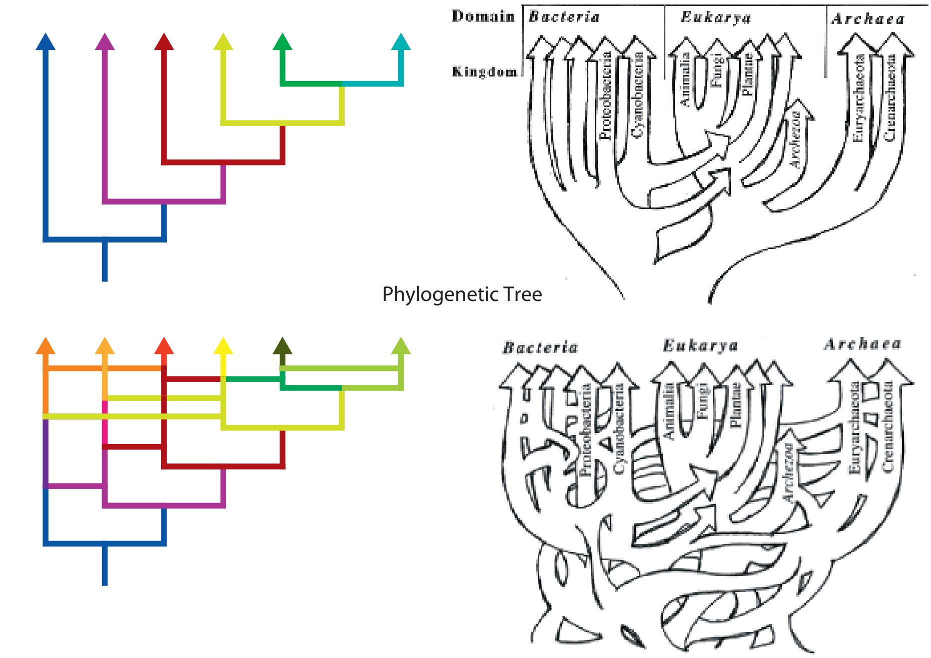
Figure 5.3 This tree appeared in Darwin's *On the Origin of Species* as a means of capturing the divergence of species. In this figure, time advances moving up the tree. The roots of the tree represent the original species that diversified according to a branching process through progeny variation and selection. The top branches constitute the modern species, and the branches that do not persist to the top represent extinct species. Source: Left: Library of Congress, Prints & Photographs Division, reproduction number, LC-DIG-ggbain-03485. Right: Darwin, C. R. 1859. *On the origin of species by means of natural selection, or the preservation of favoured races in the struggle for life*. London: John Murray.

know that some genomic material, like mitochondrial DNA, is inherited through the maternal line, while other material, like the Y chromosome in men, comes through the paternal line. Thus trees inferred from mitochondrial DNA will not agree with those inferred from the Y chromosome, as the evolutionary stories of our fathers and our mothers are different. The problem becomes more complex when different regions across chromosomes give rise to different potential trees. Genomic data has increasingly challenged the single-tree picture, as biological phenomena like species hybridization, bacterial gene transfer, and meiotic recombination have complicated the lineage of inheritance. Despite their smaller genomic size, viruses are also found to contain incompatible genomic tree histories. Frequent recombination events in HIV, for instance, have confounded attempts to reconstruct an early history of the epidemic.

In 1990, using molecular comparison, Carl Woese et al. proposed the organization of all cellular life forms into three large domains: the Bacteria, the Archaea, and the Eukarya [538]. This study showed the power of genetic information to elucidate deep phylogenetic relations that were hidden to other methods. Woese's tree, however, was based only on a small fragment of 1500 nucleotides in the 16S ribosomal RNA of prokaryotes, a tiny fraction of any organism's genome. One then wonders if the tree reconstructed from this small part of the genome can be extended to other parts of the genome, or if there exist other genes that could generate vastly different trees. Indeed, with the accumulation of genomic information, an increasingly complex picture of the relations between species is emerging, with different genes providing different incompatible tree phylogenies (see Figure 5.4), highlighting the need for new representations [147].

There are, broadly, two ways in which organisms acquire genomic material. The first, which we call here clonal evolution, is the consequence of direct transmission of genes from a single parent to the offspring. Clonal evolution is a type of vertical evolution, the direct transmission of genetic information from parents to offspring. Changes in genomic material are mediated by random mutations over multiple generations. The genomic material is inherited from a single parent, and mutations will lead to differences between a clone and its parents. This type of vertical evolution is best represented by a mathematical structure called a phylogenetic tree. The left of Figure 5.5 depicts a rooted tree where the root node at the apex represents an ancestor that propagates and diversifies over time, creating new lineages, called clades, in a branching pattern.

As has become increasingly apparent with the advent of sequencing technologies, genomic material may also be acquired through a second means: horizontal or reticulate evolutionary events. These events occur when distinct clades merge to form a new hybrid lineage. This phenomenon may be effected in a number of ways and occurs across all domains of life. This phenomenon is pervasively



Phylogenetic Tree

Phylogenetic Network

Figure 5.4 Idealized, simplistic phylogenetic trees contrast with more realistic, complex reticulate networks. On the top right is the Doolittle representation of the Tree of Life, made before the advent of sequencing technologies. It was thought that most evolution occurred through branching processes, with the notable exceptions of mitochondria and chloroplasts – believed to be symbiotic bacteria that fused part of their genome to their host's. This picture is changing as the significance of horizontal exchange of genomic information is becoming more evident. Source: [147]. From Doolittle, W. F., Phylogenetic Classification and the Universal Tree, *Science*, 1999, 284 (5423): 2124–2128. © 1999 Reprinted with permission from AAAS.

found in viruses, for instance. As we will see later in detail, viral influenza undergoes horizontal evolution through reassortment and HIV undergoes horizontal evolution through recombination. Phylogenetic trees, however, are not able to capture these horizontal evolutionary events. Representing these events graphically requires a new structure called a *reticulate network*, in which branches are allowed to both join and split. Places in the network where branches merge are known as cycles and correspond to individual reticulate events (Figure 5.5, right). The resulting network is the result of merging many different trees with different topologies.

To detect reticulate events by phylogenetic means, one must first construct a tree for each gene in the genome and then cross-reference each pair of trees for conflicts in lineal history. The simple example of the Network of Life, depicted

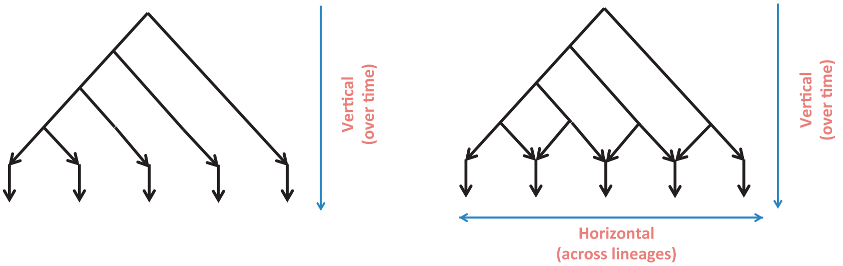


Figure 5.5 Examples of a phylogenetic tree (left) and a reticulate network (right) capturing clonal and horizontal evolution, respectively. Source: [100]. From Joseph Minhow Chan, Gunnar Carlsson, and Raúl Rabadán 'Topology of viral evolution', *Proceedings of the National Academy of Sciences* 110.46 (2013): 18566–18571. Reprinted with Permission from Proceedings of the National Academy of Sciences.

in Figure 5.4, illustrates the complexity of inferring the properties of phylogenetic networks summarizing complex data sets (from Doolittle [147]).

Some of the processes that lead to non-tree-like structures are shown in the table below.

Organism	Reticulate process	Description
Viruses	Homologous recombination	Intragenomic homologous crossover
	Reassortment	New sets of different segments in segmented viruses
Bacteria	Transformation	Acquisition of foreign DNA from environment
	Transduction	Viral-mediated exchange
	Conjugation	Exchange through cell-to-cell contact
Eukaryotes	Meiotic recombination	Crossover and gene conversion during meiosis
	Hybrid speciation	Hybridization between different species
	Endosymbiosis	Fusion of genomes of symbionts

5.2 Evolution and Topology

We now explain how to use topological data analysis to determine when evolutionary processes violate tree-like assumptions, i.e., to detect reticulate events, based on observed genomic data. First, to understand how reticulate events can be observed in genomic data, we will consider a very simple model.

Assume that we have a simplified genome with only two nucleotides, or basic informational units, 0 and 1. We will further assume that this genome is quite large

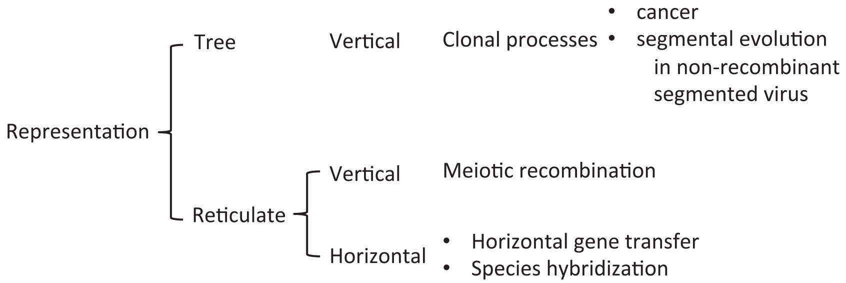


Figure 5.6 Summary of concepts used in this chapter. Clonal processes, such as tumor development or bacterial evolution without gene transfer, are well captured and represented by trees. Many processes, however, cannot be represented by a tree and require the reticulate representation. These include certain vertical processes where the genetic information is inherited from more than one parent, like in meiosis in eukaryotes. Other processes involve the transfer of genetic information between species, like in horizontal gene transfer in bacteria or species hybridizations.

and that mutations exchanging 0s and 1s occur at uniformly random positions along the genome. If the total number of bases is very large compared to the number of mutations, then, assuming that mutation sites are chosen at random, the probability that any particular site will be mutated twice is very small. In particular, as the genome length approaches infinity and the number of mutations is held constant, the probability of any site being mutated twice approaches zero. We can formalize this for genomes of finite length by imposing the constraint that any given site only mutates once; this is called the infinite-sites assumption. For this discussion, let us adopt the infinite-sites assumption and assume that an organism evolves only through a clonal process.

We now observe that certain mutational patterns are not possible given these assumptions. Suppose that we have a genome of length 2. Then we can have four possible alleles: 00, 01, 10, and 11. Consider an organism with ancestor 00. A mutation in the ancestor's first site generates 10 and a subsequent mutation in its second site generates 11. How can we generate 01 after these two mutations? The ancestral genome would have to mutate back at the first site; but this second mutation at the first site would violate the infinite-sites assumption and thus this mutational pattern would not be allowed in our model. Similarly, if the ancestor's second site mutated first, we would not be able to generate the allele 10. Therefore the presence of four alleles in the observed population is incompatible with a solely clonal evolutionary process from a single ancestor in this setup.

This observation can be turned into a test for reticulate events: checking for these four alleles is referred to as the *four gamete test*. In practice, no genome is of infinite length and so the infinite-sites hypothesis is not quite right; thus, violations of the four gamete test are possible even for strictly clonal evolution. However, if the violation is identified at multiple sites, chance becomes an unlikely explanation. For instance, in order to generate four genomes 0000, 1100, 0011, 1111, the infinite-sites model would need to be violated twice. The more violations we have, the less likely it is that our assumptions of clonal evolution are correct. So if we are confident that the infinite-sites model is a reasonable approximation, then a large number of incompatibilities casts doubt on the assumption of clonal evolution.

This raises the question of how to quantify what a "large number" of incompatibilities should be. One method is the Hudson-Kaplan test, which counts the minimum k such that there exists a partition of the data into k subsets such that within each subset all sites are compatible with the four gamete test [254]. For example, in the case of the genomes 0000, 1100, 0011, 1111, if we split the genome down the middle, and consider each half independently (00, 11, 00, 11 and 00, 00, 11, 11), then the four gamete hypothesis is no longer violated in each partition.

Besides the Hudson-Kaplan method and variations based on the four gamete test, there have been significant efforts to identify recombinants, their ancestors, and specific genomic break points, i.e., the points in the genome where recombination has occurred. Several major strategies have been developed to detect recombinants.

- Distance methods rely on differences between the genetic pairwise distances along the genome, usually using a sliding window technique. Based on some underlying model one can then evaluate the likelihood of a recombination [530].
- Phylogenetic methods are based on the idea that if a recombination has occurred, the trees inferred from different parts of the genome could have distinct topologies [432]. The same type of techniques can be used to identify genes that have been transferred when orthologous genes from different species are more similar to each other than expected, given the species' evolutionary relationship [27].
- Compatibility methods search for phylogenetic incongruence in a site-by-site basis, and, in general, do not require the phylogeny of the sequences to be known [414, 470].
- Substitution methods search for clustering of substitutions along the genome using some summary statistics in different phylogenetic partitions [414, 485].
- Linkage disequilibrium methods are one of the most popular techniques for studying large genomes, e.g., the human genome. The main idea in such methods is that if in a section of a genome there has not been recombination, the presence of a substitution is informative (linked) to nearby regions. Recombination breaks this linkage. We will describe in Section 5.8 how linkage is used to study recombination in the study of large numbers of human genomes.

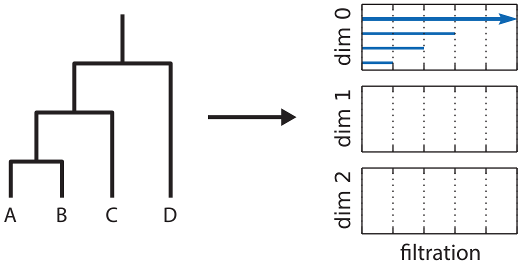


Figure 5.7 Tree topologies are contractible. When computing persistent homology, one can observe that there are no bars in barcodes of dimension bigger than zero. Source: [162]. Reprinted with permission: © EAI European Alliance for Innovation 2016.

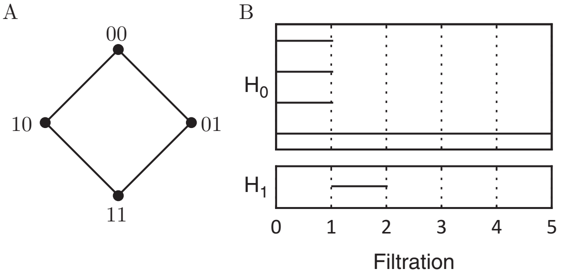


Figure 5.8 A simple reticulation event involving four genomes with only two sites and two bases 0 and 1. (A) If the four possible states 00, 01, 10 and 11 are present (four gamete test) one can suspect that a site has mutated twice or there has been a recombination between these sites. In the case of large genomes where mutations in the same site are considered highly improbable (infinite site models), the four gamete test is used in many statistical tests for the identification of recombination events and specific recombination sites. (B) When we apply persistent homology to the Hamming distance between these different small genomes, one clearly identifies an interval $[1, 2)$ in the first homology persistent diagram. The non-trivial homology classes in dimension one and higher can be used as indicators of the presence of recombination or multiple mutations in the same site. Source: [162]. Reprinted with permission: © EAI European Alliance for Innovation 2016.

A summary of these methods and accompanying software can be found at the end of the chapter in Section 5.12.

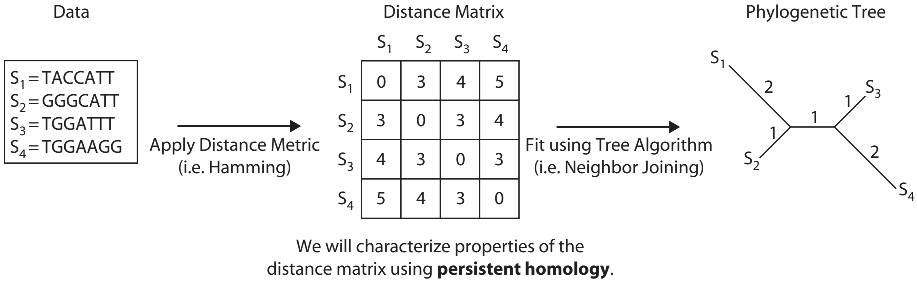
The methods used to identify recombinant sequences can suffer from prohibitive computational costs in large data sets. These methods are designed for the specific task of identifying recombinants and breakpoints. They use quantified measures of the violation of the tree assumption to infer these events. A natural question that then arises is whether there are better descriptors of the data in the event of recombination. The first hint that topological data analysis might be useful comes from the observation that trees are contractible (see Figure 5.7).

The four genomes $\{00, 01, 10, 11\}$ can be considered the fundamental and simplest model of recombination. Topologically the set forms a loop, as shown in Figure 5.8, i.e., the Vietoris-Rips complex (recall Definition 2.1.6) contains a complex with a non-trivial loop. Four of the six pairs of gametes are separated by a Hamming distance of 1, while the other two are separated by a distance of 2. (Here recall from Example 1.2.5 that the Hamming distance counts the number of positions at which the strings are different.) At a filtration distance of 1, the four pairs become connected, forming a loop. At a filtration distance of 2, the remaining two pairs become connected, destroying the loop. Thus, we have non-vanishing H_1 homology on the filtration interval $[1, 2)$. This simple example suggests that persistent homology provides a method for counting the number of incompatibilities and, at the same time, determining the scale of each incompatibility in terms of the distance between the alleles.

Each interval in the barcode can be interpreted as a sign of a recombination event involving a set of sequences including the common ancestor, parental, and recombinant strains. The interpretation and identification of the recombinant and parental strains could be complicated or impossible unless given further information. This is analogous to the problem of finding a root of a phylogenetic tree if no information about ancestral states is provided. Persistent homology can provide a simple way to estimate the number of incompatibilities.

For our purposes, we can assume the genomes of organisms in an evolving population forms a metric space (X, ∂_X), which we never directly observe. Instead, we observe a sample of data points (i.e., sequenced genomes of cells) that lie in X. Restricting the metric ∂_X on X to our sample gives us a distance between points. Considering genomes as a string of characters makes it easy to define distances, e.g., the Hamming distance. Metrics currently used in biological applications are based on different models of how mutations can occur. For instance, one can modify Hamming distances to account for the possibility of back mutations after some time (Jukes-Cantor distances [279]), account for the fact that different substitutions can occur with different probability (Kimura models [298]), allow different bases to occur at different frequencies [173, 231], among many possible refinements. Thus, we have a finite metric space generated by genomic sequences separated from each other by some genetic distance, to which we can apply the techniques of topological data analysis.

Recall from Section 4.7.1 that some finite metric spaces can be derived from a weighted tree, with the distances between two leaves calculated by adding up the weights associated to the edges connecting them (see Figure 5.9). Tree-like spaces can be used to represent clonal processes, with internal nodes representing unsampled ancestors.



Data

S1=TACCATT
S2=GGGCATT
S3=TGGATTT
S4=TGGAAGG

Distance Matrix

	S1	S2	S3	S4
S1	0	3	4	5
S2	3	0	3	4
S3	4	3	0	3
S4	5	4	3	0

We will characterize properties of the distance matrix using **persistent homology**.

Figure 5.9 Pipeline for analyzing genomic data using persistent homology. Starting from a sample of sequences, one can compute distances reflecting the similarity between organisms. These distances provide a finite metric space, that, in some cases, can be summarized by a phylogenetic tree whose leaves correspond to points in the metric space. Distances between branches can be estimated by the addition of weights along the shortest path.

Obviously, not every metric space has this tree-like metric property. In general, finite metric spaces that can be represented by weighted trees are only a small subspace of all finite metric spaces. Indeed, Lemma 4.7.2 described the required four point condition satisfied by metric spaces generated by trees.

But when there is no underlying tree explaining the data, we can capture and represent evolutionary processes beyond trees using topology. A phylogenetic tree can be continuously deformed into a single point. The same action cannot be performed for a reticulate network without destroying the loops or cycles in the structure. The active hypothesis then is that the presence of these holes results directly from horizontal evolutionary events. This idea can be formalized into the following theorem [100].

Theorem 5.2.1. *Let (M, ∂_M) be any tree-like finite metric space, i.e., a space satisfying the four point condition, and let $\epsilon \geq 0$. Then the Vietoris-Rips complex $\mathrm{VR}_\epsilon(M, \partial_M)$ is a disjoint union of acyclic complexes. In particular, $H_i(\mathrm{VR}_\epsilon(M, \partial_M)) = \{0\}$ for $i \geq 1$.*

In other words, the presence of homology above dimension zero indicates that the metric space does not satisfy tree-like metric properties. Identifying the genomes that are the generators of these homology classes selects subsets of genomes whose derived distances do not satisfy the tree condition, indicating that non-tree-like evolutionary processes have occurred within these subsets (Figure 5.10).

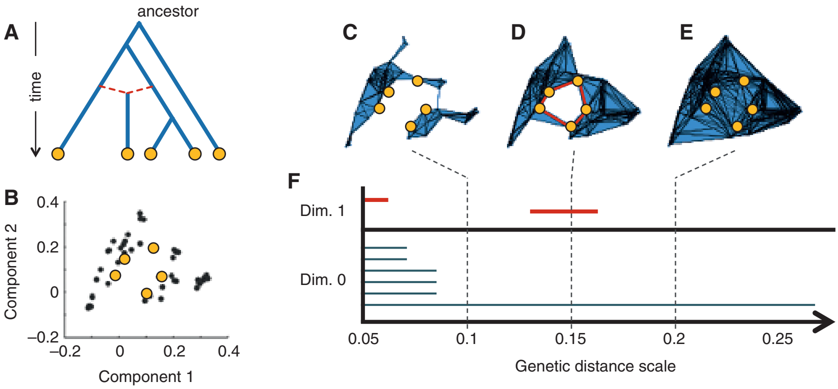


Figure 5.10 Persistent homology detects historical recombination events from population genetic data. Consider the reticulate phylogeny in panel A. Five genetic sequences sampled today (the yellow circles) developed from a single common ancestor through clonal evolution (solid blue lines) and recombinant evolution (dotted red lines). Panel B illustrates this sample within a larger sample of the population. Persistent homology is applied to this larger sample and three filtrations are shown in panels C, D and E. Panel F shows the resulting barcode. Note that these two dimensional plots (panels C, D, E), created by principal component projection, are used merely to visualize the sequences; projection is not part of the algorithm. The dimension 1 bar near the center of panel F identifies a recombination event involving the five highlighted sequences. The scale over which this bar persists captures the genetic difference between the parents of the recombinant [323]. Source: [100].

The persistent homology approach suggests a general strategy to study the space of genomes. Instead of considering trees and reticulate networks in the phylogenetic sense, we consider these structures in the context of simplicial complexes and compute their persistent homology. An additive tree is a single connected component without any loops which displays only zero dimensional topology. Reticulate structures, on the other hand, contain loops and therefore may contain non-trivial higher dimensional topology.

Recall from Section 2.3 that persistent homology can be displayed in a barcode plot where for a given filtration and dimension k, different bars represent independent k-dimensional cycles that generate non-trivial homology classes. As we have observed, the presence of non-zero homology above dimension zero indicates deviation from a tree metric. The next step is to define a quantity that captures the extent of deviation from a tree. In order to do this, we consider the distribution B_k of bar lengths of k-dimensional cycles for some $k > 0$.

Specifically, a natural measure of the deviation from a tree metric is some kind of count of the number of bars in PH_1. We define the *topological obstruction to phylogeny* (TOP) to be the L^∞ norm, or maximum, of the lengths of the bars. The work of [100] established that a filtration with non-zero TOP implies that the finite metric space is not tree-like. Another possible measure is the L^1 norm, which is equivalent to the sum of the bar lengths. In simulations of evolutionary data where we initially set a rate r of horizontal evolution, we find that of all L^p norms, the L^1 norm best correlates with r. Finally, we could also consider the L^0 norm, simply the count of the number of bars, which is also proportional to r. To approximate r, we consider either the L^1 or L^0 norm normalized by time; we define the irreducible cycle rate (ICR) to be precisely this normalization.

As we will see at the end of this chapter, the relationship between the recombination rate and the persistent homology of a sample of genetic sequences can be probed using coalescent simulations of evolution. Figure 5.11 shows how the number of persistent dimension 1 cycles, b_1, grows with the number of recombination events that occur in a simulation.

In [164], it was demonstrated how b_1, together with the birth and death scales of each cycle, can be used to estimate the population-scaled mutation rate ρ. The accuracy and precision of this estimator increases with sample size; we discuss this in Section 5.7.3.

These results suggest a map between algebraic topology invariants, such as Betti numbers and generators of homology classes, and different types of genomic exchange events (Figure 5.12). Persistent homology provides information about the obstructions to tree-like metrics due to homoplasies (shared mutations in different genomes that are not shared by their common ancestors), recombination, reassortment, or other modes of horizontal exchange of genomic material. By studying the cycles that generate higher dimensional classes (the witnesses to the violation of the tree-like assumption), we can infer what type of biological process occurred that violated the tree-like assumption. Later in this chapter, we will

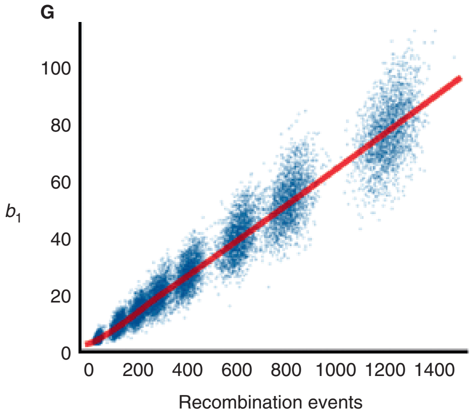
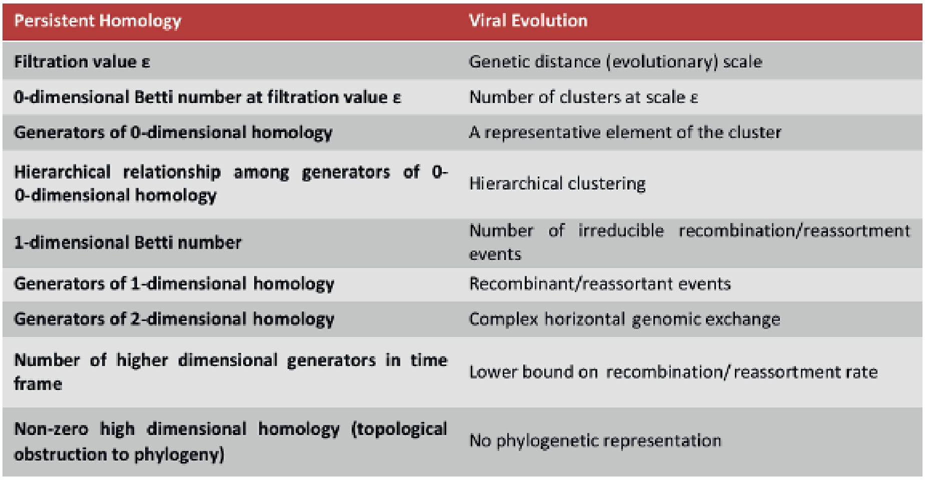


Figure 5.11 The number of one dimensional persistent homology classes, b_1, scales with number of recombination events in a coalescent simulation.



