Approaches to Ancestral Pangenomes and Their Phylogenetic Reconstruction

Xintong Zhou

Thesis submitted to the University of Ottawa in partial Fulfillment of the
requirements for the degree of
Master of Science Mathematics and Statistics[1]

Department of Mathematics and Statistics
Faculty of Science
University of Ottawa

[1]The M.Sc. program is a joint program with Carleton University, administered by the Ottawa-Carleton Institute of Mathematics and Statistics

Abstract

We investigate the problem of reconstructing ancestral pangenomes from present-day genomic data by modelling structural variation and evolutionary turnover. Our first chapter models ancestral and descendant pangenomes as sets of gene adjacencies, using phylogenetic validation — based on Dollo's law — to filter out recent innovations unlikely to have existed in any ancestor. This approach enables a meaningful reconstruction of ancestral gene order via iterative steinerization, even without optimization. In a second chapter, we complement this framework with a probabilistic tree model in which discrete objects (e.g., genes or features) are transmitted down a hierarchical phylogeny and replaced, respecting Dollo, with a certain probability. We derive theoretical expectations for the retention and overlap of ancestral objects across nodes and assess the accuracy of steinerization-based reconstruction in simulated datasets. Our simulations demonstrate that while theoretical predictions align with observed retention under low replacement rates, random divergence among novel objects introduces noise in deeper or faster-evolving trees. Together, these studies provide some promising approaches to understanding the limits and potential of ancestral reconstruction in a pangenomic landscape.

Acknowledgements

I would like to sincerely thank my supervisor, Professor David Sankoff, for giving me the opportunity to pursue graduate studies. Without his support, I do not believe I would have had the chance to complete a degree in statistics. Many times, even late at night, I reached out to him with questions about my thesis, and he always responded with patience and guidance. I feel very fortunate to have had the chance to work closely with him. This thesis would not have been possible without his supervision, and I consider it one of my proudest accomplishments.

Contents

List of Figures

List of Tables

Preface

The research reported in this thesis started out as an attempt to extend the notion of pangenome graphs [1] using the linearization of directed graphs [2, 3]. The linearization approach, with its goal of unitary genomes, was quickly seen as antithetical to pangenome evolutionary inference. Instead, I explored the reconstruction of ancestral pangenomes through two complementary models: one based on structural variation and adjacency-based phylogenetic validation, pertinent to the "core" pangenome [4], and the other grounded in probabilistic modelling of object transmission along a phylogenetic tree, more relevant to the "accessory" or "dispensable" pangenome. Motivated by challenges in capturing ancestral genomic structure amidst evolutionary turnover, the work combines combinatorial and statistical methods to simulate, analyze, and evaluate reconstruction procedures. Together, these approaches aim to clarify the theoretical and practical limits of ancestral inference in pangenomic systems.

Chapter 2 reproduces a paper [5] submitted to the Twenty-second RECOMB Satellite Conference on Comparative Genomics, accepted after evaluation by four referees for presentation at the conference. This took place in Seoul, South Korea on April 24, 2025. The accepted papers from the conference have been edited by Giltae Song, Pusan National University, for publication in a volume in *Lecture Notes in Computer Science*, number **15666**, entitled *Comparative Genomics*, with publication date in 2026, although the issue was actually published online in September, 2025.

Chapter 3 will be combined with the previous chapter in an expanded version, invited for inclusion in a special issue of *Journal of Computational Biology*.

Chapter 1

Introduction

This thesis is an attempt to bring together three disparate sets of evolutionary models. The first of these is molecular phylogenetics, a highly specialized kind of hierarchical clustering based on nucleic acid or protein-sequence data, which attempts to infer the set of speciation events, and sometimes their timing, that gave rise to the phylogeny, or family tree, of a number of present-day species. Second is pangenomics, the study of the variation that distinguishes the genomes of the individual members of a species. This focuses on differences in the gene content and the order of these genes in the genomes of individuals. It is largely statistical in nature but often has a combinatorial aspect. Finally, ancestral reconstruction is a relatively recent enterprise that attempts to estimate ancestral genomes, their gene content and gene order, concurrently with or following phylogenetic analysis.

All species where the genomes of many individuals have been sequenced show variation and their pangenomes have or could be constructed. In the context of the phylogenetics of a number of species each represented by a pangenome, it does not seem appropriate to simply reduce each pangenome to a linear order and then proceed with a traditional phylogenetic analysis of these linearized genomes. It stands to reason that ancestral genomes would also have been variable and that the notion of a pangenome would be applicable to them. After all, it is not a new idea that an ancestral population may be more or less heterogeneous with respect to the genomes of individuals or groups. This is explicit in the modern recognition of incomplete lineage sorting [6], but it was understood earlier, such as in the description of species as clouds or quasispecies of more-or-less closely related individuals [7]. Today, although the concept of an ancestral pangenome is recognized and described discursively [8], its inference in terms of a set of distinct genomes has not been attempted. Thus we may ask whether it is feasible to develop phylogenetic methods that could analyse the separate pangenomes of a number of species in order to reconstruct ancestral pangenomes.