Persistent Homology	Viral Evolution
Filtration value ε	Genetic distance (evolutionary) scale
0-dimensional Betti number at filtration value ε	Number of clusters at scale ε
Generators of 0-dimensional homology	A representative element of the cluster
Hierarchical relationship among generators of 0- 0-dimensional homology	Hierarchical clustering
1-dimensional Betti number	Number of irreducible recombination/reassortment events
Generators of 1-dimensional homology	Recombinant/reassortant events
Generators of 2-dimensional homology	Complex horizontal genomic exchange
Number of higher dimensional generators in time frame	Lower bound on recombination/ reassortment rate
Non-zero high dimensional homology (topological obstruction to phylogeny)	No phylogenetic representation

Figure 5.12 Rough dictionary between TDA notions and evolutionary concepts. Source: [100]. From Joseph Minhow Chan, Gunnar Carlsson, and Raúl Rabadán, ‘Topology of viral evolution’, *Proceedings of the National Academy of Sciences* 110.46 (2013): 18566–18571. Reprinted with Permission from Proceedings of the National Academy of Sciences.

explore this relationship in detail through a series of examples in the viral, bacterial, and eukaryotic worlds.

Remark 5.2.2. In some cases the map between homological invariants and evolutionary phenomena can be made more explicit [323]. This is the case for “galled trees,” directed acyclic graphs that differ from trees by a few isolated recombinations. In that case, the homology in dimensions bigger than one vanishes, generalizing Theorem 5.2.2. These “galled trees” can be constructed by pasting tree-like and isolated recombination events that correspond to operations in the associated finite metric spaces [323].

5.3 Viral Evolution: Influenza A

5.3.1 Influenza A

Influenza A is a segmented single-stranded RNA orthomyxovirus that infects different hosts of many species. Indeed, the highest genetic diversity of these viruses is found in birds, mostly waterfowl, of the order of Anseriformes (ducks, swans and geese), Passeriformes, and Charadriiformes (including gulls). Waterfowl are the virus’s natural reservoir, perpetuating the vast biodiversity of influenza, including all different subtypes. But influenza A has also been found in pigs, seals, and other mammals including, of course, humans. Classification of influenza viruses is traditionally made by the antigenic properties of the proteins displayed in the viral envelope hemagglutinin (HA), ranging from H1 to H16, and neuraminidase (NA), ranging from N1 to N9. Recently, new types of influenza viruses have been identified in bats in Central America [500], leading to two new hemagglutinin types (H17 and H18) and two new neuraminidase types (N10, N11). Of course, it is possible that, as surveillance programs get more extensive, new related viruses will be found in other hosts (Figure 5.13).

Infection in humans and other mammals usually occurs in the upper respiratory tract, lasts a couple of weeks, and is associated with symptoms that vary from fever, sore throat and other cold-like symptoms to more serious complications that can result in death. It has been estimated that near half a million deaths are associated to influenza infections every year around the world. Transmission of human influenza occurs mostly through the air, in the form of droplets of water released from coughs or sneezes, and through fomites, surfaces that carry infectious particles. These modes of transmission seem to be more effective at low temperatures and low humidity, factors that are probably relevant for the seasonal pattern observed in human influenza. Illness associated with influenza infection is most common in winter – from November to April in the Northern Hemisphere, and from May to October in the Southern Hemisphere.

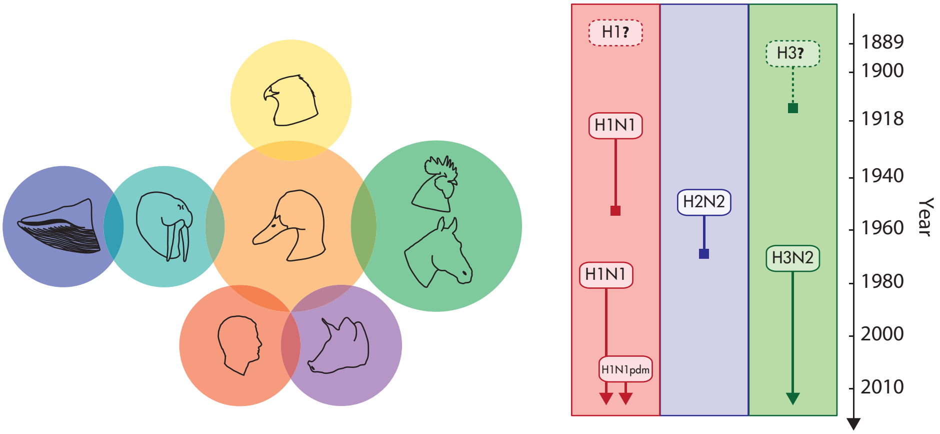


Figure 5.13 Left: Influenza A infects many different species, mostly birds and mammals. The greatest diversity of the virus can be found in waterfowl. Occasionally, viruses can jump species and infect other hosts. Influenza A has been reported in humans, swine, horses, seals, camels, bats and even whales. Right: Twentieth century influenza pandemics. Pandemics are caused by viruses containing genes from other species. Although there is some speculation about pandemics in the nineteenth century, the first well-characterized influenza pandemic was the so-called Spanish flu in 1918. Since then influenza pandemics have occurred every 30 years, with the last pandemic originating in swine in 2009. The Influenza A virus infects different species and generates pandemics.

In contrast to mammals, most birds show no clinical signs of infection by the virus, which replicates in their gut and sheds into the water through feces [529]. However, mutations occasionally occur that increase its pathogenicity, resulting in a highly pathogenic avian influenza (HPAI), which causes a multi-organ systemic disease that can kill birds. Large surveillance programs are dedicated to detecting HPAI outbreaks; HPAI transmission to humans is a chief concern.

What factors allow a virus to be transmitted between individuals or species? In both humans and waterfowl, the virus must recognize specific molecules on the surface of the cell in order to fuse with it and infect it. These molecules vary in different cells and hosts; however, the recognition of monosaccharide residues on epithelial cells by viral hemagglutinin is a common pathway. Avian influenza interacts with an α-2,3-sialic acid, prevalent in the intestinal tract of birds. In contrast, human influenza binds the α-2,6-sialic acid predominant in the human upper respiratory tract, begetting the flu-like symptoms of cough, sore throat, and rhinorrhea. Pig trachea contains both types of sialic acids. This unique feature of swine supports the mixing vessel theory that pigs provide a bridge for influenza from avian host to human, allowing the virus to adapt to recognize α-2,6-sialic acid through reassortment [269]. Host switching from waterfowl to human, however, does not require a swine intermediary. The HPAI H5N1 virus, for instance, infected 18 people and killed six in 1997 after first appearing in Guangdong in 1996 then spreading rapidly to poultry in Hong Kong. That year, Hong Kong culled more than 1 million poultry.

Since 2003, a number of sporadic H5N1 outbreaks with suspected poultry intermediaries have taken place among humans and other mammals, causing 860 human infections and 454 deaths as of February 2019 – a staggering mortality rate of nearly 60%. The fulminant progression of H5N1 infection most likely results from its specificity for α-2,3-sialic acids, which are present at a low concentrations in the human lower respiratory tract. Infection in the lower respiratory tract leads to the more flagrant symptoms of viral pneumonia. As such, avian H5N1 demonstrates high pathogenicity and productive infectivity in humans, but an inability for sustained transmission between humans. Given the high mortality rate of infection, it is a matter of the utmost importance to determine whether these viruses could become transmissible among humans like seasonal influenzas.

Recently, in the laboratory setting, teams led by Kawaoka [266] and Fouchier [240] demonstrated the pandemic potential of non-seasonal strains. They engineered H5N1 by mutating specific sites (site-directed mutagenesis) and passing the virus along ferrets (which share similar sialic acid distributions to humans) until they generated strains capable of transmission. Similarly, Zhu et al. showed that the 2013 H7N9 strain, which infected 131 humans and caused 32 deaths in two months in the Jiangsu province of China [191], infected and was transmitted between ferrets, suggesting that human to human transmission of H7N9 has most likely already occurred [548]. These outbreaks underscore the need for further investigation into the mechanisms of viral evolution and the adaptation of animal viruses to humans.

Influenza viruses are enveloped and nearly 100 nm in diameter. Their genome is 13,000 bases long and is composed of eight segments of single-stranded antisense RNA (Figure 5.14). Each segment encodes one or two viral genes. Antisense RNA is the complement of the RNA that codes for proteins; thus it cannot be directly translated into functional protein. In order for the influenza genome to express protein, positive-sense strands must be produced from the template of the antisense strands. Further complexity arises when the virus attempts to make new *virions*, the infectious particles that allow the virus to be transmitted outside of the host cell. The replicating virus must duplicate its original antisense RNA and, in order to do so, it must polymerize new strands of ribonucleotides complementary to the template of the positive-sense RNA. Influenza carries its own polymerase complex, which it uses for all of its RNA replication; in fact, the three longest genes of influenza (PB2, PB1, PA) code for the three proteins directly involved in replicating genomic material. The polymerase complex interacts directly with viral RNA and the nucleoproteins (NPs) that attach to it. An RNA segment, together with a copy

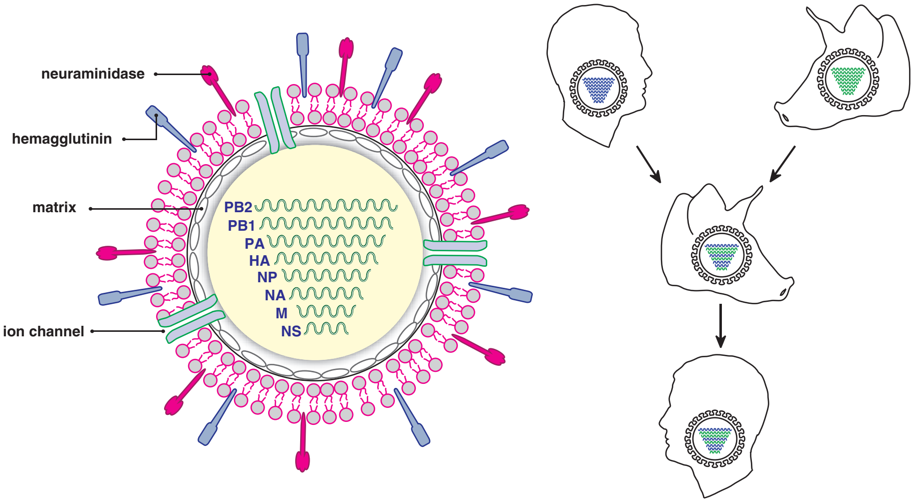


Figure 5.14 Influenza A is an antisense single-stranded RNA virus whose genome is composed of eight different segments containing one or two genes per segment. This virus contains an envelope borrowed from the infected cell that expressed two viral proteins, hemagglutinin and neuraminidase. When circulating viruses co-infect the same cell, new viruses can be created that contain segments from both parents. This phenomenon, called reassortment, can lead to dramatic adaptations to novel environments, and it is thought to be one of the contributing factors to human influenza pandemics.

of the polymerase complex and several NP proteins, forms the ribonucleoprotein (RNP) particle that is released into the cell cytoplasm and packaged in the virion or viral particle. The virion consists of a shell (capsid) of membrane proteins (MPs) and M2 proteins that form tetramers with ion channel activity. These ion channels help modulate pH within the virions and regulate the release of viral RNA into infected cells. Two other proteins, non-structural N1 and NS2, are found in infected cells but are absent, or have low expression, in virions. The existence of other proteins in alternative reading frames has been proposed; however, these proteins do not have a well-characterized role in the life cycle of the virus [507, 547].

Influenza evolves by accumulating mutations at a high rate. Estimates of evolutionary rates, or changes per unit time, indicate that influenza, like many other RNA viruses, evolves at a rate of $\sim 10^{-3}$ per nucleotide per year. This brisk evolutionary rate poses a significant challenge in the development of vaccines. Current vaccines for influenza rely on leveraging the antigenic response to epitopes (the sections of proteins recognized by antibodies) in hemagglutinin. However, these epitopes change as the virus accumulates mutations. The World Health Organization updates the composition of the vaccine with the hope that the updated vaccine will more faithfully resemble circulating strains. To help the WHO, national and international organizations put significant effort into collecting genomic and antigenic data from circulating strains of influenza. These large collections – more than 100,000 genomes, currently – constitute excellent material on which to test the mathematical and computational methods described in this book.

Substitutions (i.e., point mutations) accrued by influenza can be viewed as small changes in the nearly continuous evolution of its genome. However, point mutations are not the only way that influenza evolves; more dramatic change can occur. As discussed, influenza genomes consist of eight different segments. These segments are the viral analogue of chromosomes. When two viruses of different strains co-infect the same host cell, they can generate a progeny containing novel combinations of segments taken from both parental strains [416, 417]. This phenomenon, called reassortment, shuffles the genomic material of different strains and constitutes the underlying mechanism behind influenza pandemics.

A pandemic influenza is a viral strain that was initially endemic to animal hosts like waterfowl and swine that obtained the requisite mutations to infect and adapt to human hosts, thereby spreading on a global scale. Mutations necessary for human adaptation can be easily acquired by incorporating segments from viruses already adapted to human hosts through reassortment. Mutations and reassortments can introduce changes in the antigenic properties of the strain, which, in turn, can render antibodies raised against previously circulating viruses ineffective. The mutational change of seasonal influenza, referred to as *antigenic drift*, contrasts with the more dramatic reassortment in pandemic strains that creates entirely new viral genomes, referred to as *antigenic shift*.

In modern history, the most calamitous example of an influenza pandemic was the H1N1 (Spanish flu) epidemic of 1918. H1N1 claimed the lives of 50 to 100 million people worldwide [276]. As it disseminated throughout post-war Europe, it justified drastic public health measures including the widespread shuttering of theaters, schools and churches. A physician working at Camp Devens, a military base west of Boston, related the dramatic effects of the pandemic strain to a friend in a letter on September 29th, 1918 [210]:

> *This epidemic started about four weeks ago, and has developed so rapidly that the camp is demoralized and all ordinary work is held up till it has passed. . . These men start with what appears to be an attack of la grippe or influenza, and when brought to the hospital they very rapidly develop the most vicious type of pneumonia that has ever been seen. Two hours after admission they have the mahogany spots over the cheek bones, and a few hours later you can begin to see the cyanosis extending from their ears and spreading all over the face, until it is hard to distinguish the coloured men from the white. It is only a matter of a few hours then until death comes, and it is simply a struggle for air until they suffocate. It is horrible. . . We have been averaging about 100 deaths per day, and still keeping it up. . . We have lost an outrageous number of nurses and doctors . . . It takes special trains to carry away the dead. For several days there were no coffins and the bodies piled up something fierce, we used to go down to the morgue (which is just back of my ward) and look at the boys laid out in long rows. It beats any sight they ever had in France after a battle. Good-by old Pal, God be with you till we meet again.*

The genome and the virus itself were isolated from bodies buried in a mass grave in the permafrost of a remote Inuit village in Brevig Mission (called Teller Mission in 1918) on the Seward Peninsula of Alaska [423]. 85% of the adults that were buried in the mass grave died within the span of five days in November, 1918. In 1997, several of the bodies were exhumed. The viral sequence of this strain was recovered and can be found online under the name A/Brevig Mission/1/18 (H1N1). Despite knowledge of the sequence, many questions about the 1918 pandemic strain remain:

What was its original host?

Where and when did it first infect humans?

And why was it so pathogenic?

After a couple of waves of worldwide infection, the pandemic-causing strain became a seasonal influenza virus.

The next human pandemic, the so-called Asian flu, occurred in 1957. A descendant of the 1918 H1N1 pandemic strain, still circulating in humans, acquired three segments of avian origin (PB1, HA, NA), forming the H2N2 strain and causing a pandemic (Figure 5.13). The H2N2 virus circulated in humans, replacing the H1N1 virus, until the next pandemic. In 1968, a new reassortant, H3N2, which contained H2N2 and avian segments (PB1, HA), was identified in South Asia and rapidly spread across the world. H3N2 still circulates in humans today and is a major cause of morbidity associated with influenza. Interestingly, H1N1, which had not been found circulating in humans since the pandemic of 1957, reemerged in 1977 and co-circulated with H3N2.

In 2009, a swine-origin novel H1N1 virus marked the first pandemic of the twenty-first century (Figure 5.15). In mid March 2009, reports came from Mexico regarding an outbreak of respiratory illness. In April, two cases were documented in the United States in children from Southern California [200]. The CDC was alerted to the first case on April 13th: a ten-year-old boy who lived in San Diego County. The patient had fallen ill with fever, cough and vomiting on March 30th. None of his family members shared his symptoms. In the other case, a nine-year-old girl developed a respiratory illness in Imperial County. The CDC identified a new strain of influenza related to viruses circulating in swine and characterized and published its genome. Since neither of the children had been in contact with pigs or each other – they lived 130 miles apart – the CDC suggested that the virus was already circulating in humans. A few days later, cases emerged in Texas. Within the following month, infection had struck every continent. The World Health Organization declared the strain a pandemic on June 11th, 2009.

The 2009 pandemic resulted from a reassortment between different influenza viruses circulating in swine [474, 506]. The pandemic virus showed relation to viruses isolated in swine more than a decade ago in North America and Asia. It is still unclear how, where and when these viruses developed into a human pandemic, and where the virus was circulating in the year before the pandemic. The most widely accepted conjecture is the *hidden pig herd hypothesis*, which proposes that incomplete surveillance missed strains in untested swine herds, and recent reports suggest that these viruses circulated in pigs in Mexico [241, 348].

The recent ancestors of a pandemic virus provide invaluable information about the set of minimal genomic alterations that can transform a zoonotic agent into a human pandemic. Understanding the origins of infectious strains can help us define scientifically based rules for the risk assessment of new strains and for the implementation of public health measures that might help avoid or mitigate future pandemics.

5.3.2 Reassortments in Influenza through TDA

We have seen that dramatic changes in the genetic makeup of an influenza virus can occur through reassortments, i.e., when two or more diverse viruses co-infect the same cell and create new viruses containing genomic material from the parental strains. Figure 5.16 shows three parental viruses with genomes comprising three

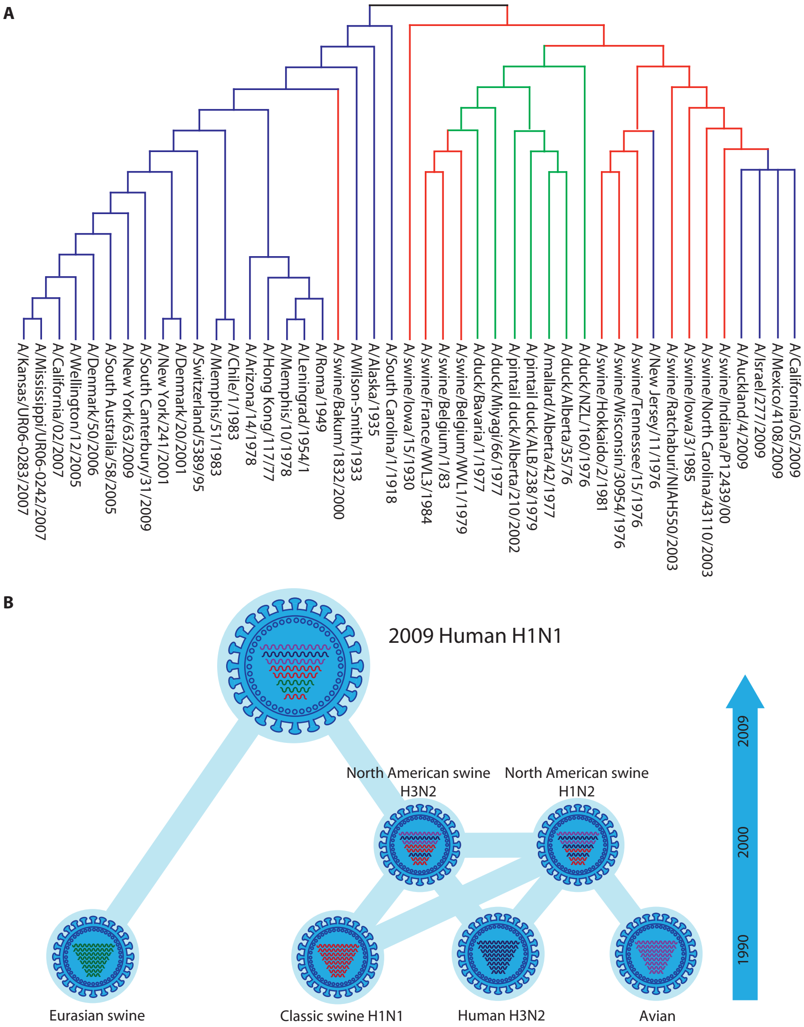


Figure 5.15 Origins of H1N1 2009 pandemic virus. Using phylogenetic trees, the history of the HA gene of the 2009 H1N1 pandemic virus was reconstructed. It was related to viruses that circulated in pigs potentially since the 1918 H1N1 pandemic. These viruses had diverged since that date into various independent strains, infecting humans and swine. Major reassortments between strains led to new sets of segments from different sources. In 1998, triple reassortant viruses were found infecting pigs in North America. These triple reassortant viruses contained segments that were circulating in swine, humans and birds. Further reassortment of these viruses with other swine viruses created the ancestors of this pandemic. Until this day, it is unclear how, where or when these reassortments happened. Source: [506]. From *New England Journal of Medicine*, Vladimir Trifonov, Hossein Khiabanian, and Raúl Rabadán, Geographic dependence, surveillance, and origins of the 2009 influenza A (H1N1) virus, 361.2, 115–119. © 2009 Massachusetts Medical Society. Reprinted with permission from Massachusetts Medical Society.

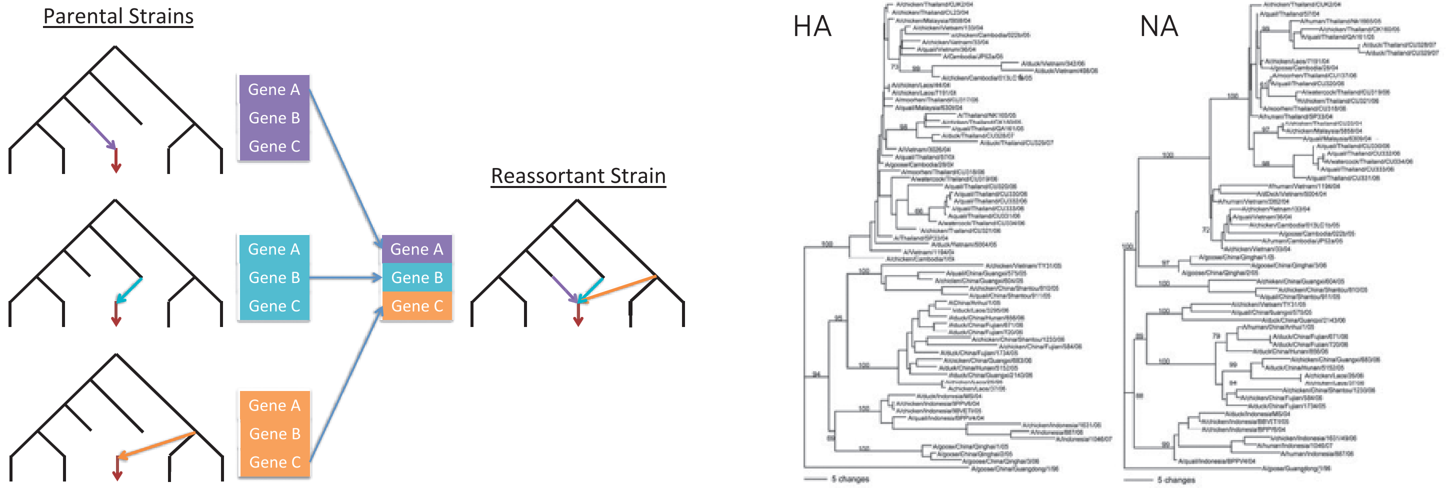


Figure 5.16 Left: Reassortments in viruses lead to incompatibility between trees. Reticulate network representing the reassortment of three parental strains. The reticulate network results from merging the three parental phylogenetic trees. Source: [100]. Right: Indeed, incompatibility between tree topologies inferred from different genes is a criterion used for the identification of events of genomic material exchange. Here we represent two genes of influenza A virus with different topologies using phylogenetic networks. From Joseph Minhow Chan, Gunnar Carlsson, and Raúl Rabadán, 'Topology of viral evolution', *Proceedings of the National Academy of Sciences* 110.46 (2013): 18566–18571. Reprinted with Permission from Proceedings of the National Academy of Sciences.

different genes and unique phylogenetic histories. All three can undergo a reassortment in which each parent donates a different gene. No single tree can capture the whole history. As such, incompatibilities between tree topologies derived from different genes may provide evidence of reassortment.

There are many interesting questions pertaining to reassortment. Imagine two different viruses infecting a cell. In principle, if each virus has eight segments, one could generate 2^8 different segmental combinations. Are these combinations all realized in nature? Is there any preference for certain combinations? Several reports have suggested that reassortments do not occur at random, but demonstrate clear preferences [206, 292, 416]. These apparent preferences may have multiple causes. The process of generating new viruses involves the packaging of eight different segments into the same virion and, although the packaging process is not completely understood, it is possible that different segments physically interact [385]. Cosegregation could also be due to selection. Given that different segments code for different proteins that work in conjunction, it is conceivable that two proteins that are co-adapted to work together will lead to offspring with higher fitness. Knowledge of these patterns may help reduce the number of potential viruses that we must consider in future pandemics.

We can study reassortments using the persistent homology framework described previously in this chapter. Let us start with a single segment: hemagglutinin [100]. To leverage persistent homology, we align our sequences, compute pairwise distances between them, and generate a finite metric space with points representing different sequences. The distance metric captures the genetic diversity present in the collection of sequences. We observe that most of the information in this metric space is contained in its zero dimensional homology with a few short bars in dimension one (see Figure 5.18 below). At this point, we can infer that a tree is a good representation of the evolution of one segment. The zero dimensional homology provides useful information about the clustering structure of different isolates. Looking at the generators of the zero dimensional classes, we can reconstruct a hierarchical clustering structure that resembles a phylogenetic tree. For example, when studying different subtypes of influenza A circulating in aquatic birds, we clearly see that the hierarchical structure derived from the zero dimensional homology correctly captures the splits between major subtypes. This phylogenetic information can be obtained easily by classical techniques that do not use persistent homology (Figure 5.17). Similarly, with our HA data, the sequences that generate zero dimensional homology can be assembled into a tree that closely resembles the unrooted phylogenetic tree created on the viral subtypes. This same analysis can be repeated for each of the eight segments of influenza (Figure 5.18). In each case, we do not recover large bars in the barcode diagram for non-zero dimensions. The few small bars at dimension one are associated with homoplasies. In cases of vanishing

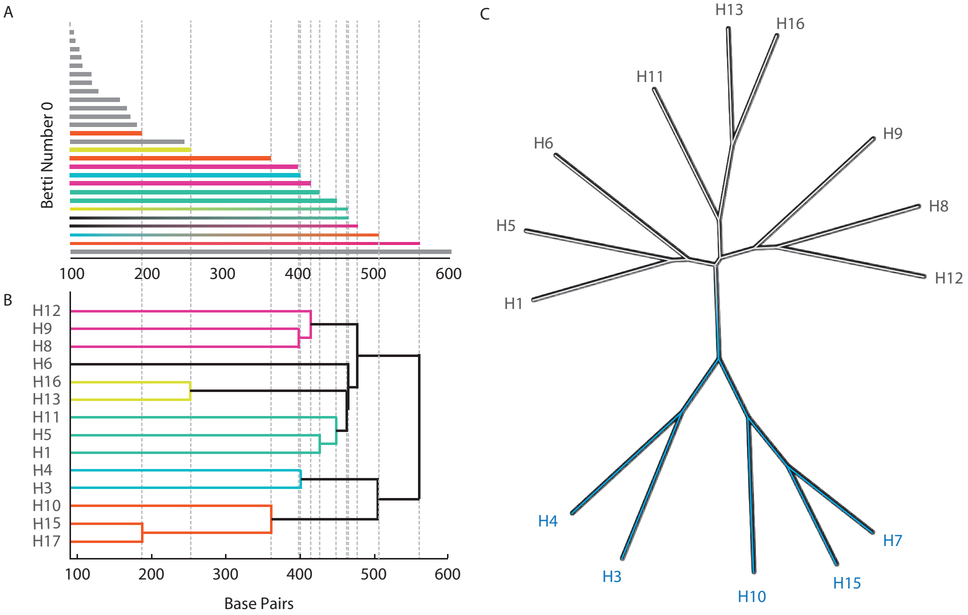


Figure 5.17 In case of vanishing higher dimensional homology, zero dimensional homology generates trees. When applied to only one gene of influenza A, in this case hemagglutinin, the only significant homology occurs in dimension zero (panel A). The barcode represents a summary of a clustering procedure (panel B), that recapitulates the known phylogenetic relation between different hemagglutinin types (panel C). Source: [100]. From Joseph Minhow Chan, Gunnar Carlsson, and Raúl Rabadán, 'Topology of viral evolution', *Proceedings of the National Academy of Sciences* 110.46 (2013): 18566–18571. Reprinted with Permission from Proceedings of the National Academy of Sciences.

higher homology, the zero dimensional homology closely follows the traditional tree structure.

However, when studying the persistent homology for several genes at the same time, large numbers of homology classes appear at dimensions one and higher, indicating pervasive reassortments. By looking in detail at the cycles in higher dimensional homologies, we can attribute these cycles to different biological processes that violate tree-like assumptions: homoplasies, recombinations or reassortments. If several sequences generate a large non-trivial class, a reassortment event likely took place among the ancestors of these isolates [100]. We can generate useful statistics based on barcode information; for instance, we can estimate how often different combinations of the eight segments cosegregate in an effort to identify preferences among the potential combinations. As an example, we rarely see cycles form with the segments that interact to form the polymerase complex PA, PB1, PB2, NP, indicating that these segments tend to cosegregate [100]. This

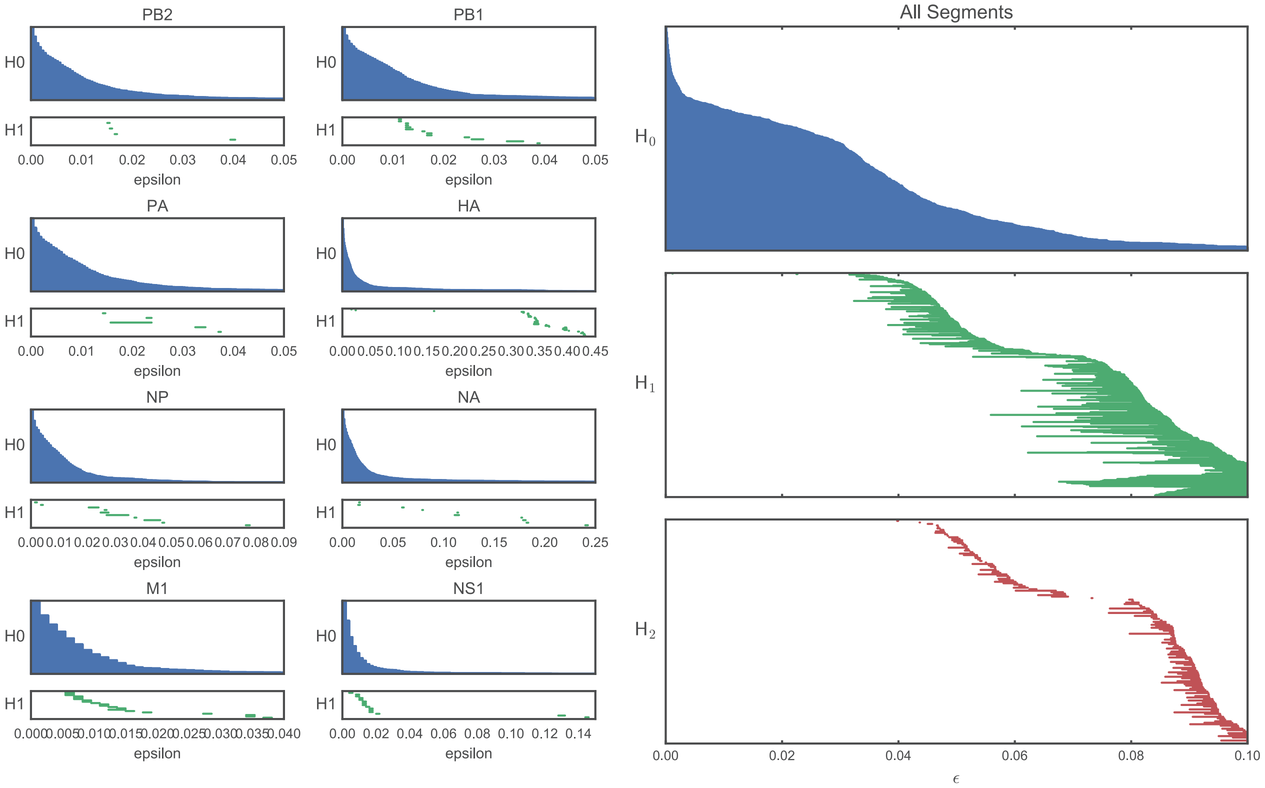


Figure 5.18 Influenza evolves through mutations and reassortment. When the persistent homology approach is applied to finite metric spaces derived from only one segment, up to small noise, the homology is zero dimensional suggesting a tree-like process (left). However, when different segments are put together, the structure is more complex revealing non-trivial homology at different dimensions (right). 3105 influenza whole genomes were analyzed. Data from isolates collected between 1956 to 2012; all influenza A subtypes.

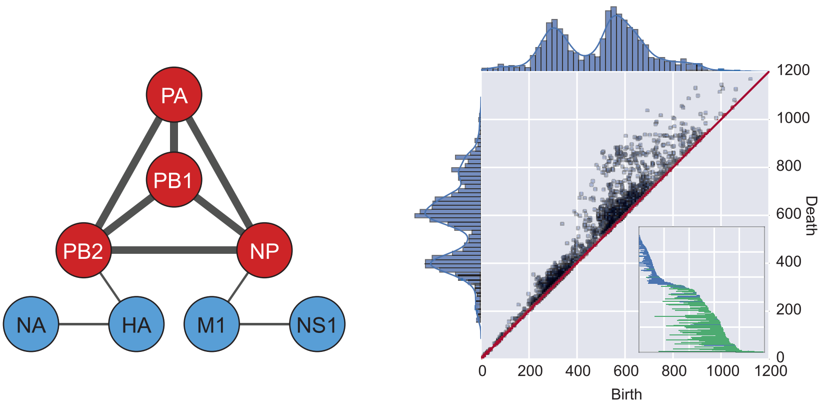


Figure 5.19 Co-reassortment of viral segments as structure in persistent homology diagrams. Left: The non-random cosegregation of influenza segments was measured by testing a null model of equal reassortment. Significant cosegregation was identified within PA, PB1, PB2, NP, consistent with the cooperative function of the polymerase complex. Source: [100]. Right: The persistence diagram for whole-genome avian flu sequences revealed bimodal topological structure. Annotating each interval as intra- or inter-subtype clarified a genetic barrier to reassortment at intermediate scales. From Joseph Minhow Chan, Gunnar Carlsson, and Raúl Rabadán, 'Topology of viral evolution', *Proceedings of the National Academy of Sciences* 110.46 (2013): 18566–18571. Reprinted with Permission from Proceedings of the National Academy of Sciences.

finding is consistent with the cooperative functioning of these proteins, which engenders negative selection against new combinations that do not cooperate as effectively (Figure 5.19).

In addition, each of the sequenced viruses (isolates) comes with information of where and when the virus was isolated, together with the hemagglutinin and neuraminidase subtype. Under the assumption that smaller cycles in the non-trivial homology classes are in some way closer genetically, one can also infer when and where the event took place and what the types of the parental strains were. Other relevant information is provided by the birth and death times of the class which provide information about how genetically distant parental viruses were. Numbers associated to one and higher dimensional classes (birth, death and size of bars in the barcode diagram) provide a useful way to summarize the type of event. The size of the bars associated to non-zero homology classes is also indicative of the type of reassortment events that could occur. The persistence diagram for whole genomes of avian flu sequences reveals bimodal topological structure (Figure 5.19, right). In other words, there are smaller bars and larger bars. Inspection of generators of different bars immediately reveals two types of reassortment processes. Small bars are generated by mixing of viruses that are closely related, belonging to the same subtype, such as two strains of H5N1 for example. Large bars, meanwhile, are generated by the mixing between the genomic material of distant viruses belonging to two different subtypes, such as H5N1 and H7N2, for example.

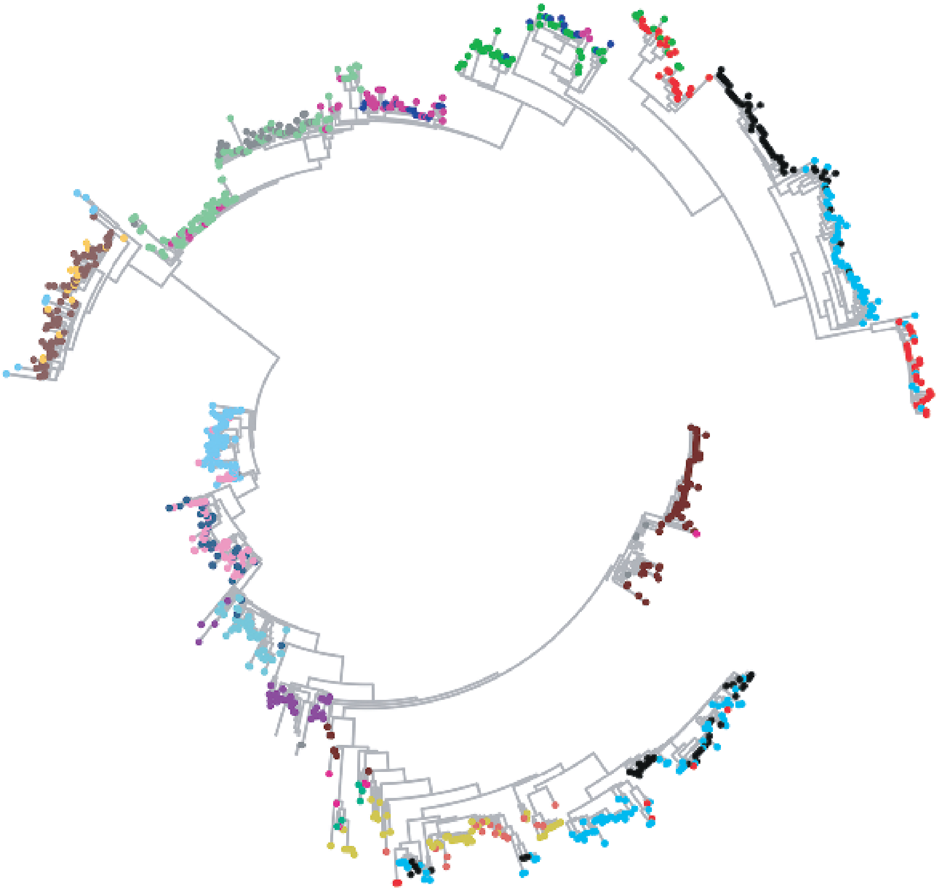
These examples show how studying finite metric spaces derived from large numbers of genomes can reveal biologically interesting phenomena and assess the flow of genomic material across different scales.

5.3.3 Influenza Virus Evolution and the Space of Phylogenetic Trees

Vaccination is probably the most effective method of reducing the morbidity associated with influenza infection. Administering a vaccine introduces a peptide with similar antigenic properties to circulating strains, causing the body to form protective antibodies against those strains. Every year, the World Health Organization selects strains for the Northern and Southern Hemispheres. Historically, it selected three different strains: two representing influenza A subtypes (H3N2 and H1N1) and one representing an influenza B subtype. Recently, a second influenza B subtype was added to make a quadrivalent vaccine containing peptides related to two influenza A and two influenza B strains. As viral genomes evolve, so does their antigenic presentation. This creates a continuous challenge to engineer new peptides that accurately represent circulating strains for use in vaccines. Ideally, one would like to have a universal vaccine able to target a wide spectrum of different strains and also future emerging strains. Interesting ideas in this vein have been put forward, but no such vaccine exists yet.

Hemagglutinin (HA) causes most of the body's antigenic response to influenza and it is the protein used in vaccines. The relation between the different isolates of the HA gene can be represented by a phylogenetic tree. Currently, more than 100,000 HA sequences can be found in public databases. With such a large sample of genomes, corresponding phylogenetic trees can become too complex to visualize or analyze. For instance, we would like to study these trees in terms of the geometry of the Billera-Holmes-Vogtmann metric space of phylogenetic trees (see Section 4.7.2). However, these spaces become increasingly complex as the number of leaves increases.

In Zairis et al., an approach involving reducing complicated trees to lower-dimensional structures by a process referred to as *tree dimensionality reduction* was proposed [545]. The idea behind tree dimensionality reduction is simple: instead of studying the properties of large trees like the one in Figure 5.20, one decomposes the large tree into a cloud of smaller trees by repeatedly subsampling the leaves of the large tree and taking the subtree determined by these leaves. In this way, one obtains a distribution of smaller trees that can capture a range of complex structural properties. This procedure has two advantages: first, it is far easier to visualize, extract, perform statistical analysis, and interpret different types of



(a) Large tree.

Figure 5.20 Evolution of influenza A virus presenting clear seasonal variation. Identifying statistical patterns in large trees is often difficult. This phylogenetic tree of the hemagglutinin (HA) segment from selected 1089 H3N2 influenza viruses across 15 seasons can be subsampled for statistical analysis in lower dimensional projections. Source: [545]. Adapted from Zairis et al., Genomic data analysis in tree spaces, arXiv: 1607.07503 [q-bio.GN].

evolutionary relationships on these smaller trees; and, second, it avoids the poor scalability of phylogenetic algorithms.

As an illustration, we describe an analysis from [545] relating HA sequences from certain seasons to those of later seasons. Zairis et al. picked random strains from five consecutive seasons from a data set of 1,089 sequences of H3N2 HA collected in the United States between 1993 and 2015. Unrooted trees were generated using neighbor-joining based on Hamming distance (a visualization of the position of these trees in tree space is shown in Figure 5.21).

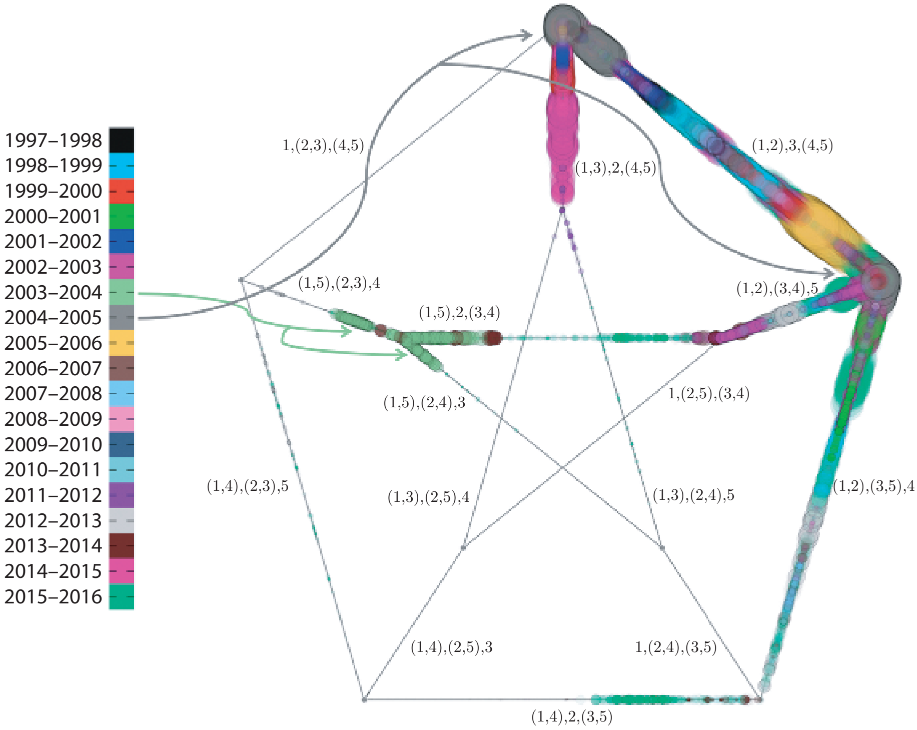


Figure 5.21 Temporally windowed subtrees in the projectivized tree metric space $\mathbb{P}\Sigma_5$. The distribution of trees derived from five-consecutive-season windows in time are superimposed on a common set of axes for projective tree space. 1089 full-length HA segments from H3N2 were collected in New York state from 1993 to 2016. Two consecutive seasons of poor vaccine effectiveness in 2003–2004 and 2004–2005 are highlighted with green and gray arrows respectively. The green distribution strongly pairs the 1999–2000 and 2003–2004 strains, hinting at a reemergence. Source: [545]. Adapted from Zairis et al., Genomic data analysis in tree spaces, arXiv: 1607.07503 [q-bio.GN].

Most of the trees showed linear evolution between seasons (the topology of the trees follows a time ordered pattern, with ancestor of strains in a season directly related to strains in the immediate previous season), indicating genetic drift as the virus's dominant evolutionary process; however, there are distinct clusters of trees in other regions of the space that indicate reemergence of strains in the 2002–2003 season genetically similar to those circulating in the 1999–2000 season.

The data was analyzed to test the hypothesis that elevated HA genetic diversity in circulating influenza predicts poor vaccine performance in the subsequent season. This amounts to staggering the seasons sampled for distributions of trees from the season of the vaccine effectiveness label, to yield an honest prediction task. Distribution features that may intuitively predict future vaccine performance

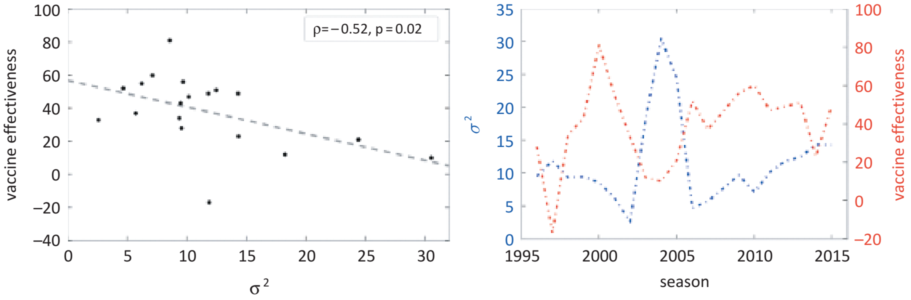


Figure 5.22 Diversity in recent circulating HA predicts vaccine failure. Negative correlation observed between vaccine efficacy in season $(t, t + 1)$ and the variance in trees generated from seasons $(t - 1, t), (t - 2, t - 1)$, and $(t - 3, t - 2)$. Source: [545]. Adapted from Zairis et al., Genomic data analysis in tree spaces, arXiv: 1607.07503 [q-bio.GN].

include the variance and the number of clusters in the point cloud. Given the limited number of temporal windows, too rich a feature space may lead to overfitting the vaccine efficacy. In Figure 5.22 we illustrate the predictions of the variance of a lagging length-3 window on vaccine effectiveness. Our notation is such that a window labeled year y would include the flu season of $(y - 1, y)$ and preceding years. The vaccine effectiveness figures represent season $(y, y + 1)$. It is clear, from both the left and right panels, that lower variance in a temporal window predicts increased future vaccine effectiveness, with a Spearman correlation of −0.52 and p-value of 0.02. The lone outlier season came in 1997–1998 [218], when the vaccine efficacy was lower than expected. In that season the dominant circulating strain was A/Sydney/5/97 while the vaccine strain was A/Wuhan/359/95.

5.4 Viral Evolution: HIV

5.4.1 Human Immunodeficiency Virus

Human Immunodeficiency Virus, or HIV, is one of the most devastating infectious diseases in modern history. Current estimates suggest 36.7 million people live with HIV today and more than 1 million die each year [244]. HIV mostly infects and destroys helper T-cells. These T-cells, also known as CD4$^+$ cells, play an essential role in the body's response to infection: they coordinate the immune response by promoting B-cells to produce antibodies and recruiting and activating neutrophils, macrophages, natural killer cells, and CD8$^+$ killer T-cells – a host of cells which neutralize invading pathogens. When CD4$^+$ T-cells die, the body's immune response is severely impaired. Pathogens that can normally be controlled by the immune system are able to infect HIV-positive patients. These "opportunistic infections" can result in the death of the infected individual. The process of CD4$^+$ T-cell depletion typically takes years and symptoms do not become evident until the cell population declines sufficiently. This clinical latent period of infection contributes to the spread of the virus through apparently healthy hosts.

HIV is a *retrovirus*. Retroviruses encode their genome in single-stranded and positive-sense RNA. When a retrovirus infects a cell, it converts its genome to double-stranded DNA in the cell's cytoplasm by first creating an antisense strand of DNA complementary to its RNA genome (cDNA) and then forming a positive-sense DNA strand complementary to the cDNA. The conversion of RNA to DNA is the opposite of the usual process in human cells, in which RNA is generated from a DNA template. It is termed reverse transcription and is facilitated by the viral enzyme reverse transcriptase (RT). After the creation of double-stranded DNA in the cytoplasm, the DNA is transported to the nucleus, where it is incorporated into the human genome. By this means, the virus gains access to the host cell's genomic machinery and its abilities to transcribe mRNA and thus the ability to translate viral proteins and replicate the viral genome (Figure 5.23).

Retroviruses are classified into two subfamilies (Orthoretroviridae and Spumaretroviridae) that include some oncoviruses, such as Rous sarcoma virus, which we will briefly describe when talking about cancer. HIV belongs to the Lentivirus genus, a taxon of retroviruses with long incubation periods before they become symptomatic and acquire the capability to infect non-replicating cells. The virions, or viral particles, of retroviruses have capsids, which surround and protect their genome, and envelopes (lipid bilayer surrounding the capsids) borrowed from the host-cell plasma membrane (Figure 5.23); specifically, HIV has a conical capsid of about 100 nm. Retroviruses contain two identical copies of the RNA genome, each around 10,000 bases in size. There are three major genes present in all retroviruses.

- The *gag* gene codes for the proteins that generate the capsid.
- The *pol* gene carries information about the enzymes necessary for replication and reverse transcription (i.e., reverse transcriptase), for integrating viral DNA into the host genome (i.e., integrase) and for cleaving viral polyproteins to activate them (i.e., protease).
- The *env* gene codes for the glycoproteins that bind to the T-cell's receptors and allow the virus to invade the host cell. Env translates directly to the polyprotein gp160, which is cleaved into two smaller proteins: gp120, which binds to the CD4 receptor and the co-receptors (CCR5 or CXCR4), and gp41, which promotes fusion of the cell membrane and viral envelope.

In addition to these three long genes, HIV has at least six smaller proteins that are involved in genomic regulation and interaction with host machinery; multiple roles

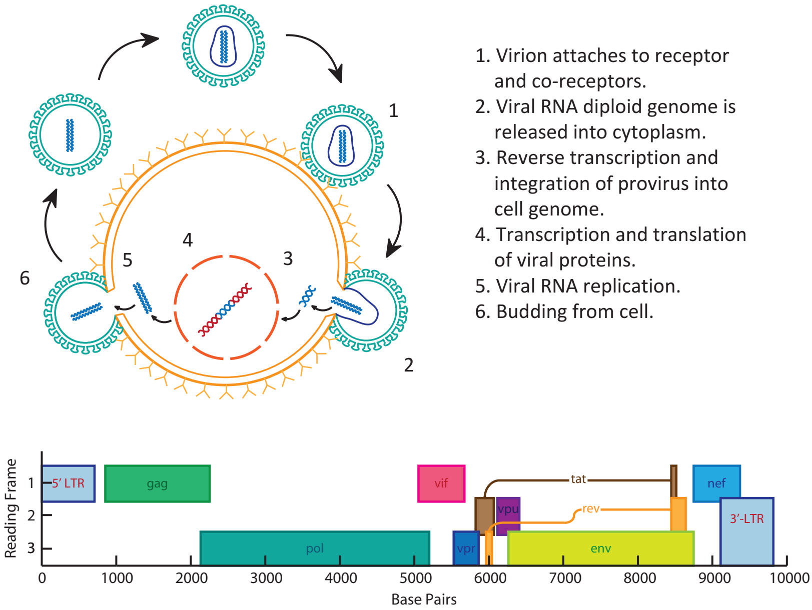

1
2
3
4
5
6
1. Virion attaches to receptor and co-receptors.
2. Viral RNA diploid genome is released into cytoplasm.
3. Reverse transcription and integration of provirus into cell genome.
4. Transcription and translation of viral proteins.
5. Viral RNA replication.
6. Budding from cell.
Reading Frame
1
2
3
5′ LTR
gag
vif
tat
nef
rev
vpu
3′-LTR
pol
vpr
env
0
1000
2000
3000
4000
5000
6000
7000
8000
9000
10000
Base Pairs

Figure 5.23 Life cycle and genomic structure of the HIV virus. Top: Life cycle of HIV. The virion attaches to CD4 receptors and co-receptors on the membrane of the CD4$^+$ T-cell, allowing for the fusion of the viral envelope with the T-cell membrane and the release of the viral RNA into the cell's cytoplasm. The viral reverse transcriptase reverse transcribes the viral genome into double-stranded DNA, which is transported into the cell's nucleus and integrated into the host's DNA. After integration, the host cell's genomic machinery treats the integrated virus, or provirus, as part of the host genome, generating mRNA and viral protein, and copies of the RNA genome. Two copies of the HIV genome are packaged in each virion and the virions bud from the host cell. In the final process of maturation, cell-free virions assemble conical capsids that stabilize their genomes. These mature virions are now able to infect other cells. Bottom: The genome of HIV consists of three large genes, gag, pol and env, common to most retroviruses, and six small genes that arise from subsequent splicing events.

have been reported for each of these proteins. The Trans-Activator of Transcription (Tat) is a small protein of around 100 amino acids that binds to cellular factors in order to increase transcription of all HIV genes, including itself, thus creating a positive-feedback loop of transcription. The regulator of the expression of virion proteins (Rev) is necessary for the synthesis, stability and transport of several viral mRNAs. The Viral protein R (Vpr) has about 100 amino acids and among other functions, transports the pre-integrated viral genome into the host's nucleus. The Viral infectivity factor (Vif) inhibits the cellular protein APOBEC3G. APOBEC proteins are cytidine deaminases, proteins that induce mutations in cytidines, that catalyze the deamination of cytidine to uridine, introducing a large number of C-to-U or C-to-T mutations in RNA or DNA respectively in localized settings.

APOBEC3G enters the virion and mutates the viral genome, resulting in hypermutated genomes causing defective viruses. Vif prevents APOBEC3G activity by targeting it for proteasomal degradation [454, 544]. Beyond blocking APOBEC3G activity, it has also been associated with the infectivity of virions. Finally, the Viral protein Unique (Vpu) has been implicated in the degradation of host-cell CD4 receptors and the release of virions.

It remains unclear exactly when, where, and how HIV became a human pathogen [453]. The disease associated to the virus, the Acquired Immunodeficiency Syndrome, or AIDS, is caused by two related retroviruses, HIV-1 and HIV-2. In the developed world, AIDS was identified through a sudden increase in rates of opportunistic infections and very rare tumors in injection drug users and men who have sex with men. The opportunistic infections included *Pneumocystis jirovecii* pneumonias, previously reported to occur in individuals with highly compromised immune systems, and the tumors included Kaposi sarcoma, later shown to be itself caused by an infection [185]. In 1983, two groups in the United States and France reported a new retrovirus associated with this immunodeficient state [36, 190]. For this work, Françoise Barré-Sinnousi and Luc Montagnier won the Nobel Prize in Physiology or Medicine in 2008 (Figure 5.24). A second HIV virus, named HIV-2, was reported in West Africa in 1986 with a similar, although not identical, genomic structure to HIV-1.

The virus was then identified in the general population living in Africa [410]. Infection rates indicated that the virus was already circulating in African

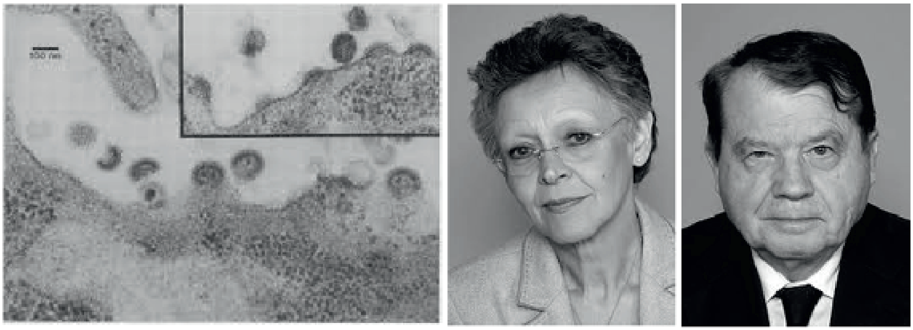


Figure 5.24 Identification of HIV as a cause of AIDS. Left: Electron microscopy of sections of HIV virus producing cells. Source: [36]. From F. Barré-Sinnoussi et al., Isolation of a T-Lymphotropic Retrovirus from a Patient at Risk for Acquired Immune Deficiency Syndrome (AIDS), *Science*, New Series, Vol. 220, No. 4599, pp. 868–871, 1983. © 1983 American Association for the Advancement of Science. Reprinted with permission from AAAS. Right: Françoise Barré-Sinnousi and Luc Montagnier, who won the Nobel Prize in Physiology or Medicine in 2008 for the discovery of the virus. Source: © The Nobel Foundation. Photo: Ulla Montan.

populations before it was identified in the Western world. More recently, sampling of HIV viruses in Central Africa has shown a higher genetic diversity compared with other viruses collected all around the world, suggesting an older African origin [516]. That was supported by retrospective studies that identified the virus in blood samples from patients in Kinshasa at the end of the 1950s [540]. A Norwegian sailor, Arvid Darre Noe, was reported to be infected with HIV-1 group O, most likely in 1961 or 1962 when working in Cameroon [186]. The closest relatives of these viruses infecting other species can be found in African primates. It is now believed that there were multiple transmission events leading to the major subclades of the virus. Some of these transmission events, such as that of group M from chimpanzees in Central Africa, led to rapid spread throughout the human population. A recent study using HIV-1 env sequence data from different countries in the Congo River basin suggests that the most recent common ancestor of all group M strains dates back to 1920 in the Democratic Republic of Congo [169]. Several societal changes occurring at that time, including the growth of African cities and the mobility of workers, have been discussed as potential factors contributing to the spread of the virus.

5.4.2 Viral Recombination in HIV

HIV is notorious for its high diversity, created and maintained not only by its high mutation rate but also by frequent recombination. Using data from patients, mutation rates of HIV have been estimated to be $(4.1 \pm 1.7) \times 10^{-3}$ per base per cell [129]. Many of these mutations, however, are lethal to the virus and only a small fraction can make functional viruses. The major causes of mutations in vivo are the reverse transcriptase and cytidine deaminases (in the process of retrotranscription), although human DNA-dependent RNA polymerase can also contribute when generating viruses from the integrated provirus. On average, mutations caused by RT only constitute 2% of all mutations; but this statistic varies to a large degree across patients. Patients that rapidly progress to the symptomatic stage experience fewer hypermutations (accumulation of a large number of mutations in a virus), suggesting that cytidine deaminases play an important role in HIV pathogenesis.

Because the genome of HIV is not segmented, reassortment does not occur. Instead, recombination is the major driver of horizontal evolution. RT's polymerase can use either genomic RNA strand as a template for reverse transcription and it can switch between strands during the process (Figure 5.25). If the two RNA strands packaged in a virion include distinct mutations or come from different parent viruses, template switching by RT can create a mosaic genome.

These recombinations can occur commonly, and recombinants can become the dominant forms circulating in large fractions of host populations. Circulating

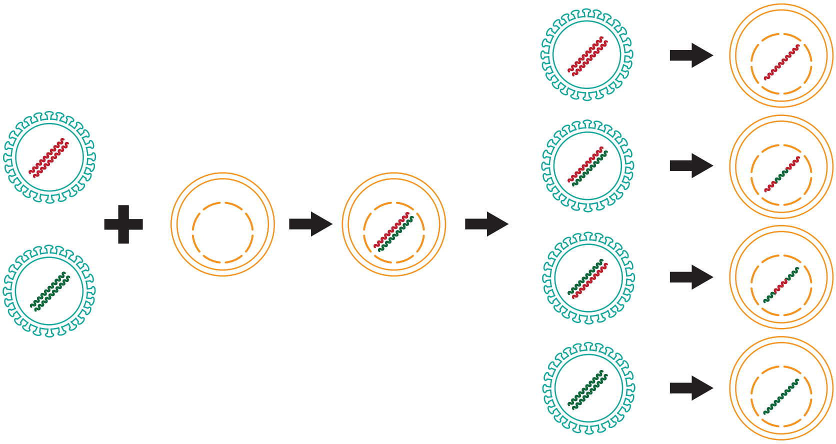


Figure 5.25 Recombination in HIV. The genome of HIV is diploid, containing two more-or-less identical copies of the RNA genome. Virions, however, can be packaged with two very different copies if two distinct HIV viruses co-infect the same cell. When reverse transcribing the RNA from these virions into a single copy of DNA, the polymerase can jump between the two strands, generating a mosaic virus containing fragments of both parental strands.

Recombinant Forms, or CRFs, are common recombinants deriving from recombination between viruses of different subtypes. The notation and naming of CRFs is complex because different "pure" parent subtypes can generate many different mosaic viruses. The breakpoint of recombination can occur anywhere along the genome and multiple breakpoints are common. Barred by frequent recombination, drawing an evolutionary tree from a single gene is virtually impossible. As expected, and in contrast to influenza, when applying persistent homology to HIV, individual genes reveal large numbers of one and higher dimensional homology classes, indicating a history of reticulate events, most likely recombination (Figure 5.26). When concatenating the large genes of the virus, large recombination events are uncovered, relating multiple parental strains of subtypes A. An example of a long bar observed in two dimensional homology is shown in Figure 5.27, revealing a complex recombination event between major HIV subtypes, B, C, D, F, and 13cpx, a complex recombinant strain.

5.4.3 Viral Recombination in Late-Stage HIV Infection

We have seen that untreated HIV can lead to an impaired immune system. However, there are other symptoms that occur in patients with long-term infections. HIV-associated dementia (HAD) is a condition associated with long-term viral progression and low $CD4^+$ T-cell counts. This condition is the most severe of

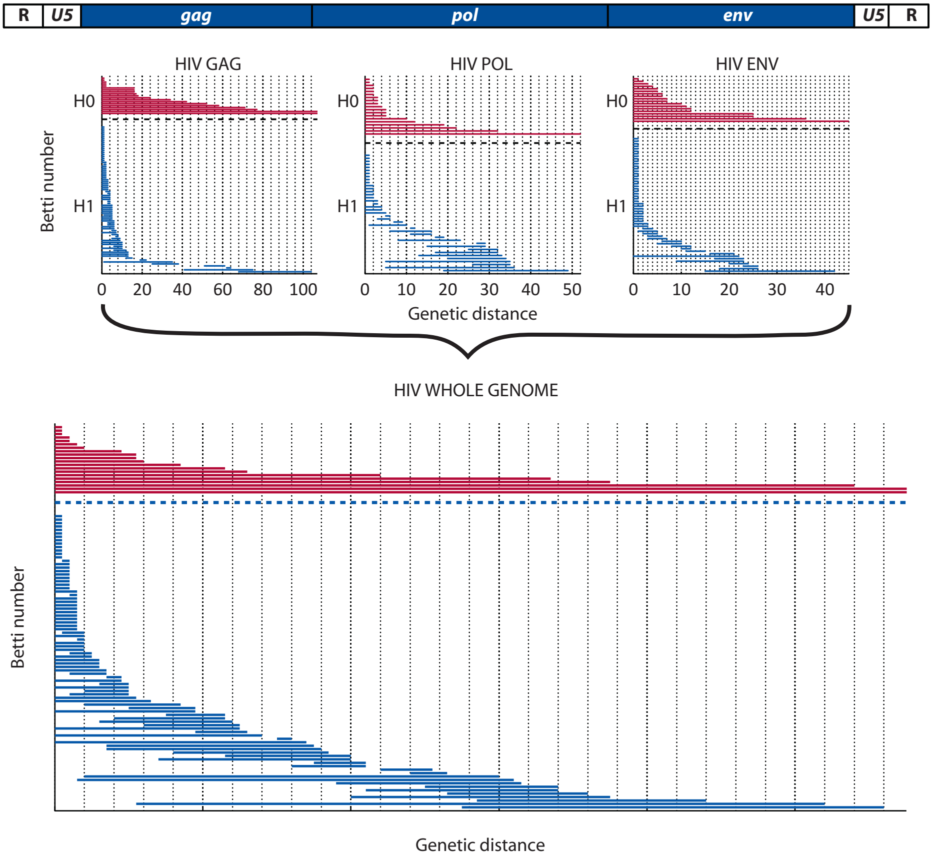


Figure 5.26 Persistent homology reveals recombination within genes and across the genome. Unlike in influenza, persistent homology barcodes of HIV reveal intragenic recombination in the three major HIV genes gag, pol and env. When concatenated and run through the persistent homology pipeline, the multi-gene fragments have homology classes in dimensions one and higher. Source: [100]. From Joseph Minhow Chan, Gunnar Carlsson, and Raúl Rabadán, ‘Topology of viral evolution’, *Proceedings of the National Academy of Sciences* 110.46 (2013): 18566–18571. Reprinted with Permission from Proceedings of the National Academy of Sciences.

the HIV-associated neurocognitive disorders, which are believed to result from exposure of the brain to high levels of HIV-1 following breach of the blood-brain barrier by HIV-infected monocytes [287]. While instituting combination antiretroviral therapy early in infection may prevent neurocognitive decline, later initiation of therapy does not appear to reverse pre-existing symptoms [503]. Understanding the nature of the viral population in the brain is therefore of ongoing interest. Virus sampled from the cerebrospinal fluid (CSF) or brain of HAD-affected individuals is often genetically distinct from that of the peripheral blood, suggesting

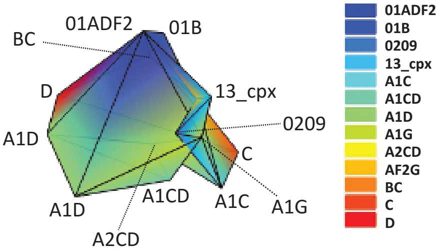


Figure 5.27 Here is a polytope representing complex recombination events with multiple parent strains. This polytope represents a two-dimensional class in persistent homology. Each vertex of the polytope represents a sequence that is colored according to HIV-1 subtype. Source: [100]. From Joseph Minhow Chan, Gunnar Carlsson, and Raúl Rabadán, 'Topology of viral evolution', *Proceedings of the National Academy of Sciences* 110.46 (2013): 18566–18571. Reprinted with Permission from Proceedings of the National Academy of Sciences.

continuous viral replication in the brain as a potential cause of HAD [311]. Moreover, viral recombination may occur more frequently within the populations found in the brains of individuals affected with severe HAD than in other HIV-infected individuals, further implicating unchecked viral replication as a cause of HAD.

In this section, we describe how tools of persistent homology can be used to characterize this viral recombination to study intra-host HIV evolution in patients with long-term viral progression. In particular, we are interested in understanding how recombinant viruses spread between different tissues. This can be done by comparing genomic sequences from the central nervous system (CNS) to sequences obtained from other tissues. Zigzag persistent homology [93], described in Section 2.5, provides a formalism to study and compare events across different populations.

Lamers et al. [310, 311] obtained tissue samples from the autopsies of 11 individuals who died from AIDS. They extracted genomic HIV DNA and amplified a 3.3 kb fragment stretching from env to the 3′ LTR by PCR, cloned it, and sequenced it. They published sequences of the glycoprotein gp120 ($\approx$ 1200 bp) found in the peripheral tissues of seven individuals and, for five of the individuals (Patients AZ, BW, CX, DY, GA), included sequences from the CNS. Patients AM and IV only had sequences from non-CNS tissues reported. A summary of the data is shown in Table 5.1.

Recall from Section 2.5 that zigzag persistence provides a formalism to describe "filtrations" where arrows can go in both directions; for example, when the data can be modeled by a mathematical object that first "builds up" (the "zig") and later "breaks down" (the "zag") [91, 93]. For sequences sampled from two related subpopulations, this framework provides a way to divide recombination events into four classes:

Table 5.1 *Summary of patient data: first column is the identifier of the patient, second and third columns are the number of sequences obtained from the central nervous system, fourth columns is the GenBank accession numbers*

Patient	# CNS sequences (unique sequences)	# Non-CNS sequences (unique sequences)	Accession Numbers
AZ	35 (33)	52 (48)	HM001587 – 1673
DY	107 (99)	59 (54)	HM002004 – 2169
BW	103 (99)	18 (18)	HM001674 – 1794
CX	162 (152)	47 (43)	HM001795 – 2003
GA	75 (73)	57 (55)	HM002170 – 2301
AM	—	225 (210)	HM001362 – 1586
IV	—	181 (177)	HM002302 – 2482

1. event occurring in the first population, but not the second;
2. event occurring in the second population, but not the first;
3. event detectable in either population alone (typical if the two populations are very closely related);
4. events involving both populations, detectable only in some union of their sequences and not in either population individually; this class represents the case of gene flow between genetically distinct populations.

Consider the reticulate phylogeny shown in Figure 5.28A, where the red nodes (left node in each numbered pair) are sampled from one population (e.g., geographic region or anatomical site) and the yellow nodes (right node in each pair) are sampled from a second population. Computing persistent homology identifies the recombination event as a topological loop that appears at particular scales (see Figure 5.28C). Visually, it is clear that a single recombination event has affected both populations, and can be seen from either population. Zigzag persistence allows us to recover this computationally. Starting from the first population alone, a loop is detected (Figure 5.28B). Complexes are built up (the “zig”) by adding sequences from the second population (Figure 5.28C) and broken down (the “zag”) by removing sequences from the first population (Figure 5.28D). The zigzag barcode captures the fact that the loop in panel B and the loop in panel D are representatives of the same homology class, indicating that the same recombination event generated them – a class 3 event. The ancestry represented in panel E contains a recombination event that brings together the red and yellow populations. The sequence of simplicial complexes starts as a single line (panel F), builds up to a square (panel G), and breaks down to a different line (panel H). As a loop

Table 5.2 *Patient status and putative recombination events indicated by persistent homology*

Patient	HAD status	Degree of neuropathology	# CNS events	# Cross-site events	# Non-CNS events
AZ	None	3	0	1	2
DY	Acute	1	2	5	1
BW	Progressive	2	3	0	0
CX	Progressive	5	8	0	1
GA	Progressive	5	5	7	8
AM	n/a	n/a	—	—	7
IV	n/a	n/a	—	—	9

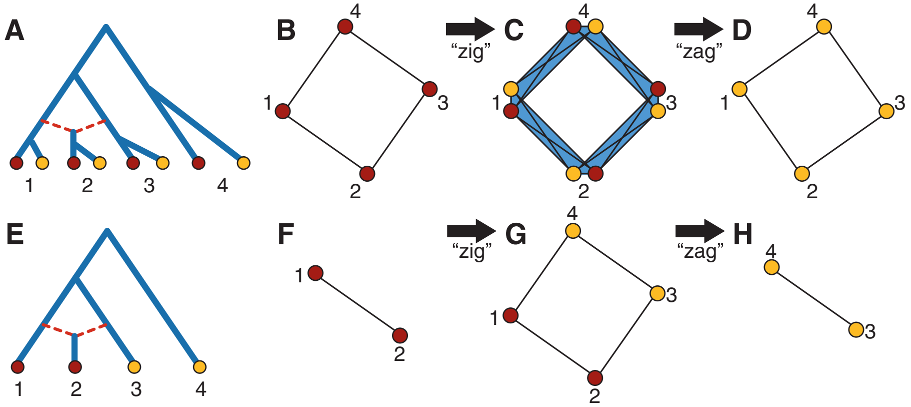


Figure 5.28 Schematic of zigzag persistence used to identify inter-population recombination (see text).

appears only when both populations are included (class 4 recombination event), this identifies exchange of genomic material between populations.

Summarizing, persistent homology was used to identify putative recombination events. Where sequences from both CNS and non-CNS sources were available, zigzag persistence was used to classify each recombination as occurring in the CNS, outside the CNS, or between CNS and non-CNS sequences (Table 5.2). The two patients exhibiting progressive HAD and the most severe neuropathology – CX and GA – also had the greatest number of recombination events localized in the CNS, suggesting that frequent viral recombination contributes to this disorder. Apart from this similarity, the viral population structure was strikingly different for the two patients: patient GA's CNS sequences were relatively more intermingled with the non-CNS sequences, with frequent recombination events occurring between the two anatomical groups. In contrast, the two groups in patient

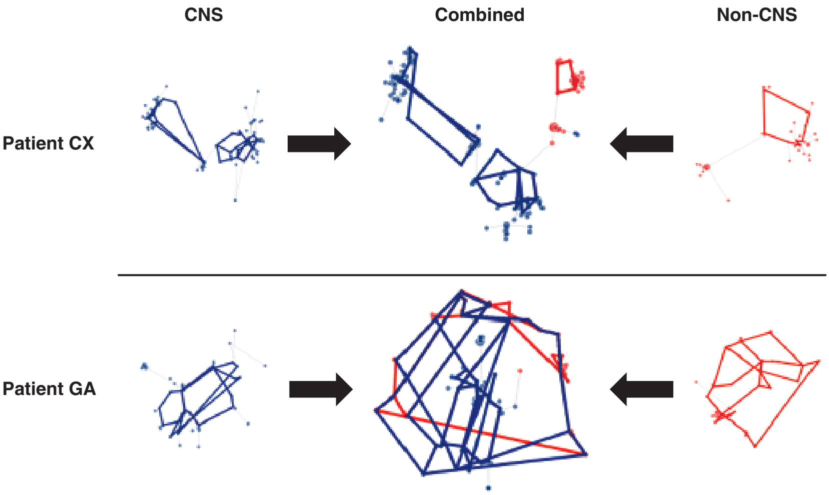


Figure 5.29 Phylogenetic networks of HIV-1 gp120 sequences obtained from patients CX and GA. Each node represents one sequence; larger nodes show sequences that were sampled multiple times. Blue nodes were sampled from the CNS; red nodes were sampled from elsewhere in the body. The position of each node is determined by the first two principal components (computed via MDS) of genetic distance (Hamming distance). The network backbone (thin gray edges) is a minimum spanning tree, and the thick red and blue edges are generators of cycles identified by persistent homology. Red cycles denote putative recombination events that involve sequences sampled fully outside the CNS; blue cycles denote events that involve some sequences from the CNS.

CX were more clearly separated. Figure 5.29 depicts phylogenetic networks of the sequences for these two patients, illustrating this difference in structure.

Since the number of sequences sampled can affect the number of cycles observed, in Table 5.3 we show $\hat{\rho}_{PH}$, an estimate of the population-scaled recombination rate, as described in Section 5.7. Again patients CX and GA stand out as having CNS populations with the highest recombination rate, suggesting that the association of HAD with severe neuropathology is not an artifact of the sampling procedure.

If the CNS and non-CNS populations are completely distinct, then the population-scaled recombination rate ρ for the combined population will equal the sum of the ρ values for each individual population. For most patients, the value of $\hat{\rho}_{PH}$ computed using all sequences is in fact less than the sum of the two $\hat{\rho}_{PH}$ values computed from the CNS and non-CNS samples. This is consistent with the two populations being partially intermingled and sharing common ancestral recombination events, such that much of the historical signal can be obtained by sampling just a single population. Patient DY was unique in that $\hat{\rho}_{PH}$ for the combined population exceeded the sum of the two individual values. Consistent with this observation, Table 5.3 $\hat{\rho}_{PH}$ *estimated from different sources*

Patient	$\hat{\rho}_{PH}$ from CNS sequences	$\hat{\rho}_{PH}$ from non-CNS sequences	Sum of both estimates at left	$\hat{\rho}_{PH}$ from all sequences
AZ	0	6.7	6.7	4.4
DY	4.0	3.0	6.9	11.0
BW	5.9	0	5.9	5.3
CX	12.7	3.7	16.3	12.5
GA	12.5	27.4	39.9	35.2
AM	—	9.2	—	—
IV	—	13.1	—	—

patient DY was also the only individual in which the majority of recombination events observed occurred between representatives of the two populations ("cross-site events" in Table 5.2). These observations suggest considerable recent traffic of virus across the blood-brain barrier in this patient, perhaps borne by increased traffic of macrophages stimulated by the *Mycobacterium avium* infection that started a year prior to death. Although there is statistically significant clustering of the two populations, it is weakest in this patient compared to the others [255].

5.5 Other Viruses

Most of our knowledge of microbes relates to human pathogens, of which there are on the order of 10^3 species, representing a tiny fraction of all microbial species. It has been estimated that there are 10^{31} viruses on this planet [158, 488], constituting the largest and most diverse biological population on Earth. About 8% of our DNA is derived from remnants of viruses that once infected our ancestors. While all cells in the three domains of life store their genomes as double-stranded DNA, viruses use RNA and DNA in different forms. The taxonomy of viruses is extremely complex as there are no common structures shared by all viruses, and there is no clear evidence that all viruses share a common origin. The Baltimore classification [28], a common classification based on the type of genomic material and replication strategy, divides viruses into seven different groups.

- Group I: double-stranded DNA viruses.
- Group II: single-stranded DNA viruses. Unlike cells, these viruses use only one strand of DNA.
- Group III: double-stranded RNA viruses.
- Group IV: single-stranded RNA viruses, with genomic material encoded in the positive-sense strand.

- Group V: single-stranded RNA viruses, with genomic material encoded in the negative-sense strand.
- Group VI: single-stranded positive RNA viruses that use reverse transcription.
- Group VII: DNA viruses that use reverse transcription.

We have seen that influenza uses negative-sense RNA for its genomic material, so it is classified in the type V group. HIV is a retrovirus, using RNA and reverse transcription, and thus it is classified as type VI. An example of a type I virus is the Epstein-Barr virus, which causes mononucleosis, and which we will encounter again when talking about cancer. This classification may be neat, but it does not provide information about the origins of viruses, and two viruses belonging to the same group may have very little in common genetically, while viruses from different groups may have related genes. Such similarities could be due to a common ancestor or to different exchange modes of genomic material.

The same persistent homology approach that we used to study reassortment in influenza and recombination in HIV can be applied to study other viruses. Flaviviridae is a family of viruses comprising several different genera, including hepaciviruses and flaviviruses. Flaviviridae are positive-sense single-stranded RNA viruses (group IV), whose ability to perform homologous recombination through RNA polymerase template switching has been debated. Sporadic recombinant strains have been detected for hepaciviruses like hepatitis C [120] and flaviviruses like dengue virus [539] and West Nile virus [409]. In some of these cases, the evidence for recombination remains controversial [426]. One can use persistent homology to study the extent of recombinations in the Flaviviridae family [100]. Comparing using different measures such as the size of the longest bar (TOP) and the number of bars in the sample time (ICR), it was found that hepatitis C showed some but lower recombination than in HIV (Figure 5.30). No high-dimensional homology was found in dengue or West Nile virus, suggesting that recombination rarely occurs in these viruses.

In type V viruses, like influenza, recombination is considered to be an even less frequent event like in Newcastle or Rabies virus. Persistent homology does not identify high-dimensional classes for rabies, while the analysis of Newcastle virus confirmed a low ICR but a non-vanishing TOP.

5.6 Bacterial Evolution

Bacteria are the most common cells on Earth and even in our bodies. From marine samples, biologists estimate that there are 3×10^{28} bacterial cells on Earth. Despite being less numerous than viruses, these prokaryotes represent more than 90% of Earth's biomass [158, 488]. The bacteria in a human's gut collectively weigh about

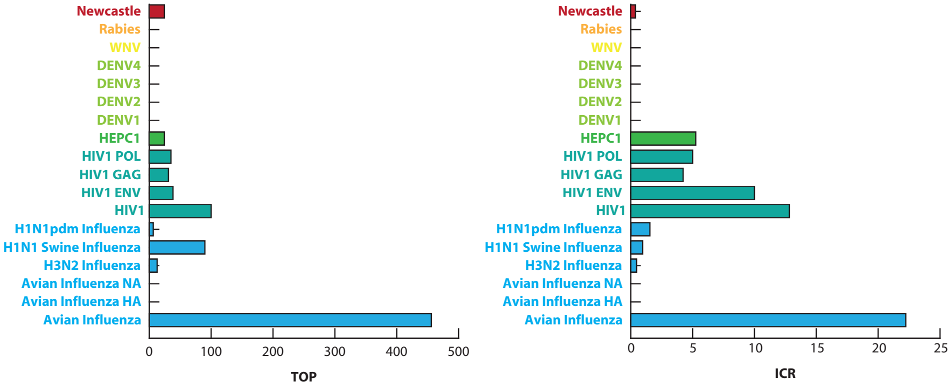


Figure 5.30 Recombination across different viruses. Left: A topological obstruction is estimated using the maximum barcode length in dimension one. Right: The rate of irreducible cycles is defined as the number of one dimensional bars in the barcode diagram divided by the time spanned by the sequence collection. Source: [100]. From Joseph Minhow Chan, Gunnar Carlsson, and Raúl Rabadán, 'Topology of viral evolution', *Proceedings of the National Academy of Sciences* 110.46 (2013): 18566–18571. Reprinted with Permission from Proceedings of the National Academy of Sciences.

a kilogram. In a gram of dental plaque there are 10^{11} bacteria. Only a small fraction of bacterial species has been characterized so far. Although large multidisciplinary efforts are under way, such as the Earth Microbiome Project (which plans to study 200,000 samples) it is unlikely that we will have a comprehensive atlas in the near future.

5.6.1 Horizontal Gene Transfer in Bacteria

Bacterial genomes vary widely in size; typically they are a few megabases long. *Mycoplasma genitalium*, an intracellular pathogenic bacterium, has one of the smallest genomes at half a megabase. *Escherichia coli*, a common bacterium living in our intestine and used in laboratories, has a genome of 4.6 megabases. Its mutation rate has been found to be 5.4×10^{-10} per base per replication, or 0.0025 per genome per replication [149, 150]. Mutation rates vary between species, but also with changes in the ambient environment. For example, it has been shown that starving bacteria have dramatically increased evolutionary rates [72, 86].

In addition to mutations, horizontal gene transfer (HGT), the exchange of genomic material in a non-vertical way, constitutes a major form of genetic innovation in bacteria. Borrowing genes through HGT allows for rapid adaptation to challenging environments [389]. As we will see, HGT has been found to be a major factor in the spread of antibiotic resistance [134]. Transfer of genetic material is well known since the work of Lederberg on bacterial conjugation in the 1940s [318] (Lederberg received the 1958 Nobel Prize for this work). Until the advent of large scale genomic studies, it was widely thought that HGT was a rare event. Now it is known that effects of HGT are found pervasively across many different bacterial species [305]. In some cases the effect of HGT is extremely dramatic, in particular when genes are imported across different domains of life. For instance, the genomes of some bacteria contain a large fraction of archaeal genes. The best known example of this borrowing is that of hyperthermophilic bacteria, which are bacteria that can tolerate temperatures near boiling, such as *Aquifex aeolicus* and *Thermotoga maritima*. In genomic analysis, HGT is usually identified through incongruent tree phylogenies, with different gene histories represented by incompatible tree topologies. The widespread effect of HGT across and within different domains of life has led some to question the existence and usefulness of representing the relationship between distant bacterial species in a Tree of Life [147].

There are three main molecular mechanisms by which HGT can occur (see Figure 5.31) [389].

- *Transformation*: the uptake of naked, free-floating DNA from the environment.
- *Transduction*: the transfer of genomic material through a virus intermediate. Viruses that infect bacteria, known as bacteriophages or phages, mediate the transduction process. The amount of DNA is limited by the size of a viral capsid, usually about 100,000 bases. Phages also can encode proteins that can help the integration of the new material into the receptor cell.
- *Conjugation*: transfer of genomic material by cell-cell contact. For this to occur, the cytoplasms of the bacteria must be connected. Bacteria often connect to each other using an appendage called a pilus. The pilus exists precisely for this role, demonstrating that HGT can be advantageous for bacteria.

HGT can be hindered by disruptions in any of the following processes: in the donor, the ability to generate genomic material in the form of free DNA or plasmids; a transportation method for the DNA, such as the existence of phages that

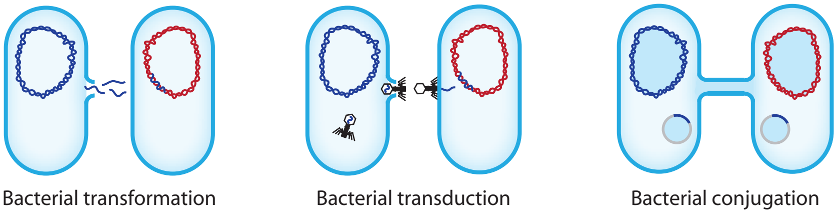


Figure 5.31 A few mechanisms of horizontal gene transfer in bacteria: transformation, transduction and conjugation.

can effectively infect both the donor and recipient; and in the recipient, the capacity to uptake and integrate the new DNA.

Experimentally, it has been shown that HGT between species decreases with increasing genetic distance [182]. In the following section, we will employ genomic data from large databases and tools from TDA to study the frequency and patterns of intra- and inter-species HGT in bacteria.

5.6.2 Pathogenic Bacteria

As previously mentioned, horizontal exchange occurs when a donor bacterium transmits foreign DNA into a genetically distinct bacterial strain; for instance, in Germany, 2011, *E. coli* acquired the Shiga toxin, typical to the *Shigella* genus, via phage-mediated gene transfer, and caused a serious outbreak of foodborne illness [435]. Control of bacterial pathogens is hampered by rampant horizontal gene transfer, which allows bacteria to acquire genes conferring resistance to commonly used antibiotics [382, 391, 497]. Genes for resistance can be transferred between strains of both the same and different species existing in the same environment. Elements of bacterial genomes demonstrating evidence of foreign origin are known as genomic islands and may be associated particularly with phenotypic effects, such as virulence or resistance to antibiotics.

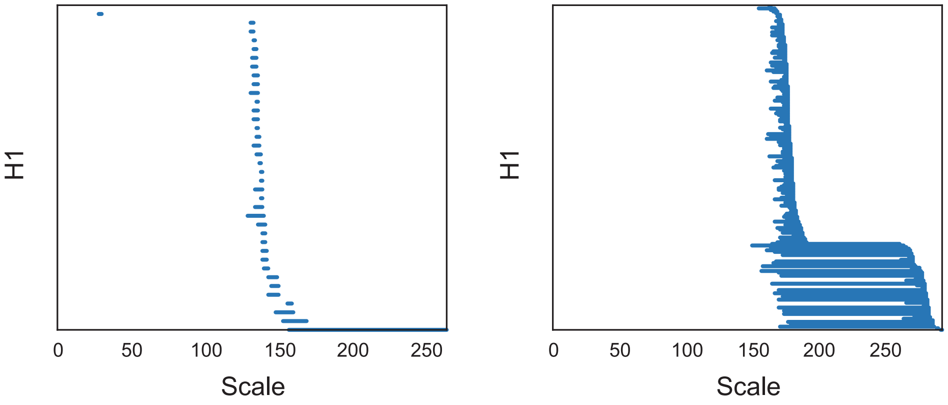
Tools from topological data analysis can help to characterize the frequency and scale of horizontal gene transfer in bacteria, elucidating issues of significant public health relevance, such as the spread of antibiotic resistance in *Staphylococcus aureus* and the human microbiome's role as a reservoir for antibiotic resistance genes.

5.6.3 Multilocus Sequence Typing Analysis

Within a single bacterial species there can be many genetically distinct strains. Different strains can have important functional differences. For example, some strains may be more virulent than others and some may be more susceptible to the immune responses generated by vaccines. Multilocus sequence typing (MLST) is a method for detecting particular bacterial strains that does not require whole-genome sequencing. It relies on the fact that strains can be identified from certain representative genomic loci selected from regions within housekeeping genes. Typically the size of each locus is about 500 base pairs.

Curated MLST data from laboratories around the world is available in large online databases. Often there are thousands of strains identified within a single pathogenic species (over 10,000 in the case of *Neisseria* spp.). MLST data can be used to study horizontal exchange of genomic material in bacteria. Because different species have different loci, one can only examine horizontal exchange within species. Furthermore, because all of the selected loci exist within a few housekeeping genes, our analysis does not provide information about events involving genes other than these housekeepers.

The data used here comes from PubMLST [277]. For each of twelve bacterial species, one can construct a pseudogenome by concatenating the typed sequence at each locus. Using the Hamming distance metric, one can calculate a pairwise distance matrix between strains and compute persistent homology on the resulting metric space. In Figure 5.32, we show the persistent homology barcodes associated to the witness complex (recall Definition 2.7.3) with 250 landmark points. We plotted the H_1 barcode diagrams for *K. pneumoniae* and *S. enterica*. Based on the observed range of recombinations, one can identify two distinct species profiles: *K. pneumoniae* recombines solely at one short-lived scale, while *S. enterica* recombines both at the short-lived scale and also at another longer-lived scale. This analysis can be repeated for each species; we plotted the results as persistence diagrams in Figure 5.33. For the bulk of pathogens, there are three major scales of recombination: one short-lived scale at intermediate distances, another longer-lived scale at intermediate distances, and a third short-lived scale at longer distances. *H.*



Klebsiella pneumoniae

Salmonella enterica

Figure 5.32 Barcode diagrams reflect different scales of genomic exchange in *K. pneumoniae* and *S. enterica*. Source: [161]. Reprinted by permission from Springer Nature: Emmett K. J., Rabadán R. (2014) Characterizing Scales of Genetic Recombination and Antibiotic Resistance in Pathogenic Bacteria Using Topological Data Analysis. In: Ślęzak D., Tan A. H., Peters J. F., Schwabe L. (eds) *Brain Informatics and Health*. BIH 2014. Lecture Notes in Computer Science, vol 8609. Springer, Cham. © Springer International Publishing Switzerland 2014.

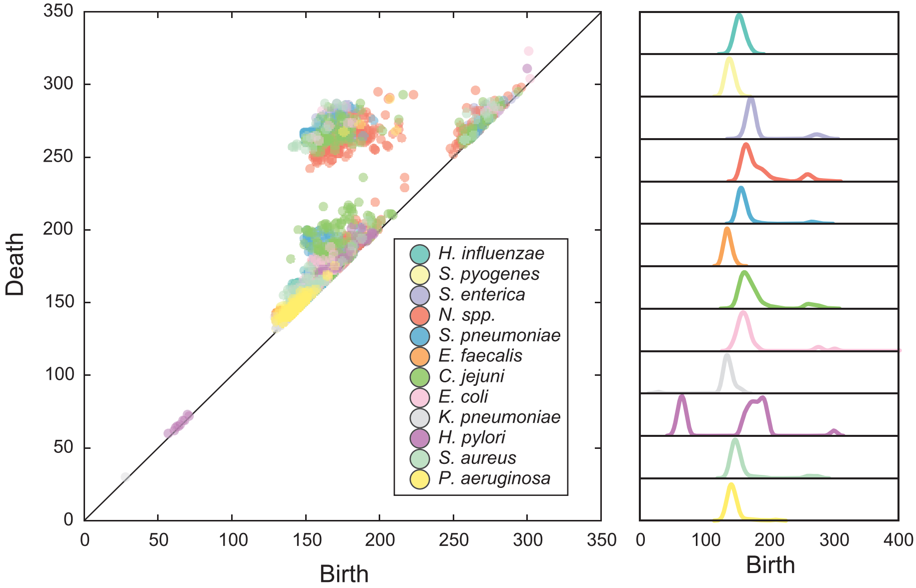


Figure 5.33 On the left, the H_1 persistence diagram for the twelve strains of pathogens selected for this study MLST profile data. Observe three scales of recombination. On the right, the birth time distribution for each strain. There is an earlier scale of recombination present in *H. pylori* not observed in the other species. Source: [161]. Reprinted by permission from Springer Nature: Emmett K. J., Rabadán R. (2014) Characterizing Scales of Genetic Recombination and Antibiotic Resistance in Pathogenic Bacteria Using Topological Data Analysis. In: Ślęzak D., Tan A. H., Peters J. F., Schwabe L. (eds) *Brain Informatics and Health*. BIH 2014. Lecture Notes in Computer Science, vol 8609. Springer, Cham. © Springer International Publishing Switzerland 2014.

pylori is a clear outlier, tending to recombine at significantly lower scales than the other pathogens.

A relative recombination rate can be defined by counting the number of H_1 loops across the filtration and then dividing by the number of samples for that species. The results of this analysis are shown in Figure 5.34, which demonstrates that there exist a wide range of recombination profiles among bacterial species. *S. enterica* and *E. coli* have the highest recombination rates, while *H. pylori* recombines at a substantially lower rate than the others. This analysis suggests that *H. pylori*'s core genome is comparatively impervious to recombination except by closely related strains.

5.6.4 Protein Family Analysis

MLST data can provide information about the exchange of genomic material in typed loci within related species. In order to study horizontal exchange between

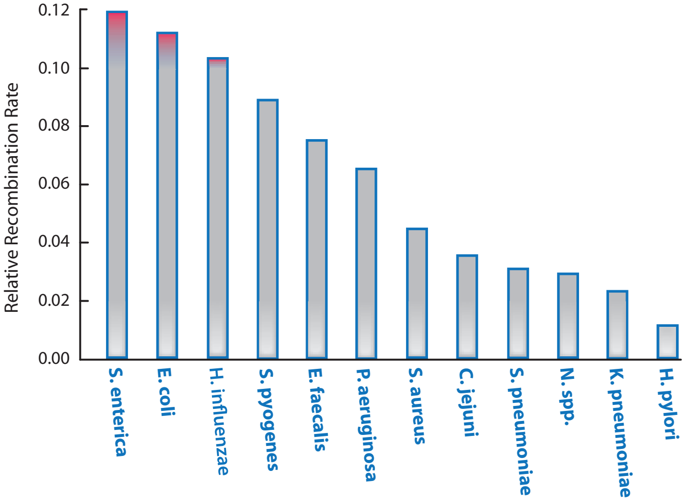


Figure 5.34 Relative recombination rates computed by persistent homology from MLST profile data. Source: [161]. Reprinted by permission from Springer Nature: Emmett K. J., Rabadán R. (2014) Characterizing Scales of Genetic Recombination and Antibiotic Resistance in Pathogenic Bacteria Using Topological Data Analysis. In: Ślęzak D., Tan A. H., Peters J. F., Schwabe L. (eds) *Brain Informatics and Health*. BIH 2014. Lecture Notes in Computer Science, vol 8609. Springer, Cham. © Springer International Publishing Switzerland 2014.

different species, one needs data that are relevant across bacterial species. One approach is to consider the presence or absence of protein families among different bacterial species. Protein families are proteins with similar sequence and function. The presence of a member of a protein family in a strain could be due to a horizontal gene transfer event between strains or species.

The presence or absence of protein families can be converted into a binary vector for each bacterial strain. One can use FigFam protein annotations in the Pathosystems Resource Institute Center (PATRIC) database, one of the most comprehensive databases for genomic annotations, including pathogenic strains [527]. When this analysis was performed FigFam contained over 100,000 protein families comprising over 950,000 unique proteins [350]. Binary vectors describing the presence or absence of protein families were used to calculate a distance matrix and compute the persistent homology in this space. Figure 5.35 shows the persistence diagram relating the scale and structure between species. Different species have a far more diverse topological structure in this space than in the MLST space, as well as a wide range of recombination scales. The large scales of exchange in *H. influenzae* suggest it is readily capable of acquiring novel genetic material from quite distantly related strains. It is known that HGT in *H. influenzae* can lead to the acquisition of virulent factors [199]. Furthermore, it has been observed that differences between

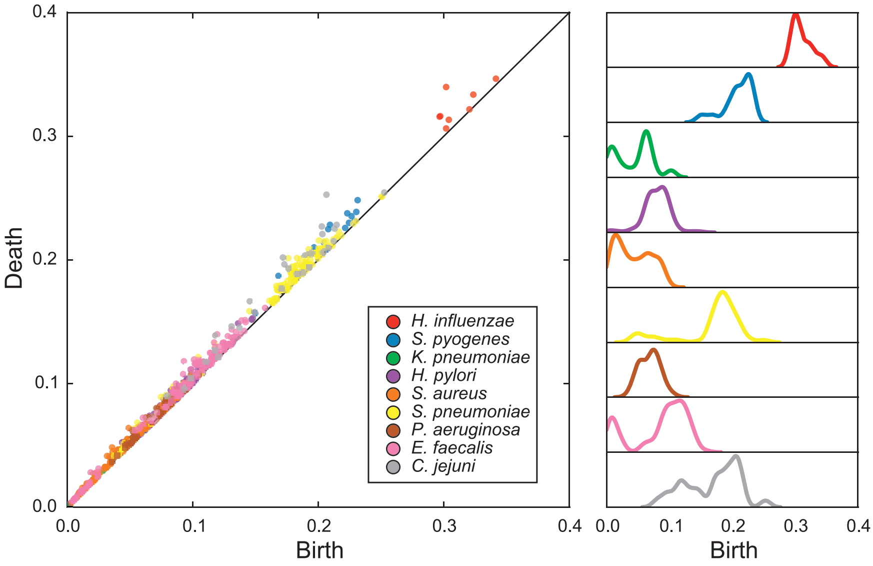


Figure 5.35 Persistence diagram for a subset of pathogenic bacteria, computed using the FigFam annotations compiled in PATRIC. Compared to the MLST persistence diagram, the Figfam diagram has a more diverse scale of topological structure. Source: [161]. Reprinted by permission from Springer Nature: Emmett K. J., Rabadán R. (2014) Characterizing Scales of Genetic Recombination and Antibiotic Resistance in Pathogenic Bacteria Using Topological Data Analysis. In: Ślęzak D., Tan A. H., Peters J. F., Schwabe L. (eds) *Brain Informatics and Health*. BIH 2014. Lecture Notes in Computer Science, vol 8609. Springer, Cham. © Springer International Publishing Switzerland 2014.

H. influenzae strains are more commonly associated to recombination than to point mutations [345].

5.6.5 *Antibiotic Resistance in* Staphylococcus aureus

S. aureus is a gram positive bacterium found commonly in the upper respiratory tract and nostrils. Some strains are capable of causing severe infections in high-risk populations, particularly in a hospital setting. Therefore, the emergence of antibiotic resistant *S. aureus* is a significant clinical concern. Methicillin resistant *S. aureus* (MRSA) strains are resistant to β-lactam antibiotics, which include cephalosporin and penicillin. The gene *mecA*, part of Staphyloccoccal cassette chromosome mec *(SCCmec)*, codes for a dysfunctional penicillin-binding protein 2a (PBP2a), prohibiting the β-lactam primary mechanism and causing resistance [273]. Characterizing the spread of resistance within the *S. aureus* population is clearly of critical clinical import.

To address this question, one can use FigFam annotations in PATRIC, as described in the previous section. PATRIC contains genomic annotations for 461 strains of *S. aureus*, collectively spanning 3578 protein families. One can perform a clustering analysis using Mapper [268]. By selecting as filter function the first two singular values, it can be observed that the resulting graph structure exhibits two main clusters with a thin "bridge" connecting them, as shown in Figure 5.36. These two clusters accord with previous phylogenetic studies which used multilocus sequence data to identify two major population groups [124].

142 of the 461 strains of *S. aureus* in PATRIC carry the *mecA* gene. When we color based on an enrichment for *mecA*, a stronger enrichment can be observed in the cluster on the right (Figure 5.36). This analysis would suggest that β-lactam

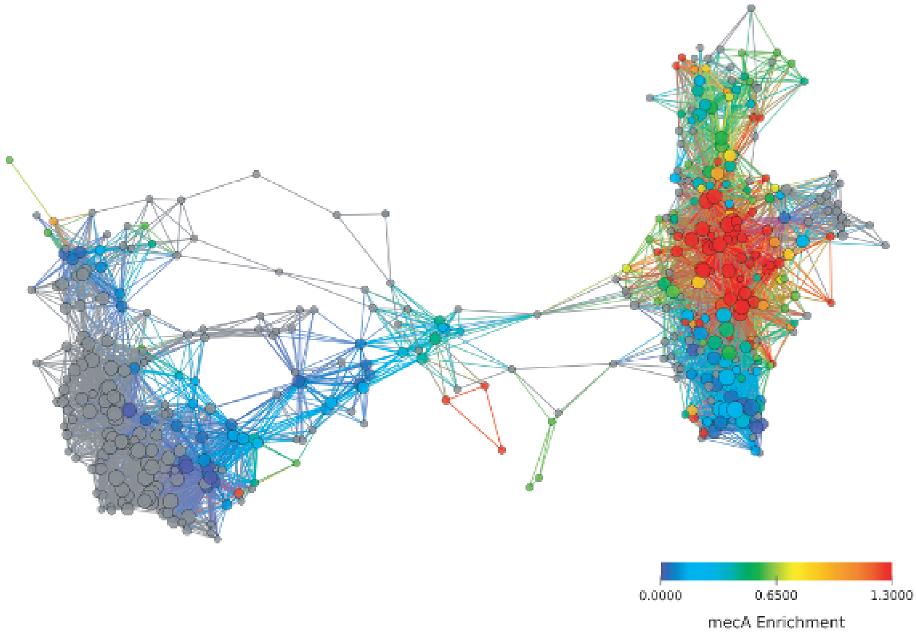


Figure 5.36 The FigFam similarity network of *S. aureus* constructed using Mapper as implemented in Ayasdi Iris. One can use a Hamming distance metric and primary and secondary metric SVD filters (res: 30, gain 4×, eq.). Node color is based on strain enrichment for *mecA*, the gene conferring β-lactam resistance. Two distinct clades of *S. aureus* are visible, one of which already shows significant drug resistance. The growing enrichment for *mecA* in the second clade is clinically worrisome. Source: [161]. Reprinted by permission from Springer Nature: Emmett K. J., Rabadán R. (2014) Characterizing Scales of Genetic Recombination and Antibiotic Resistance in Pathogenic Bacteria Using Topological Data Analysis. In: Ślęzak D., Tan A. H., Peters J. F., Schwabe L. (eds) *Brain Informatics and Health*. BIH 2014. Lecture Notes in Computer Science, vol 8609. Springer, Cham. © Springer International Publishing Switzerland 2014.

resistance has already become dominant in that clade, likely as a result of selective pressures. More strikingly, one observes that while *mecA* enrichment was not as strong in the second cluster, there was a distinct path of enrichment emanating along the connecting bridge between the two clusters and into the less enriched cluster. This suggests the hypothesis that antibiotic resistance has spread from the first cluster into the second cluster via strains intermediate to the two and will likely continue to appear in the second cluster.

5.7 Persistent Homology Estimators in Population Genetics

Mathematical models provide a way of generating data that can be used to tune inference procedures. In population genetics, there are simple models that can simulate the generation of mutations and recombination in populations of genomes. In Appendix B we describe some of the commonly used models of population genetics, including the Wright-Fisher, Moran, and coalescence models. In this section, we will study one of the most popular models, the coalescent model with recombination. With only two parameters, the mutation and recombination rates, one can generate large amounts of simulated data. Using this data, we will construct estimators based on persistent homology.

5.7.1 Coalescent Process

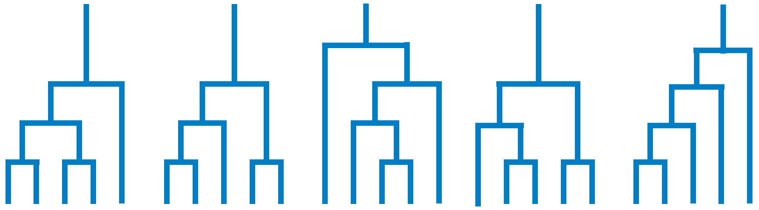
The coalescent process is a stochastic model for generating genealogies, evolutionary histories represented by lines of descent from a common ancestor, for a collection of individuals sampled from an evolving population (see Appendix B). These genealogies can then be used to simulate new, synthetic genetic sequences. Coalescence processes and the attendant coalescent theory underlie many methods commonly used in population genetics.

Starting with a sample of n individuals from a present-day population, each individual's lineage is traced backward in time by randomly choosing a member of the previous generation as the individual's parent. Two individuals may, by chance, be assigned the same parent, in which case their lineages merge. This stochastic process ends when the lineages of all sampled individuals have merged at a single most recent common ancestor.

In this process, if the total population size N is sufficiently large, then the expected time before a coalescence event, in units of $2N$ generations, is approximately exponentially distributed:

$$P(T_k = t) \approx \binom{k}{2} e^{-\binom{k}{2} t},$$

where T_k is the time that it takes for k separate lineages to collapse $k-1$ lineages.



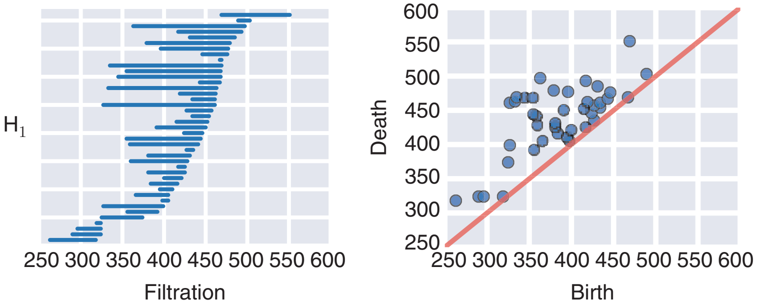


Figure 5.37 Two representations of the same topological invariants, computed using persistent homology. Left: Barcode diagram. Right: Persistence diagram. Data was generated from a coalescent simulation with $n = 100$, $\rho = 72$, and $\theta = 500$. Source: [164]. From Emmett et al., Parametric inference using persistence diagrams: A case study in population genetics, arXiv: 1406.4582 [q-bio.QM].

After generating a genealogy, the genetic sequences of the sample can be simulated by placing mutations on the individual branches of the lineage. The number of mutations on each branch is Poisson distributed with mean $\frac{\theta t}{2}$ where t is the branch length and θ is the population-scaled mutation rate. In this model, the average genetic distance between any two sampled individuals – the number of mutations separating them – is θ.

Coalescence models can be extended to include recombination events, allowing different genetic loci in a sampled individual to come from different lineages within the genealogical structure. Recombination is modeled as a splitting event in which an individual, rather than being a direct descendant of only a single parent, descends from two separate lineages – and occurs at a rate determined by a population-scaled recombination rate ρ. Thus evolutionary histories are no longer represented by a contractible tree, but, due to the combined splitting and joining actions, by an *ancestral recombination graph* which may have loops and other non-trivial, higher-dimensional topology.

5.7.2 Statistical Model

A persistence diagram generated by a coalescence simulation with recombination is shown in Figure 5.37. The information in the diagram can be used to infer the

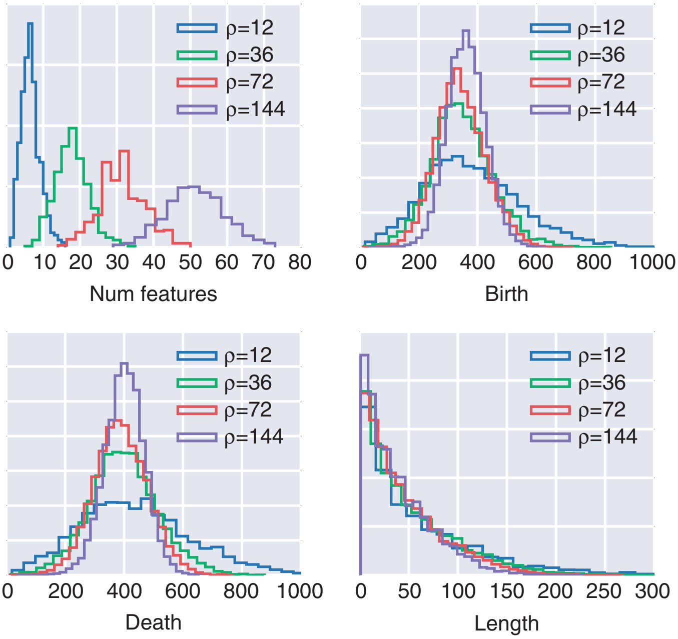


Figure 5.38 Distributions of statistics defined on the H_1 persistence diagram for different model parameters. Top left: Number of features. Top right: Birth time distribution. Bottom left: Death time distribution. Bottom right: Feature length distribution. Data generated from 1000 coalescent simulations with $n = 100$, $\theta = 500$, and variable ρ. Source: [164]. From Emmett et al., Parametric inference using persistence diagrams: A case study in population genetics, arXiv: 1406.4582 [q-bio.QM].

parameters θ and ρ (the mutation and recombination rates, respectively) that generated the data. Here, inference is based only on the detected H_1 invariants, but the idea can be readily generalized to higher dimensions. We consider the following properties of the persistence diagram: the total number of features, K; the set of birth times, $(b_1, \ldots, b_K)$; the set of death times, $(d_1, \ldots, d_K)$; and the set of persistence lengths, $(l_1, \ldots, l_K)$. In Figure 5.38 the distributions of these properties for four values of ρ are shown, keeping fixed $n = 100$ and $\theta = 500$.

It is immediately evident that the number of features K increases with ρ, consistent with the basic intuition that recombination events generate non-trivial topology in the model. The means of the birth and death time distributions depend only very weakly on ρ and are slightly smaller than θ, suggesting θ defines a natural scale in the topological space; however, higher values of ρ dramatically reduce variance of the distributions. Finally, the distribution of persistence lengths is independent of ρ.

Examining Figure 5.38, we can observe that the distribution can be approximated by $K \sim \text{Pois}(\zeta)$, $b_k \sim \text{Gamma}(\alpha, \xi)$, and $l_k \sim \exp(\eta)$. Death time is given by $d_k = b_k + l_k$, which is incomplete gamma distributed. The parameters of each distribution are assumed to be an a priori unknown function of the model parameters, θ and ρ, and the sample size, n. Keeping n fixed, and assuming each other parameter in the diagram is independent (a strong assumption), we can define the full likelihood as

$$p(D \mid \theta, \rho) = p(K \mid \theta, \rho) \prod_{k=1}^{K} p(b_k \mid \theta, \rho) p(l_k \mid \theta, \rho).$$

Simulations over a range of parameter values suggest the following functional forms for the parameters of each distribution. The number of features is Poisson distributed with an expected value

$$\zeta = a_0 \log\left(1 + \frac{\rho}{a_1 + a_2\rho}\right).$$

Birth times are gamma distributed with shape parameter

$$\alpha = b_0\rho + b_1$$

and scale parameter

$$\xi = \frac{1}{\alpha}(c_0 \exp(-c_1\rho) + c_2).$$

These expressions appears to hold well in the regime $\rho < \theta$, but break down for large ρ. The length distribution is exponentially distributed with shape parameter proportional to mutation rate, $\eta = \alpha\theta$. The coefficients in each of these functions are calibrated using simulations, and could be improved with further analysis. This model has a simple structure and standard maximum likelihood approaches can be used to find optimal values of θ and ρ.

5.7.3 Coalescent Simulations

We describe results associated to the simulation of a coalescent process with sample size $n = 100$ and $l = 10{,}000$ loci. The mutation rate, θ, was varied across $\theta = \{50, 500, 5000\}$. The recombination rate, ρ, was varied across $\rho = \{4, 12, 36, 72\}$. The output of the process is a set of binary sequences of variable length (the length is dependent on θ). The Hamming metric yields a pairwise distance matrix between sequences. Computing persistent homology and using the model described in Section 5.7.2 produces estimates of θ and ρ. Results are shown in Figure 5.39, where we plot estimates and 95% confidence intervals from 500 simulations. We observe an improved ρ estimate at higher mutation rate. This is expected, as increasing θ is essentially increasing sampling on branches in the genealogy. We also observe tighter confidence intervals at higher recombination rates, consistent with the

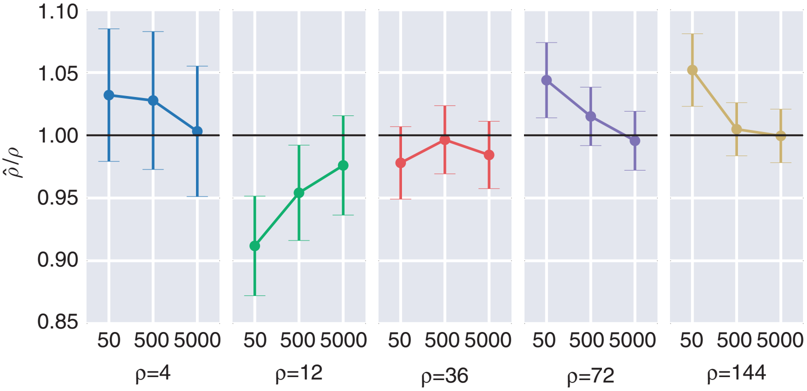


Figure 5.39 Inference of recombination rate ρ using topological information. The recombination rate ρ is estimated for five values {4, 12, 36, 72, 144} at three different mutation rates {50, 500, 5000}. Mean estimates over 500 simulations and 95% confidence interval are shown. Source: [164]. From Emmett et al., Parametric inference using persistence diagrams: A case study in population genetics, arXiv: 1406.4582 [q-bio.QM].

behavior seen in Figure 5.38. See [259] for follow-up work on estimating recombination rates for coalescent models and further discussion of the relationship between topological invariants and population genetics.

5.8 Recombination Landscape in Humans

Sexual reproduction is a non-tree-like event, essential to ensuring genetic diversity of offspring and preserving genome integrity. Cells in sexually reproducing organisms contain two copies of most chromosomes (autosomes). Each copy differs slightly in sequence, but has the same overall structure. Humans have 22 pairs of chromosomes, as well as sex chromosomes – a pair of X chromosomes for females and an X and Y chromosome for males. Each of these 23 pairs of chromosomes is inherited from a different parent. In the process of meiosis, cells become haploid, i.e., containing only one chromosome of each pair, with different regions randomly selected from the paternal or maternal copy.

Meiosis occurs through two rounds of division. In division I of meiosis, a diploid cell containing a paternal and maternal copy of each chromosome duplicates (see Figure 5.40). Homologous chromosomes are then paired in a structure that is called a *bivalent*. This is where the process of recombination takes place. In a nutshell, meiotic recombination begins with a double-strand break in one of the parental chromosomes catalyzed by a particular protein, Spo11, and the broken strands from this chromosome are partially degraded. To repair this chromosome, the intact

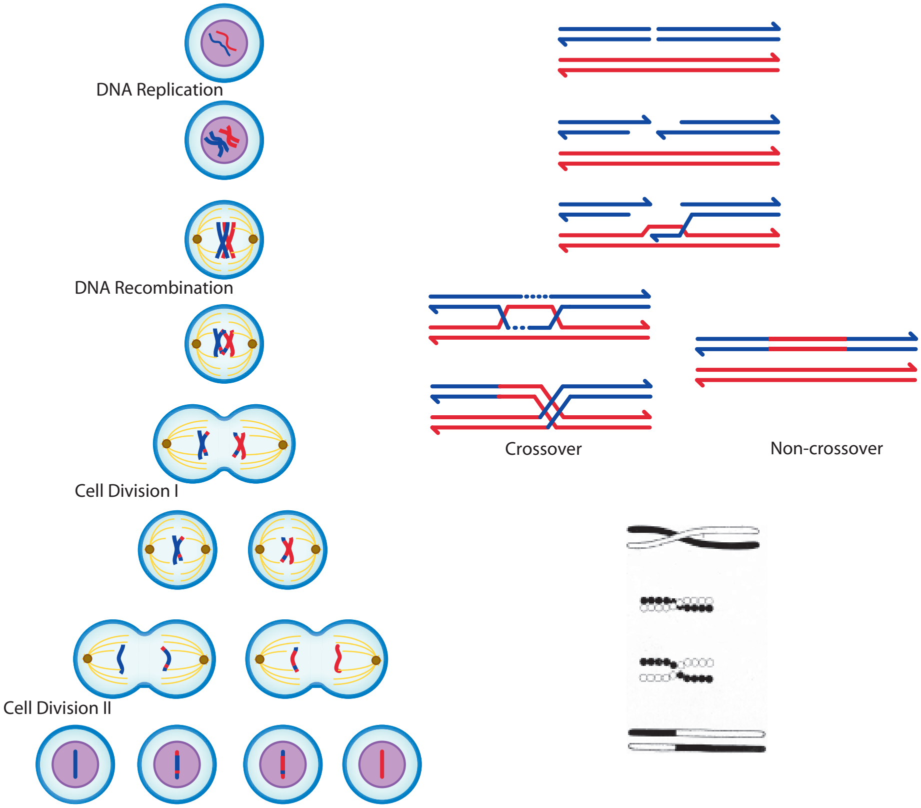


Figure 5.40 Left: Process of meiosis through two rounds of division. Top right: Cartoon of recombination. Bottom right: Illustration from the 1916 book of Morgan explaining crossovers. Source: [359].

chromosome strands are used as templates. The final recombination product could result in crossover resulting in a new chromosome generated from both parental chromosomes. Also it could lead to a non-crossover event (associated to gene conversion, or partial replacement of a DNA region by a homologous sequence), where part of the genomic material from one parent is used in the other strand. Finally, the cell divides and two of the homologues are then contained in each daughter cell. In division II of meiosis, cells divide again without further duplication of the genomic material. At the end, there are four haploid cells derived from the initial diploid cell. Each of the cells contains genomic material from the paternal, the maternal, or a recombinant of both.

Given that meiotic recombination is such a fundamental process in eukaryotic evolution, involving break and repair of genomic material, it is not surprising that it is a highly regulated process. Since the work of Morgan using the fruit fly, *Drosophila melanogaster*, as a model, we have a quantitative understanding of how often chromosomal crossovers occur in meiosis (bottom right panel in Figure 5.40, obtained from [359]). Morgan was able to establish a link between the probability of crossover and how far away in chromosomal position two different loci were. In humans, recombinations occur at an average rate of one crossover per chromosome per generation. A more quantitative measure of these rates can be obtained by estimating the probability that a crossover event will occur between two different loci in a chromosome. One defines a *centi-Morgan* (cM) distance in genomic position with a 1% chance of recombination per generation. The average rate of recombination in humans is about 1 cM per megabase.

However, genetic versus chromosomal distance approaches do not allow a high-resolution mapping below millions of bases, as it will require many generations to track many meiotic events. Pedigree and linkage disequilibrium analysis provide a much more refined view of where recombination occurs [288]. Pedigree analysis studies families of related individuals along several generations. Linkage disequilibrium (LD) is a measure of how the variability of two genomic loci is associated. If there is no recombination between loci, two mutations in the same chromosome will be always traveling together. If recombination occurs very frequently, the presence of a particular allele provides very little information about nearby mutations. The simplest measure of LD is $D_{ij} = f_{ij} - f_i f_j$, where f_{ij} is the frequency of observing two alleles i and j together, and f_i is the frequency of observing the allele i.

It has been found that recombination occurs preferentially at narrow genomic regions known as recombination hotspots [18, 37, 396]. In mammals, recombination hotspots are specified by binding sites of the meiosis-specific H3K4 trimethyltransferase PRDM9 [38, 372, 400]. However other factors play a role too. The recombination landscape in eukaryotes is actually the result of a hierarchical combination of factors that operate at different genomic scales. High-resolution mapping of meiotic double-strand breaks (DSBs) in yeast and mice [181, 294, 397, 466] reveals fine-scale variation in recombination rates within hotspots as well as frequent recombination events occurring outside hotspots [397].

Population-based recombination maps are a valuable tool in the study of recombination in humans [344, 371]. Due to the number of genomes published by such consortia as the 1,000 Genomes Project [122] and ENCODE [123], it is now possible to produce exquisitely fine-scale mapping and annotation of human recombination. Chromatin immunoprecipitation (ChIP-seq), bisulfite, or RNA sequencing methods, as well as other high-resolution data sets reveal a wide variety of distinct biological features associated with small genomic regions. These can aid in connecting locations where recombination occurs with other molecular and biological phenomena.

Establishing compelling statistical associations with such narrow and often clustered biological features, and analyzing the very large numbers of sequences in these data sets, is becoming a crucial challenge for traditional methods of recombination rate estimation (such as methods based on linkage disequilibrium). Robust and scalable methods to detect and quantify rates of recombination at different scales are particularly useful.

5.8.1 Fine-Scale Resolution of Human Recombination

The persistent homology estimators of recombination introduced in the previous section can be easily implemented on a sliding window and therefore adapted to the very long eukaryotic genomes [87, 88]. The sliding window cuts the genome into small overlapping segments. One can estimate the local recombination rate $\rho(x)$ using the persistent homology estimators of recombination rates in a sliding window centered around a genomic position x. There are different implementations of the procedure that determine the size of the window. A constant window size may not be desirable, because mutation rates can vary over different genomic regions; and thus one might get windows that contain no polymorphic sites, in which no recombination could be detected. Choosing variable window sizes to fix the number of polymorphic sites per window avoids this problem. The number of polymorphic sites per window defines the genomic scale at which recombination is observed.

Figure 5.41 captures a snapshot of the sliding window near the cytogenic band 1q24.1 of human Chromosome 1 [87]. We describe estimates of recombination rates from 647 individuals genotyped for the 1,000 Genomes Project [122], with windows containing 14 polymorphic sites. The sliding window approach assigns a finite metric space for each position x, containing 647 points, one for each individual. The one dimensional homology estimator of $\rho(x)$ based on b_1 on a sliding window allows inference of local recombination rates at x. Recombination maps reflect a landscape with peaks showing recombination hotspots and valleys showing low recombination regions.

The 1,000 Genomes Project provides genotype data of nearly 38 million single nucleotide polymorphisms (SNPs). The data is phased, meaning the sequences' locations include the specific chromatid on which they were found. The individuals genotyped by the 1,000 Genomes Project came from seven different populations: European-American, Han Chinese, Finnish, British, Japanese, Tuscan and Luhya (a Bantu ethnic group in Kenya). Each of these populations has a different recombination map that can be compared to the others. The median effective population recombination rates detected for non-African populations had $\rho \sim 0.6$ per kbp. The effective population recombination rates were substantially higher in the African population, consistent with its larger effective population size and supporting the out-of-Africa human expansion model [494]. While the recombination maps agree at a global scale, there are population-specific variations. In particular, the Luhya

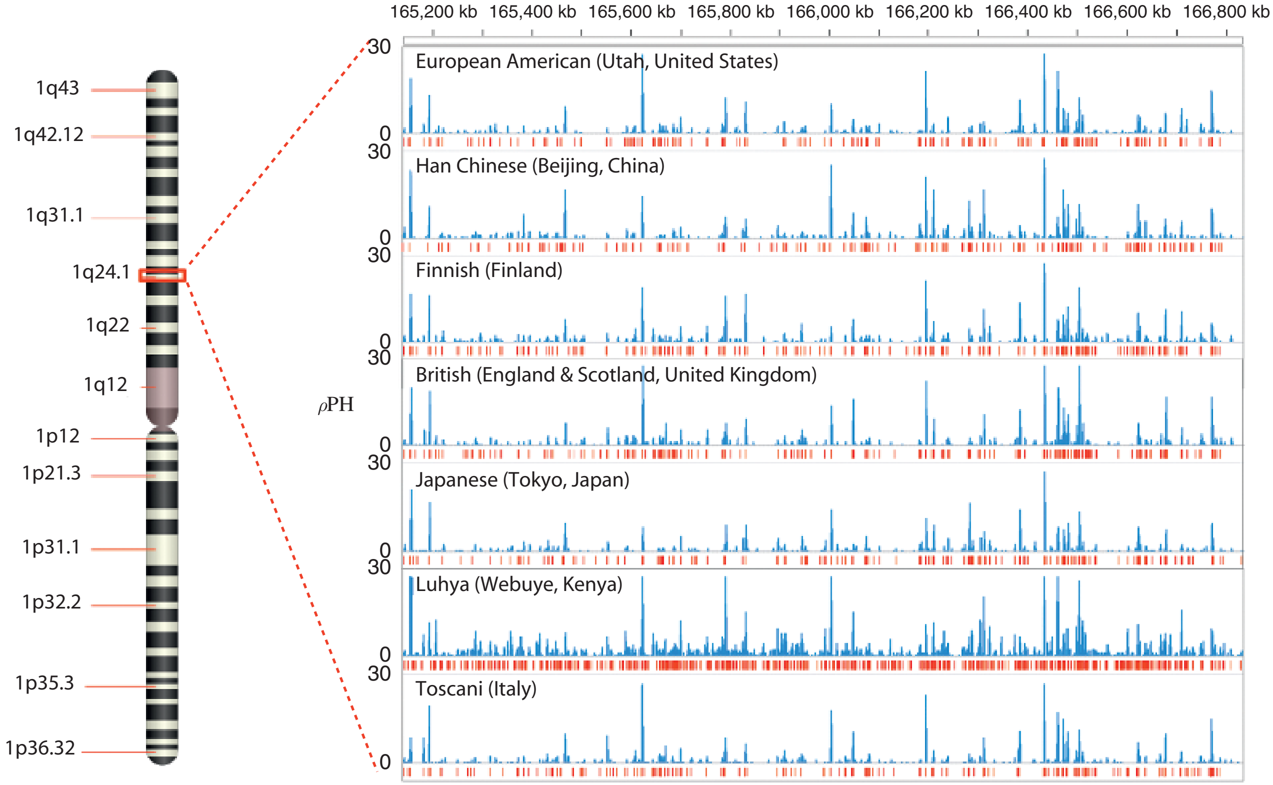


Figure 5.41 Positionwise recombination rate estimates across seven distant human populations for the cytogenic band 1q24.1. Blue bars represent recombination rates estimated with a variable-size sliding window containing 14 segregating sites. Below, red segments represent genomic regions where a 500 bp window detects recombination ($b_1 > 0$). Source: [87]. From Pablo G. Cámara et al., 'Topological data analysis generates high-resolution, genome-wide maps of human recombination', Cell Systems 3.1 (2016): 83-94. Reprinted with permission from Elsevier.

present a more unique recombination landscape. Topological methods can be used to describe gene flow across populations, where migration and admixture appear as high-dimensional loops in the evolutionary space.

The fine-resolution maps of recombination connect population maps to specific genomic locations that can inform us about specific molecular processes associated to recombination hotspots. Recently, high-throughput methods have catalogued binding sites of different proteins, epigenetic marks, and gene expression across genomes [123]. For instance, one can ask what proteins bind to the genome in locations where recombination occurs. The persistent homology recombination landscape recapitulates known proteins and epigenetic marks associated to recombination, such as the meiosis-specific histone 3 lysine 4 (H3K4) trimethyltransferase PRDM9, CpG hypomethylation and H3K4 trimethylation. Comparing persistent homology estimators of recombination to binding sites from ChIP-seq data of 118 transcription factors, these binding sites are depleted in high recombination loci on average (see Figure 5.42) [87]. In addition, this analysis led to the discovery of previously unreported transcription factors associated to recombination regions, such as members of the E2F protein family, important regulators of cell cycle progression and differentiation. These proteins bind to sites of RNA polymerase II and different regulatory subunits of the MLL/MLL1 protein complex [87].

5.9 Gene Trees and Species Trees

In the previous sections, we have described how different genomes from individual organisms are related. Sometimes, like in clonal processes, the relation between different genomes can be well represented by a phylogenetic tree. However, phylogenetic trees have been traditionally used to capture relations beyond individual organisms, describing relationships between different taxa (e.g., kingdoms, genera, or species). For instance, when Darwin in 1859 proposed his model for the origin of species he had in mind a branching process with different species as leaves (panel A in Figure 5.43). We have to remember that while genomes are ascribed to individual organisms, individuals within a species have slightly different genomes. In a strict sense, the genome of a species, such as the human genome, does not exist, only the related genomes of organisms within a species. If there is no single genome for a species, how can one construct a *species tree*, a tree where leaves are labeled by species? More formally, given a set of genomes G from organisms belonging to a set of species S, how can we find a representation between different species (or other taxa)? When does it make sense to talk about a species tree?

The construction of a species trees is not straightforward, as there are not only technical questions but also profound conceptual obstacles to this enterprise. First,

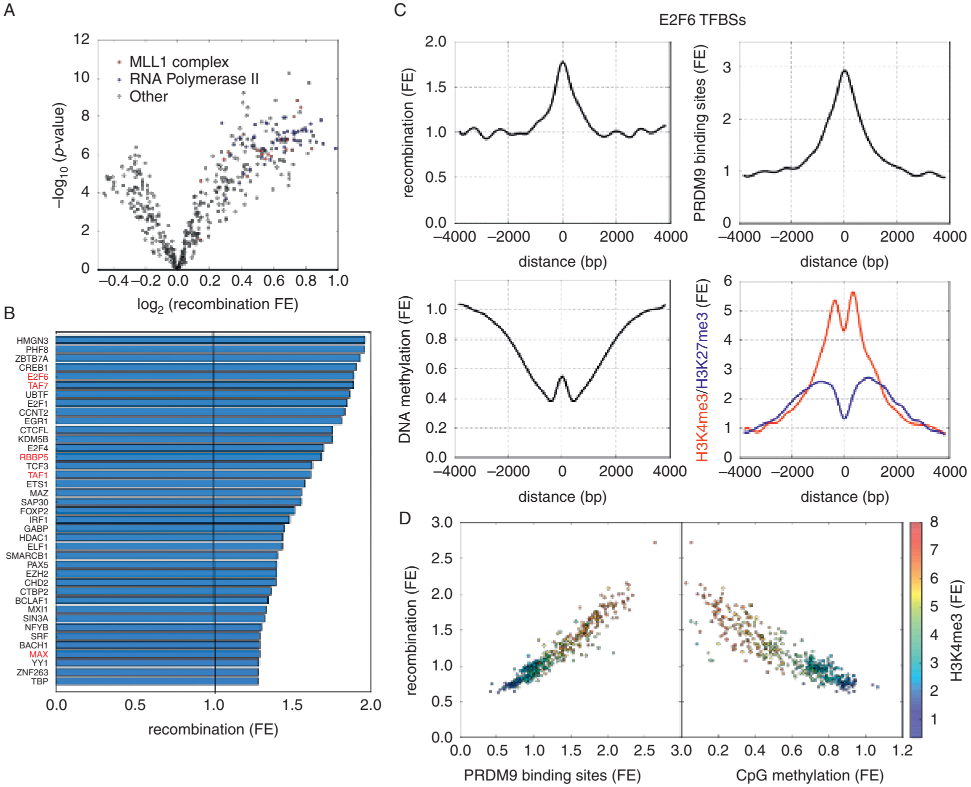


Figure 5.42 Relation of recombination to TFBSs (transcription factor binding sites). (A) Logarithm in base 10 of p-value against logarithm in base 2 of the recombination enrichment at TFBS, for each TF and cell line. RNA Polymerase II and TFs that form part of MLL complexes are indicated explicitly. Enrichments are taken with respect to neighboring genomic regions. (B) Recombination enrichment at TFBSs for each TF with respect to the whole-genome average. Only TFs with the highest enrichments are shown. TFs that may form part of MLL complex are indicated in red. (C) Recombination, predicted PRDM9 binding sites, CpG methylation, H3K4me3 and H3K27me3 enrichments at binding sites of E2F6. (D) Recombination enrichment against enrichment for PRDM9 binding sites (left) and sperm CpG methylation (right). Color scale represents H3K4me3 enrichment. Source: [87]. From Pablo G. Cámara et al., ‘Topological data analysis generates high-resolution, genome-wide maps of human recombination’, Cell Systems 3.1 (2016): 83–94. Reprinted with permission from Elsevier.

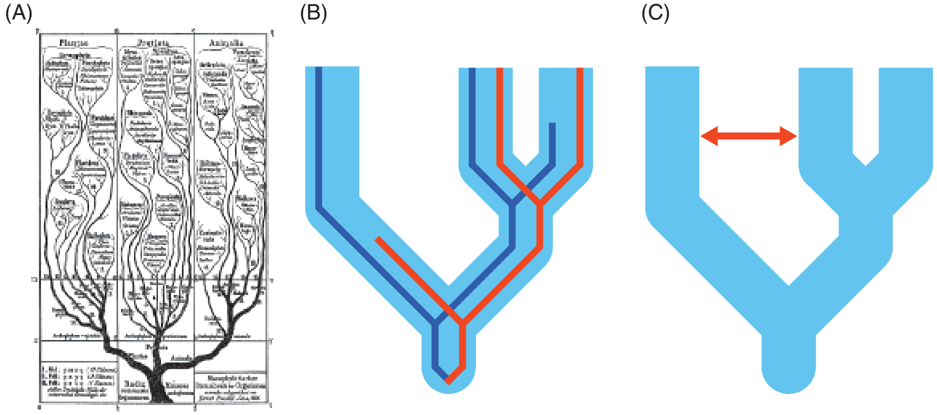


Figure 5.43 Phylogenetic trees have been traditionally used to describe the relation between species. (A) Tree of life by Haeckel, Generelle Morphologie der Organismen (1866) with species and higher taxa assigned to branches and leaves. In a strict sense, there is no genome of a species or any higher taxa, and different genomes and different genomic regions could generate (slightly) different trees. Incomplete lineage sorting is a common phenomenon that generates different tree topologies. Source: From E. Haeckel, Generelle morphologie der organismen. Allgemeine grundzüge der organischen formen-wissenschaft, mechanisch begründet durch die von Charles Darwin reformirte descendenztheorie, Berlin, G. Reimer, 1866. (B) and (C) Incomplete gene sorting can occur if a locus in an ancestral species is polymorphic (has more than two alleles). Suppose that it divides first into two lineages and then one of those further divides into another two. The alleles could then be fixed differently in each of the lineages. Incomplete lineage sorting generates "gene trees" (trees from the allele) that present a different tree topology than the species tree. These tree incompatibilities are represented in a "fat tree" that can capture different topologies such as the ones occurring during incomplete gene sorting (B) or represented by arrows representing horizontal gene transfer events (C).

and the most serious obstacle, relates to the assignment of a given genome to a particular species. Even in metazoa, where phenotypic differences are significant, it is sometimes unclear how to define species. The problem becomes acute in bacteria and viruses. From an empirical genomic point of view, given genomic data from two organisms, how can we determine whether they belong to the same species? This assignment problem is linked to the definition of species. A species is commonly and informally understood as the largest group of interbreeding individuals capable of producing fertile offspring. However, there is no consensus on a precise species definition that can incorporate sexual and asexual organisms, that can consider more complex phenomena such as ring species (populations that can breed with nearby populations but not those far apart), that can consider hybrids, among others. More importantly, from a pragmatic genomic point of view, it is not unusual to sequence the genome of an organism without knowing the mating possibilities with others. Genomic based definitions of species are based on genomic data, for instance, based on arbitrary cutoff values for species definitions, e.g., a 95% average nucleotide identity as a potential criterion to define whether two bacteria belong to the same species. In viruses, the problem is even more acute. According to the International Committee on Taxonomy of Viruses (ICTV) a virus species is a "polythetic class of viruses that constitute a replicating lineage and occupy a particular ecological niche." A "polythetic class" means a group of organisms with several properties in common but not necessarily a single defining property. Thus, in a way, the very definition of viral species is artificial.

Even if we have a good species assignment, a serious second obstacle appears. We have seen in the previous sections of this chapter that a variety of biological processes (reassortments, recombinations, horizontal gene transfers, etc.) generate genomic relations that are not well captured by trees. When these processes occur between members of different species, there will not be a tree representation. If exchange of genomic material is rare one could envision a tree with small corrections reflecting non-tree-like processes (Figure 5.43). But even if non-tree-like processes did not occur between members of different species, the history of a particular genomic region could be different from the history of another region, a phenomenon that is referred to as *incomplete lineage sorting*.

Incomplete lineage sorting could happen when lineages divide before polymorphisms fix in the population. If variant alleles are kept in the different descendant populations and fix independently in each population, the final tree of these alleles could be different from the species tree (Figure 5.43B). These tree incompatibilities are informally represented in a "fat tree" that reflects the tree topological ambiguity due to incomplete lineage sorting. The problems of defining species, of assigning species and of finding good *species summarizations* (tree or others to be defined) are linked. If organisms across different species frequently exchange genomic material it is difficult to establish clear species boundaries.

Sometimes, under the assumption that there are some smaller genomic regions where trees are good approximations, one can consider the problem of reconstructing a *species tree* as finding a good "summarization" of a set of trees. In the literature, this problem is referred to as finding the species tree from a set of smaller genomic region trees, called *gene trees*, although the latter do not necessarily refer to genes and could refer to other genomic regions. In this context, Billera, Holmes, and Vogtmann [55] applied the structure of the CAT(0) spaces described in Section 4.7. Here, the *species tree* problem is posed as the problem of finding the centroid of a set of *trees* of 12 primates, including *Homo sapiens*.

5.10 Extensions: Median Complex and Topological Minimal Graphs

Thus far, our approach has been to start with genomic data, compute a finite metric space, and relate the topological properties of the metric space to biological phenomena, such as reassortment, recombination, or horizontal gene transfer. We have constructed estimators of recombination rate based on statistical properties of persistent homology summaries. We have also deconstructed large genomes using a sliding window. Beyond these methods, there exist other constructions that can be useful. In particular, starting from the genomic data, we can increase the number of genomes by adding inferred genomes that could be associated to potential ancestors. From this extended data set, one can again define a finite metric space, whose topology could increase the sensitivity for recombination detection, at the expense of increasing the complexity and introducing spurious non-interpretable events in a biological context.

Let us consider a few simple examples for which the four gamete test (the presence of all four different alleles in two loci) indicates a non-tree-like event, and how persistent homology can or cannot detect the reticulate event.

Our first set of examples [162] consists of four genomes of length two and two bases represented by 0 and 1: $s_1 = 00$, $s_2 = 10$, $s_3 = 01$, and $s_4 = 11$. One can easily verify that the four gamete test finds an incompatibility between the first two sites, as the four possible gametes are present (Figure 5.8). This event is precluded in an infinite-sites model without recombination, where mutations in the same site are rare. Persistent homology captures this event, as it identifies a bar $[1, 2)$ in the first homology persistent group using Hamming distance. We can vary this simple example by taking four genomes of length three and two bases represented by 0 and 1: $s_1 = 000$, $s_2 = 100$, $s_3 = 010$, and $s_4 = 111$. One can again easily verify that the four gamete test finds incompatibility between the first and second sites. However, the barcode in dimension one (or higher) does not show any bar. In this simple example, it is easy to identify the reason: if s_1 is the common ancestor to other sequences, s_2 and s_3 can be considered to be the parents of s_4, a direct descendant of a reticulate event. In this case, it is easy to infer that there was an ancestral recombinant sequence, $s_r = 110$, which was not sampled in our data set (Figure 5.44A). If this missing sequence was present in our data set, we would have recovered a bar in the first homology group representing the event. This simple case shows that incomplete sampling of the process can significantly lower the sensitivity of persistent homology to detect potential recombinant events.

Another example can be found in the article by Song and Hein [476]. In this case there are five genomes with four sites: $s_1 = 0000$, $s_2 = 1100$, $s_3 = 0011$, $s_4 = 1010$, and $s_5 = 1111$. There are multiple incompatibilities between sites that can be easily identified using the four gamete test (1 and 3, 1 and 4, 2 and 3, and

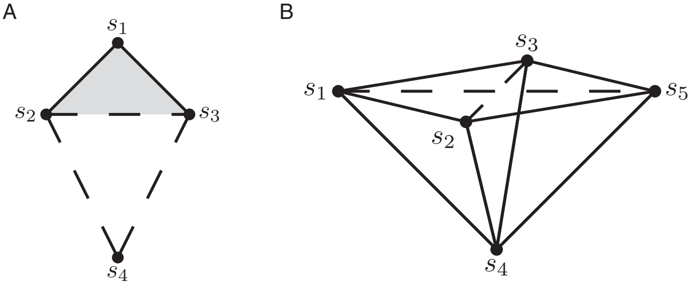


Figure 5.44 Two simple cases where the Vietoris-Rips complex applied to a distance matrix between sampled genomes fails to identify a potential recombination. (A) In this example an ancestral sequence has not been considered in the sample. If considered, that recombination is identified. (B) In more complex cases, multiple recombinations can lead to a degeneracy. Source: [162]. Reprinted with permission: © EAI European Alliance for Innovation 2016.

2 and 4). Song and Hein [476] show that at least two recombinations are needed to explain this data. When applying persistent homology using the Vietoris-Rips complex with Hamming distance as the metric, one finds that the barcode diagram does not show any event in dimension larger than zero, failing to capture potential recombinations. Close inspection shows that this is a special case, where s_4 sits at the same distance from the other four sequences. If other ancestral states could have been sampled or if s_4 had not been present in the data set, persistent homology could have detected the reticulations.

These examples show that persistent homology using Vietoris-Rips complexes applied to genetic distances is limited as a method to identify potential reticulate events. This is not surprising, as sequence data is much richer than a distance matrix; as in standard phylogenetic approaches, methods based only on distances constitute a first approximation.

Working directly with sequences provides a much more powerful data structure that can capture all potential reticulations. For instance, from our original sequence data, we can construct many distance matrices by subsampling sets of sites. If the phylogeny is truly tree-like and the infinite-sites model holds, none of these subsamples generates any non-tree-like structure, and the persistent homology barcodes for each of the subsamples should be empty for dimension bigger than zero. If the four gamete test is satisfied for a particular subset of data, the four alleles should be present and so a bar in dimension one. Subsampling sites provides a very powerful tool to increase sensitivity at heavy computational expense, as the number of potential subsets is exponential with the number of sites. There is also the problem of interpretability of the results: can we infer what could have been the recombination history, or just the minimal number of recombination events, from subsamples of data? There are several alternatives though that take advantage of the fact that homologous recombination occurs between nearby sites in the genome, or in other words, the four gamete test is more informative between nearby sites in the genome. This type of information is fundamental for interpretability and is used in most standard tests of recombination, for instance, the Hudson-Kaplan test [255].

In what follows we propose several approaches to increase the sensitivity and interpretability of persistent homology for the identification of reticulate events. The first approach, the median complex, is based on the idea of adding potential ancestral states. The second infers associated graphs, named topological minimal graphs, as explicit representations of potential histories.

5.10.1 The Median Complex Construction

In order to increase the sensitivity of persistent homology methods for identification of recombination, we apply an old insight in the field, namely that adding information about ancestral inferred states can make it easier to identify potential recombinant events. The idea is to add extra points to the original data, some of which can be mapped to potential ancestral states. The *median graph* (also called the Buneman graph) is a graph constructed with inferred median points. It was introduced as a way of capturing all maximum parsimony evolutionary trees [81] and has been the object of study for phylogenetic network inference [30, 31].

Associated to the median graph is a collection of filtered complexes referred to as the *median complex*. The persistent homology of the median complex can be computed by considering from the finite metric space consisting of the original data plus the new points imputed from the median procedure. If the original data is tree-like, there is no persistent homology information in the median complex above dimension zero; this is consistent with the persistent homology of the data [100]. But high-dimensional classes in the median complex can capture recombination events not visible in the persistent homology of the underlying complex. The major drawback to the use of median graphs is the large number of imputed points, which complicate computation and obscure the biological interpretation.

The *median sequence* $m(a, b, c)$ of three binary sequences a, b, and c is a sequence with the majority consensus at each site. For instance, take the example shown in Figure 5.46 with three sequences with three sites each: $a = 000$, $b = 110$, and $c = 011$. The median of the first site is a 0, in the second a 1 and in the third a 0, so the median sequence is $m = 010$, different from any a, b, and c. Note that the median sequence is not sensitive to sites that are specific to one of the original sequences.

Notice that in the example previously shown in panel A of Figure 5.44, the median of $s_2 = 100$, $s_3 = 010$, and $s_4 = 111$ is precisely the missing sequence

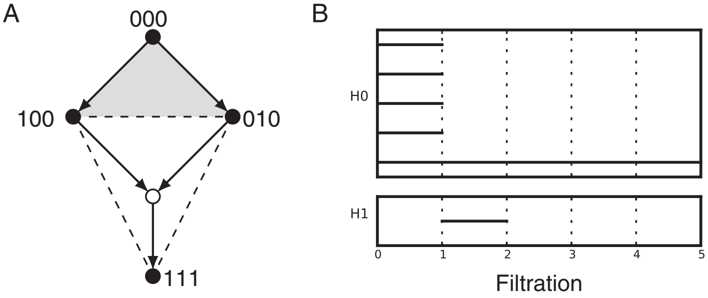


Figure 5.45 One can infer some interesting sequences by applying the median operation. The newly inferred median node (in white) can be interpreted as the missing recombinant between s_2 and s_3 and the ancestor of s_4. Adding the median sequence to the original set one can identify a new one dimensional persistent class in the interval $[1, 2)$. The median operation does not generate other new sequences. Source: [162]. Reprinted with permission: © EAI European Alliance for Innovation 2016.

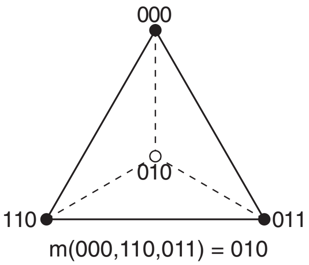


Figure 5.46 The median sequence is constructed for triples of sequences by taking the most common allele at each site. The process can be iterated, adding more median sequences to the original data until no new sequences can be added by this procedure (the median closure). Source: [162]. Reprinted with permission: © EAI European Alliance for Innovation 2016.

$s_r = 110$. Applying the median operation to subsets of size 3 of the augmented set of sequences does not generate any new sequences (Figure 5.45). Computing the persistent homology of the Vietoris-Rips filtration of the original set of sequences together with s_r uncovers the missing one dimensional loop in the interval $\epsilon = [1, 2)$ generated by s_1, s_2, s_3, and the newly reconstructed s_r.

For every triple of sequences one can define the median. The median of a triple may or may not be in the original set. This procedure can be repeated by adding the new median sequences to our original set of sequences S and iterating until there are no more new sequences added. The final set of original sequences and their medians, and successive medians, and so forth, constitute the *median closure* $\bar{S}$:

$$\bar{S} = \{v \,|\, v = m(a, b, c) \in \bar{S} \; \forall \, a, b, c \in \bar{S}\}.$$

The median closure is closed under the median operation.

Instead of constructing the Vietoris-Rips complex using the distances from the original set of sequences, one can construct the complexes using the median closure $\bar{S}$. We will refer to the resulting Vietoris-Rips filtration as the median complex.

Let us revisit our second example (Figure 5.44B). The median operation adds four new sequences, as displayed in Figure 5.47. Now persistent homology applied to the median closure identifies four one dimensional persistent intervals in the barcode diagram, all in the interval $\epsilon = [1, 2)$. In this example the minimal number of recombinations that is needed to explain the data is two, as found by Song and Hein [476]. The number of intervals found in persistent homology is now higher than the minimal number of recombinations. As a consequence, the interpretation of these homology classes as potential recombination is a central problem with the use of the median constructions.

Observe that we can compute the Vietoris-Rips complexes of two finite metric spaces: the one we previously explored with the original data (called here the leaf complex) and the median complex. If the original data is derived from a tree-like structure, the two complexes will provide no high-dimensional persistent homology. Counting bars in PH_1 for the leaf complex frequently underestimates reticulate evolution because of incomplete sampling, while counting bars in PH_1 for the median complex usually overestimates reticulate events. The median complex is in some sense an upper bound on probable recombination histories; although it does not contain within it all possible recombination graphs, as there are infinitely many

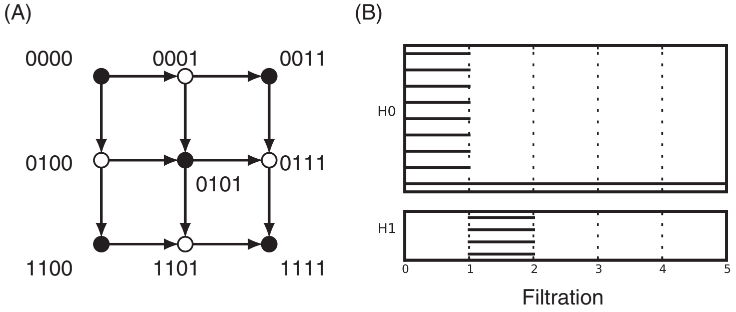


Figure 5.47 The median complex increases the power to identify potential reticulations, but complicates interpretability. When applying the median operation to the example of Song and Hein, one finds four median vertices (in white). Persistent homology identifies four one dimensional loops in this case. Song and Hein found that this data could be explained by a minimum of two recombinations. Source: [162]. Reprinted with permission: © EAI European Alliance for Innovation 2016.

complicated ancestral recombinant graphs (ARGs, see Section 5.10.2), it contains within it all maximum parsimony trees.

We now illustrate the use of the median complex via some examples using real data. In the following examples, the Vietoris-Rips complexes built from the distance matrices of the original data do not show any non-trivial homology. However, it is easy to verify using the four gamete test that there are potential recombinations. The median closure in these examples adds new median sequences that enrich the original data, increasing the power of homology to identify potential recombinations.

The first example is a classic data set in population genetics from Kreitman [309] of eleven sequences from the alcohol dehydrogenase (Adh) locus of the fruit fly *Drosophila melanogaster*. The original eleven sequences consist of 43 polymorphic sites. The median closure adds more than 30 median sequences to the original data set (see Figure 5.48). While persistent homology on the original data set of 9 sequences fails to identify any higher dimensional homology, the median complex identifies 32 bars in dimension one homology and 3 in dimension three.

Hybridization, the process of generating new species by the genetic mixing of two different species, is very common in plants. Common plants, such as wheat, are the result of hybridization and artificial hybrids are very commonly used in crops. Huber and colleagues [250] collected data from the maturase gene (*matK*) in nine species from the genus *Ranunculus*. The median closure added 23 new median vertices to the original data (Figure 5.49). Persistent homology on the original data does not identify any non-trivial class, but the median complex shows 17 one-dimensional and 3 three-dimensional classes.

5.10.2 Topological Minimal Graphs and Barcode Ensembles

In the last two decades, there have been many efforts to produce frameworks to represent non-tree-like events. Phylogenetic networks try to represent inconsistencies among trees as graphs [30, 31, 32, 33, 260, 261, 262, 263, 362]; however, the biological interpretation of these networks is often unclear. Other constructions, sometimes referred to as *explicit networks*, try to provide potential reconstruction of past events that lead to tree inconsistencies. The ancestral recombination graphs, or ARGs, provide a potential historical explanation in terms of mutation and recombination events. Mutations appear as events along the branches of the graph and recombinations appear as merges between parental branches. ARGs do not consider homoplasies due to convergent evolution [208, 209, 220].

In principle, there is an infinite number of ARGs that are consistent with the data. Population genetics models, like coalescence with recombination, can assign probabilities to them [209, 253]. Finding the minimal ARG, i.e., an ARG with the

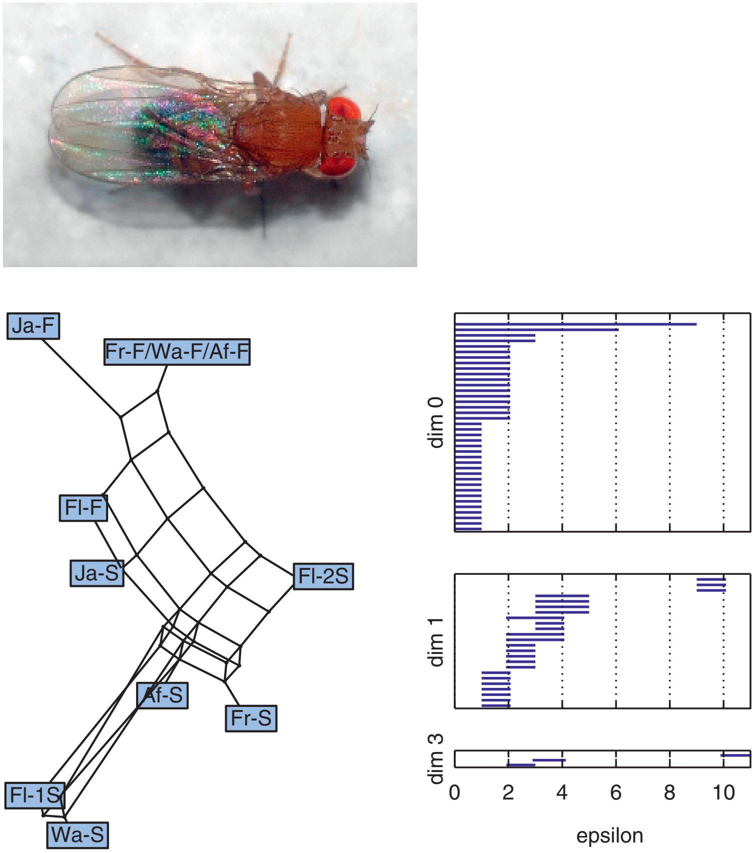


Figure 5.48 Left: The alcohol dehydrogenase (Adh) locus of the fruit fly *Drosophila melanogaster* provides a well studied set of sequences with recombination. Source: Reprinted with permission from André Karwath under the Creative Commons Attribution-Share Alike 2.5 Generic license. Right: Persistent homology on the median closure of the original data identifies one- and three-dimensional homology structures due to recombination events in the population. Source: [162]. Reprinted with permission: © EAI European Alliance for Innovation 2016.

minimal number of mutations and recombinations, is an extremely computationally intensive task. Indeed, finding a minimal ARG has been shown to be an NP-hard problem [65, 66, 522], and an infeasible approach to large data sets. There are, however, several approaches that can approximate minimal ARGs, including heuristic methods [356], branch and bound [477], galled trees [219, 221], and sequentially Markov coalescent approaches [421].

Here, we will present another approximation called a *topological ARG* or *tARG* [88], which is closely related. These capture ensembles of minimal recombination

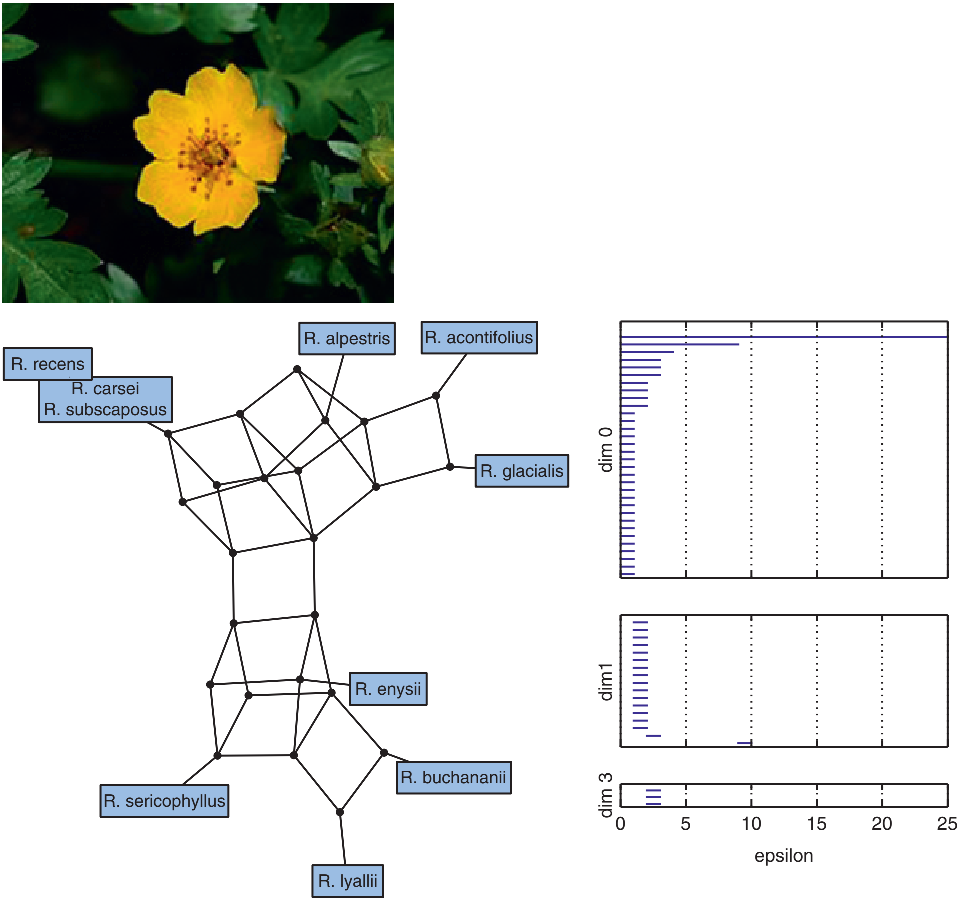


Figure 5.49 Hybridizations are commons in plants. Reprinted with permission from Walter Siegmund under the terms of the GNU Free Documentation License, Version 1.2 or any later version published by the Free Software Foundation. Left: The median closure of data collected by Huber and colleagues from the maturase gene (*matK*) in nine species from the genus *Ranunculus*. Source: Wikipedia. Right: The median complex allows us to identify 17 one-dimensional and 3 three-dimensional homology classes. Source: [162]. Reprinted with permission: © EAI European Alliance for Innovation 2016.

histories. tARGs, like minimal ARGs [220, 356], are interpretable, explicit phylogenetic representations. But unlike minimal ARGs, they can be constructed in polynomial time.

An ARG is an explicit representation of a potential history of mutations and recombinations that, starting from an ancestor, is able to generate the sampled sequences. Let us consider a sample of n sequences with m binary characters that can take values in states 0 or 1.

Definition 5.10.1. An *ARG* is a labeled directed acyclic graph N with $n+1$ external nodes, corresponding to the n sequences, and a unique root node. There are two types of internal nodes:

1. Tree nodes, of in-degree one.
2. Recombination nodes, of in-degree two.

Each node in $\mathcal{N}$ is labeled by an m-length binary sequence, subject to the following constraints:

1. External leaf nodes are labeled by the original sequences.
2. Tree nodes are labeled by sequences that differ from the parent node in certain positions; these represent mutations.
3. Recombination nodes have sequences attached that are formed by taking the first k sites from the sequence of one of the parent nodes and appending the last $m-k$ sites from the other parent node. These labels represent recombination of the parent sequences.

We are particularly interested in ARGs satisfying minimality conditions. A *minimal ARG* is an ARG that contains the minimal number ($R_{\min}$) of single-crossover recombinations required to explain the binary sequence data [220].

Definition 5.10.2. An *ultra-minimal ARG* is a further restricted type of minimal ARG, that minimizes the function

$$D(\mathcal{N}) = \sum_{r=0}^{R_{\min}} d_r,$$

where d_r is the Hamming distance between the two sequences in the rth recombination. Examples of ultra-minimal ARGs are shown in Figure 5.50.

The condensed graph of an ARG is the graph resulting from collapsing edges that connect identically labeled nodes. Condensed graphs can be embedded into m-dimensional hypercubes and their diagonals.

Definition 5.10.3. A *topological ARG* (or *tARG*) associated to a set of condensed ultra-minimal ARGs $\{G_i = (V, E_i)\}$ explaining $\mathcal{S}$ and having the same set of vertices V is defined as the undirected graph $G = (V, E)$, with vertices V and edges $E = E_1 \cup \ldots \cup E_l$, resulting from the union of all condensed ultra-minimal ARGs.

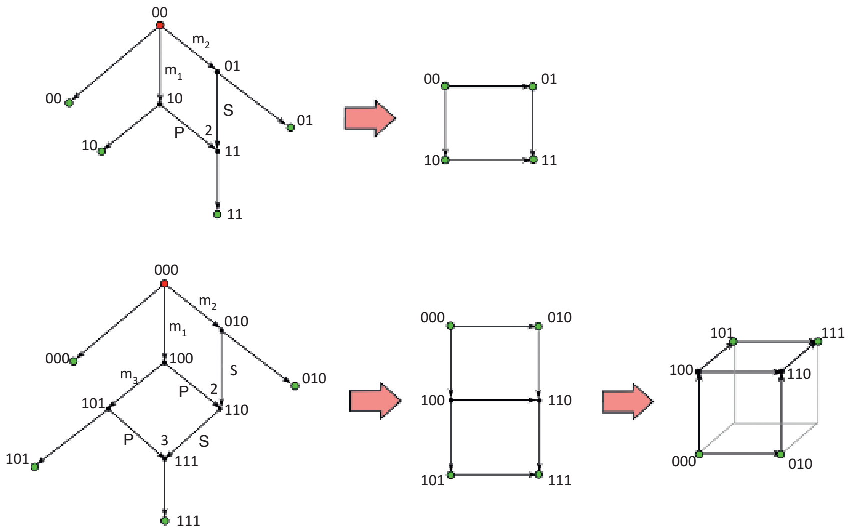


Figure 5.50 ARGs and condensed graphs. Examples of ultra-minimal ARGs and a condensed graph resulting from collapsing the unlabeled edges. The root and sampled nodes are marked in red and green. Edges in a recombination node can be annotated depending on their contribution as prefix (P) or suffix (S). Source: [88]. From Pablo G. Cámara, Arnold J. Levine, and Raúl Rabadán, 'Inference of ancestral recombination graphs through topological data analysis', PLOS Computational Biology 12.8 (2016). doi: 10.1371/journal.pcbi.1005071.

A tARG captures the possible parsimonious histories (see Figure 5.51 for examples). One advantage of tARGs over minimal ARGs is that a tARG is completely determined by its vertices, in the sense that the tARG can be computed entirely from the vertices and their labels.

Given a sample of genetic sequences, our goal is now to obtain information about the associated ultra-minimal ARGs that explain $\mathcal{S}$, without explicitly constructing them (see Figure 5.52).

We now explain how to make inferences about recombination events in topological ARGs associated to sequence data by applying persistent homology. The persistent homology of the metric space determined by the original sequence data under the Hamming distance captures information about the genetic distance between recombining parental sequences. In particular, one-dimensional classes in persistent homology correspond to loops in the tARG corresponding to the data. That is, the tARG provides a framework for explaining the persistent homology barcodes.

However, the size of the barcode provides only a lower bound on the number of recombination events in the tARG, and the larger the length of the sequences, the

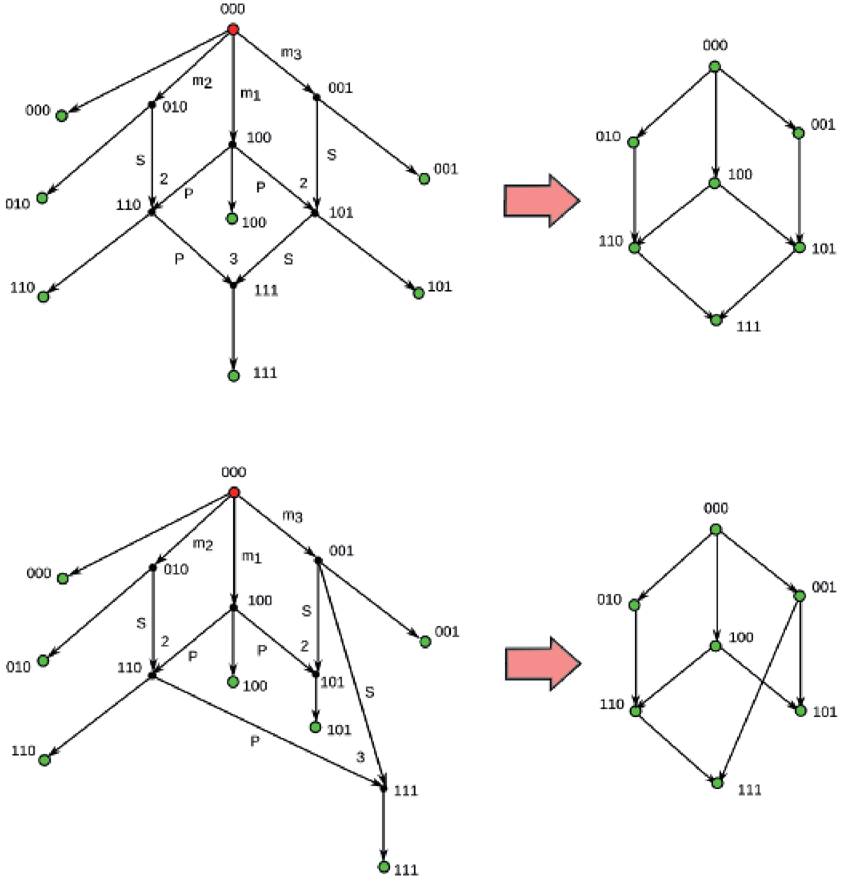


Figure 5.51 Ultra-minimal ARGs. Here we show two examples of ARGs that contain three recombinations, in this case the minimum number of recombination events. This is the minimal number required to characterize a sample of seven sequences with only three sites. Both ARGs are minimal ARGs. Source: [88]. From Pablo G. Cámara, Arnold J. Levine, and Raúl Rabadán, 'Inference of ancestral recombination graphs through topological data analysis', PLOS Computational Biology 12.8 (2016). doi: 10.1371/journal.pcbi.1005071.

worse this bound gets. A standard technique for handling this issue is to partition the sequences and reassemble local estimates. By a partition of the sequences we mean sets of subsequences specified by fixing a collection of indices $0 = i_0 < i_1 < i_2 < \ldots < i_k = m$. Now for each $0 \leq j < k$, we consider the set of sequences determined by taking the characters in positions between i_j and i_{j+1} in the original sequences. Given a partition of the sequence data into distinct intervals, one can associate a barcode that captures information about recombination events with breakpoints in each interval. Taking the union of the barcodes of a partition usually captures more recombination events than the barcode associated to the union

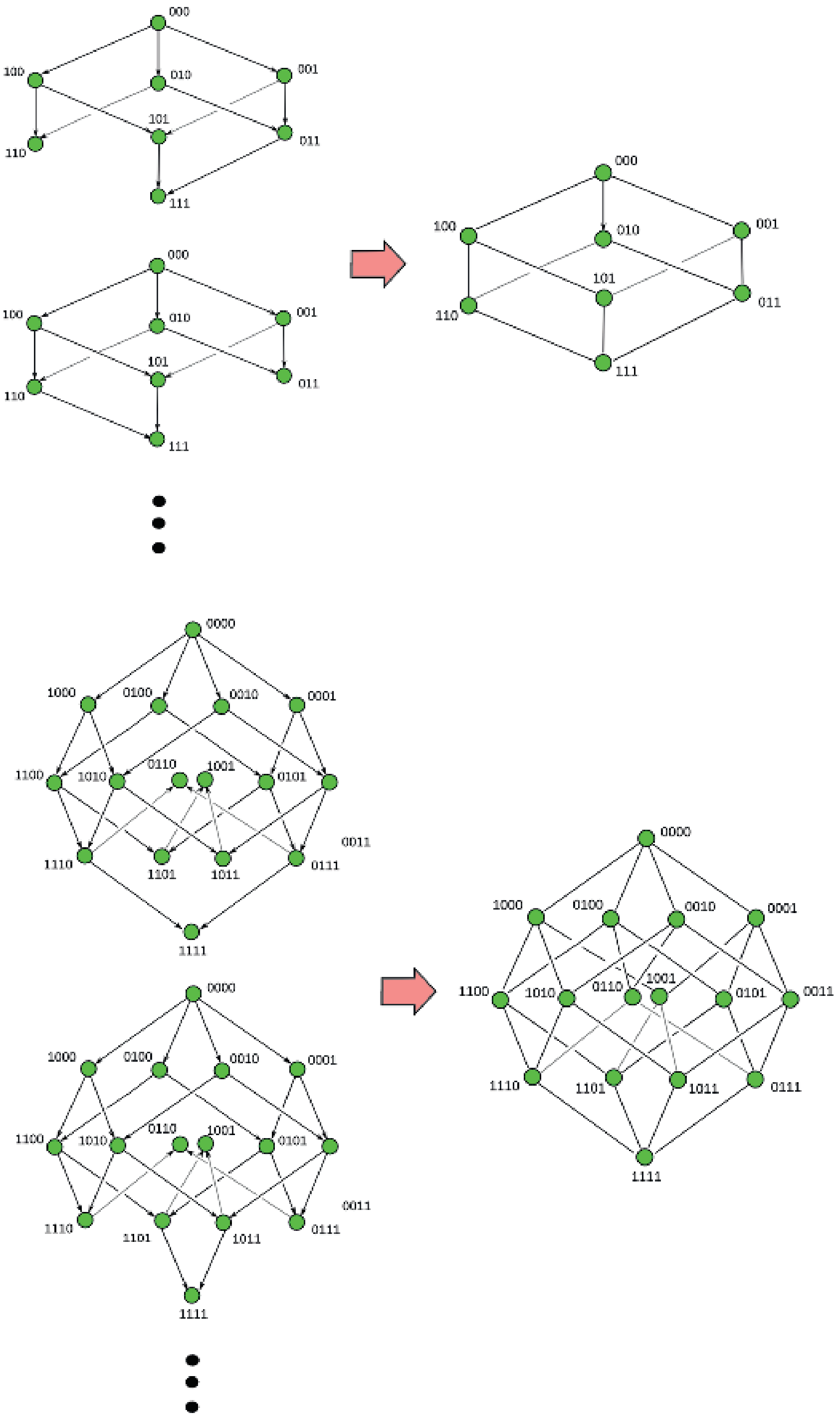


Figure 5.52 Topological ARGs. The tARG, shown on the right, can be different from the original condensed ultra-minimal ARGs, shown on the left. Source: [88]. From Pablo G. Cámara, Arnold J. Levine, and Raúl Rabadán, 'Inference of ancestral recombination graphs through topological data analysis', PLOS Computational Biology 12.8 (2016). doi: 10.1371/journal.pcbi.1005071.

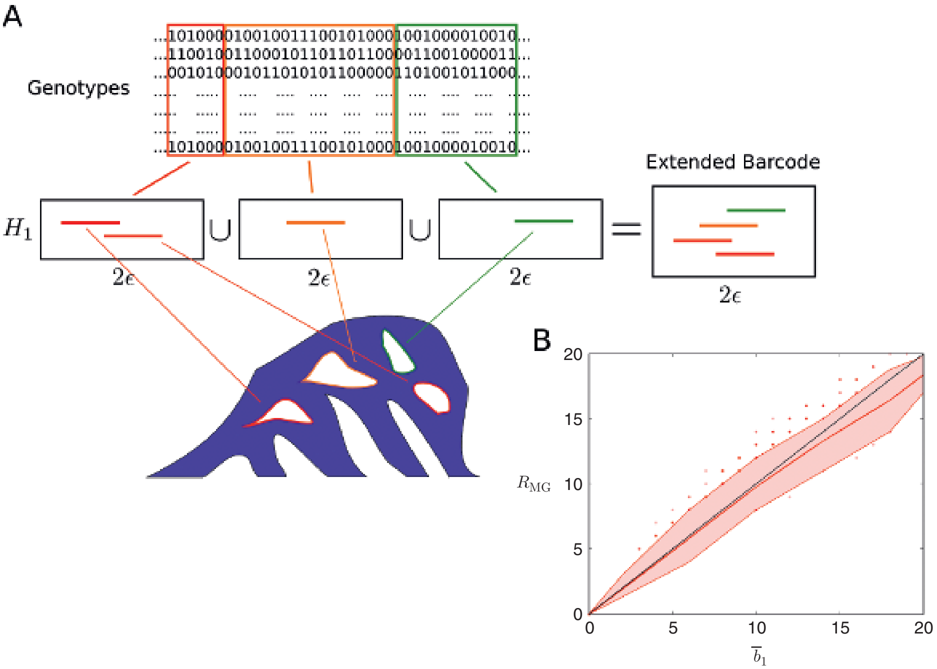


Figure 5.53 Barcode ensemble of a sample. (A) A schematic representation of the barcode ensemble of a genomic sample. Persistent homology is computed for each genomic partition of the sequences. Barcodes associated to different genomic intervals capture different recombination events with breakpoints contained within their respective partitions. The union of these barcodes builds the barcode ensemble. The total number of intervals in the barcode ensemble is denoted as $\bar{b}_1$. The genomic partitions are chosen such that $\bar{b}_1$ is maximized. (B) Comparison of the lower bounds $\bar{b}_1 \leq \overline{R}_{\mathrm{min}}$ and $R_{\mathrm{MG}} \leq R_{\mathrm{min}}$ in coalescent simulations. Values of $\bar{b}_1$ and R_{MG} are plotted for simulated samples of 40 sequences with 12 segregating sites, sampled from a population under the coalescence model with recombination. 4000 samples were simulated in total. The colored band represents the interdecile range, whereas the central line represents the mean. The values of $\bar{b}_1$ and R_{MG} are strongly correlated (Pearson's $r = 0.98$, $p < 10^{-100}$). At high recombination rates, $\bar{b}_1$ tends to be larger than R_{MG}, as cases where $\overline{R}_{\mathrm{min}} > R_{\mathrm{min}}$ occur more frequently. Source: [88]. From Pablo G. Cámara, Arnold J. Levine, and Raúl Rabadán, 'Inference of ancestral recombination graphs through topological data analysis', PLOS Computational Biology 12.8 (2016). doi: 10.1371/journal.pcbi.1005071.

of the two genomic intervals. By systematically exploring all possible partitions of the genetic sequences in a data set, it is possible to find a partition that maximizes the total number of bars in the barcodes, referred to as the *barcode ensemble* and denoted by $\bar{b}_1$ (see Figure 5.53). A detailed explanation of the algorithm to compute barcode ensembles can be found in [88]. In simulated data, $\bar{b}_1$ is a good

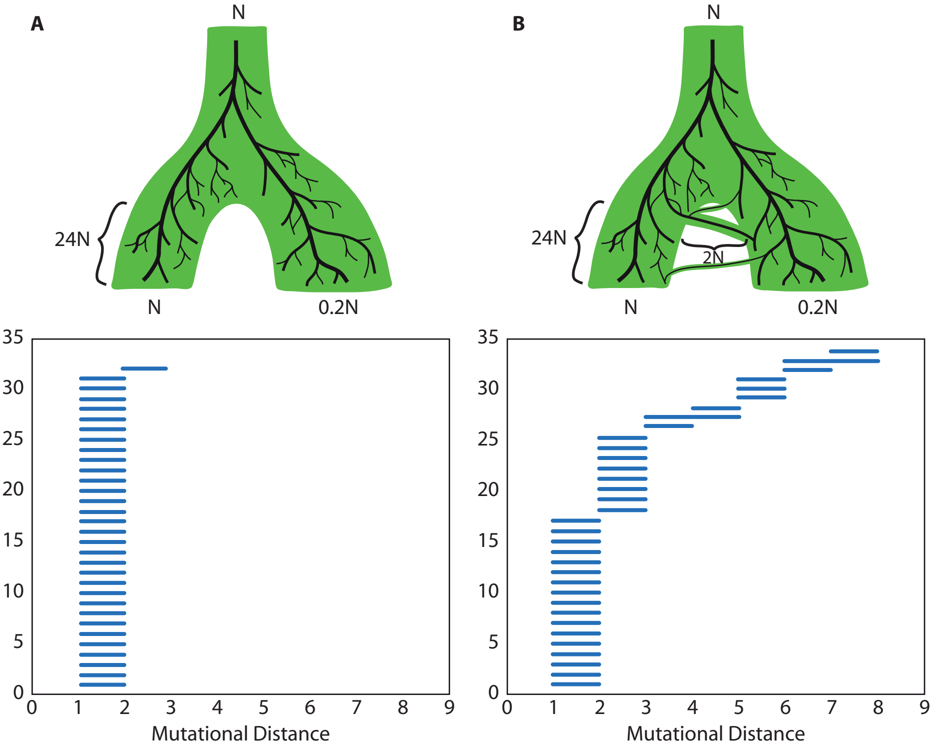


Figure 5.54 Barcode ensemble of two divergent sexually reproducing populations. The case in (A) assumes the two populations are completely isolated. All recombination events present in the barcode ensemble involve genetically close parental gametes. The case in (B) allows migration between the populations at a low rate. Some of the recombination events present in the barcode ensemble involve distant parental strains, leading to larger death times ϵ_d. The total number of detected recombination events is similar in both cases and uniform across the entire genome. Intervals with the location of the recombination breakpoints are indicated for each recombination event, where positions refer to segregating sites. Source: [88]. From Pablo G. Cámara, Arnold J. Levine, and Raúl Rabadán, 'Inference of ancestral recombination graphs through topological data analysis', PLOS Computational Biology 12.8 (2016). doi: 10.1371/journal.pcbi.1005071.

approximation of the minimum number of recombination events, as Myers and Griffiths described [370]. Barcode ensembles not only provide $\bar{b}_1$, but also richer information that bounds the genetic distances between recombining sequences.

Let us consider two cases of sampling two sexually reproducing populations with effective population sizes N and $N/5$ that diverged $24N$ generations ago. In the first case, the two populations were completely isolated from each other (Figure 5.54A). In the second case, a migration occurs between the two populations at a low rate (Figure 5.54B). We can detect the presence of gene flow by studying the barcode ensembles. Whereas the number of total detected recombination events was very similar, reflecting the fact that both examples had the same recombination rate, migration was reflected in the existence of large scale loops. Specifically, these correspond to migration events followed by recombination.

The phenotypic variation and geographical distribution of finches on the Galapagos Islands inspired Darwin's theory of the origin of species. It is believed that these finch species originated from a common ancestor 1.5 million years ago [406]. Recently, genetic information was collected from 15 different species of finches from the Galapagos archipelago and the Cocos Islands [312]. With information about homozygous single-nucleotide variants in a nine megabase genomic region of 112 finch samples, we computed a barcode ensemble. The one dimensional barcode ensemble (Figure 5.55A) indicates 13 potential recombination events. Interestingly, the majority of these events involve individuals from multiple species and include *Certhidea* samples (Figure 5.55B), the *Certhidea* being the most ancestral lineage in the data set. Barcode ensembles provide support for genetic introgression, meaning the acquisition of genetic material from one species by another through hybridization [312].

5.11 Summary

- Vertical evolution is the direct transmission of genetic material from parent to offspring.
- Horizontal evolution refers to other modes of acquisition of genomic material that are not vertical. Phylogenetic trees cannot represent these events, which are better summarized by a network with loops (reticulate evolution). At all levels of life (virus, bacteria, eukaryotes) there are reticulate events. Viruses have recombination and reassortment. Bacteria have transformation, transduction and conjugation. Eukaryotes have recombination.
- Finite metric spaces can be constructed by comparing genomic sequences. Persistence homology captures properties of these spaces. In particular, if genomic data is derived from trees, the only non-trivial homology is in dimension zero. Or equivalently, there is a topological obstruction to constructing trees when homology is found in dimensions higher than zero.
- The number of loops tells us how frequently reticulate events occur. We have seen, for instance, that HIV has a high rate of recombination relative to other viruses, and that *H. pylori* has a low rate of horizontal gene transfer among bacteria.
- The scale of bar tells us how different the species are, and persistent homology generators identify individuals whose ancestors were involved in reticulate

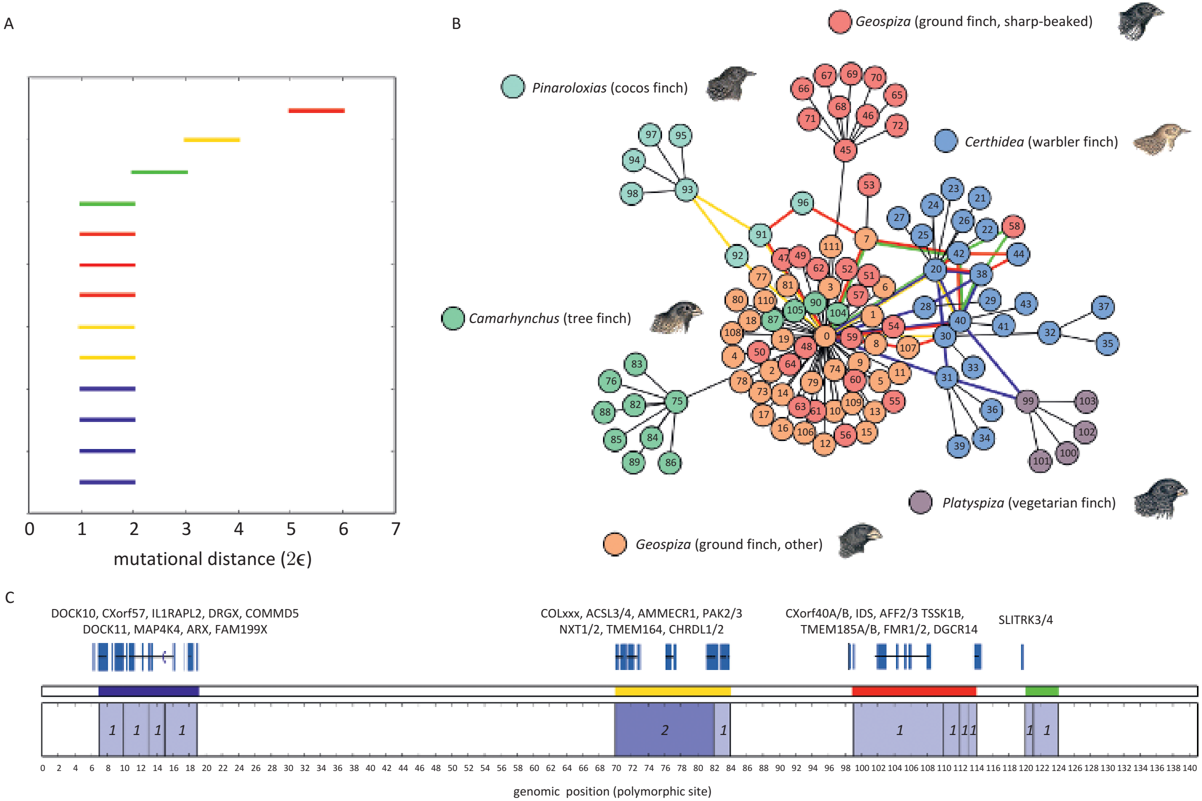


Figure 5.55 Barcode ensemble and partially reconstructed tARG of a sample of 112 Darwin's finches. The barcode ensemble is shown in (A), based on 140 homozygous SNPs present in a 9 megabase scaffold. In total, 13 recombinations or gene flow events were captured in the barcode ensemble at different genetic scales. Bars are colored according to the position of the corresponding recombination breakpoint in the genome, as depicted in (C). We also indicate the number of recombination events detected at each genomic interval, as well as some of the orthologous genes – distinct alleles thought to be passed down from a common ancestor – present in regions where recombination events were detected. The reconstructed tARG is presented in (B). Loops in the reconstructed tARG are outlined using the same color code as (A). We also include leaf nodes that do not participate in recombination, clustering them with a nearest-neighbor algorithm based on genetic distance. Edge lengths are arbitrary. Source: [88]. From Pablo G. Cámara, Arnold J. Levine, and Raúl Rabadán, 'Inference of ancestral recombination graphs through topological data analysis', PLOS Computational Biology 12.8 (2016). doi: 10.1371/journal.pcbi.1005071.

events. Barcodes in dimension one and above provide valuable information about size and frequency of genomic exchange events.

- In segmented viruses, like influenza, reassortment can lead to the formation of novel viruses by combining segments from different parental strains. Other viruses, like HIV, recombine generating high diversity.
- Persistent homology can be used to estimate actual recombination rates by fitting models involving topological features to values generated by simulations (e.g., created by the coalescent model of evolution). These models can be used to estimate recombination rates across the human genome.
- A persistent homology sliding window approach in large genomes provides fine detail on recombination rates in specific locations in the genome.
- Other constructions, like median complexes and barcode ensembles, increase the sensitivity for identification of non-tree-like events.

5.12 Suggestions for Further Reading, Databases, and Software

Here is a recommendation of a few books that explore topics related to the ones described in this chapter.

- *Gene Genealogies, Variation and Evolution*, by Jotun Hein, Mikkel Schierup and Carsten Wiuf [237], is an excellent primer in coalescence theory, with dedicated chapters on population genetics models, coalescence, ancestral recombinant graphs, and linkage disequilibrium. There is also a chapter on applications to human evolution, population structure, and migrations.
- *Phylogenetic Networks*, by Daniel Huson, Regula Rupp and Celine Scornavacca [262], provides a very clear survey of methods for inference of phylogenetic networks.
- *Viruses*, by Arnold Levine [328], is a lucid introduction for the neophyte to the world of viruses, including a description of molecular biology and historical accounts of HIV, influenza, and some other common human pathogens.
- *The Evolution and Emergence of RNA Viruses*, by Eddie Holmes [246], is a beautiful account of different evolutionary aspects of RNA viruses, insightful and very complete, with very interesting speculations about the deep origins of RNA viruses.
- *Principles of Virology: Molecular Biology*, by Jane Flynt, Lynn Enquist, Vincent Racaniello, and Anna Skalka [179] is a very clear two-volume description of the main principles of viral entry, replication, propagation, etc. Highly recommended for the reader who wants to delve deeper into how viruses operate.

- Recent review of applications of topological data analysis to genomic data have been written by P. Cámara [87] and the authors of this book [502].

Data used in the examples in this chapter and related data can be found in diverse databases.

- Influenza genomes annotated by subtype, day and geographic location can be found in the Influenza virus resource, in NCBI (National Center for Biotechnology Information, www.ncbi.nlm.nih.gov/genomes/FLU/), the Global Initiative on Sharing All Influenza Data (GISAID, http://platform.gisaid.org/), and the Influenza Research Database (IRD, www.fludb.org/).
- HIV genomes and immunological data can be obtained in Los Alamos HIV database (www.hiv.lanl.gov/content/sequence/HIV/mainpage.html). HIV genomes used in the study of HIV associated dementia can be found with consecutive GenBank accession numbers HM001362 to HM002482.
- For other viruses, Los Alamos HIV database also has compiled annotated genetic data from Hepatitis C Virus (HCV) and Hemorrhagic Fever Viruses (HFV) Databases (mostly Ebola). More general information can be found in the National Center for Biotechnology Information viral genome database (www.ncbi.nlm.nih.gov/genome/viruses/).
- The 1,000 Genome Project data can be found in The International Genome Sample Resource (IGSR, www.1000genomes.org).
- MLST bacterial data can be found in PubMLST [277].
- Data sets and software used along this chapter can be found in https://github.com/RabadanLab.

Here is a list of some software that can be used to detect recombinations and breakpoints.

- RDP in http://web.cbio.uct.ac.za/~darren/rdp.html
- GARD in www.hyphy.org/.
- BARCE in www.topali.org/.
- Simplot in https://sray.med.som.jhmi.edu/scroftware/simplot/
- PhylPro in https://cran.r-project.org/web/packages/stepwise/index.html
- Recco in http://recco.bioinf.mpi-inf.mpg.de/
- MaxChi in http://web.cbio.uct.ac.za/~darren/rdp.html.
- Chimaera in http://web.cbio.uct.ac.za/~darren/rdp.html.
- GeneConv in http://web.cbio.uct.ac.za/~darren/rdp.html;https://www.math.wustl.edu/~sawyer/geneconv/.

- 3seq Substitution in http://web.cbio.uct.ac.za/~darren/rdp.html, http://mol.ax/software/3seq/.
- PhiPack in www.splitstree.org/.
- SiScan in http://web.cbio.uct.ac.za/ darren/rdp.html; http://mateo.fourment.free.fr/software.html.
- TREE in https://github.com/MelissaMcguirl/TREE.