Thus we will explore the notion of phylogenetic analysis of pangenomes, where

the root ancestor and all the intermediate ancestors are also pangenomes.

In this thesis, we explore two avenues to pangenome reconstruction. The first focuses on the global structure of the individual genomes making up a pangenome and their evolution through chromosomal rearrangement disrupting gene order, independently on separate lineages in the phylogenetic tree. To sharpen this focus, deletion or substitution of genes are not considered. Instead, gene adjacencies and breakpoints between genes play a major role. Our approach to the generation of modern pangenomes produces a great surfeit of adjacencies. The problem is to identify which ones derive from the original ancestral pangenome.

The major technique in the reconstruction is "phylogenetic validation", whereby adjacencies are attributed to an intermediate ancestor in the tree only if the adjacency is present in at least two of the three lineages incident to this ancestor. This criterion plays a key role in the "steinerization" procedure of optimizing all the intermediate ancestors simultaneously. Our method does not go all the way to reconstructing

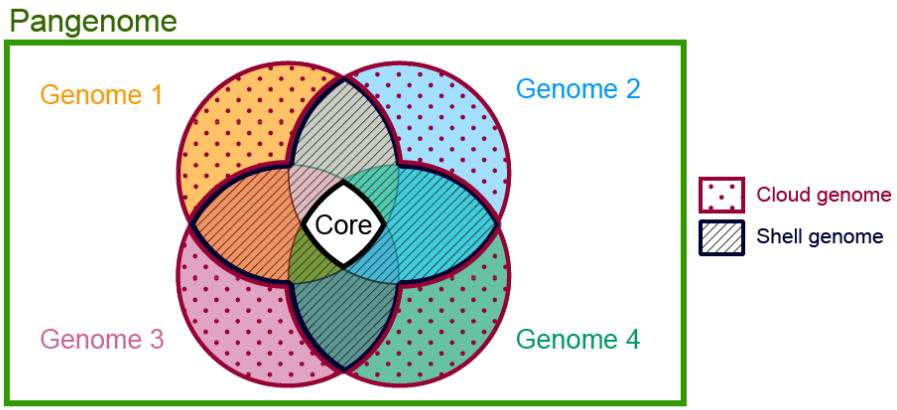

Figure 1.1: Three sets of genes: core, shell and cloud genome. The core genome comprises the genes that are present in all genomes analyzed. The shell genome consists of the genes shared by the majority or a large proportion of genomes. The gene families present in only one genome or are described as dispensable or cloud genome. From Wikipedia article "Pan-genome"[9].

the ancestral pangenome. It does reconstruct a set of adjacencies, some of which are incompatible with a single genome and constitute important evidence of the ancestral pangenome. Where traditional inference of ancestral genomes — individual genomes, not pangenomes — is based on the abhorrence [11] of conflicting adjacencies in the search for strict linearity of chromosomes, our hospitality towards these conflicts is an approach that will allow algorithms to specifically target multi-genome pangenomes. In the Conclusions chapter, we propose approaches to achieving this inference.

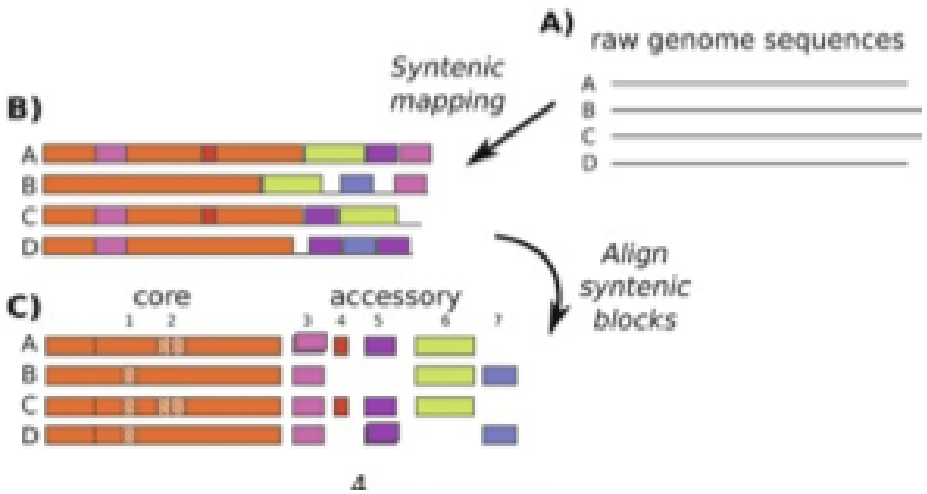

Figure 1.2: Schema of whole-genome alignment of genomes within a pangenome. Extensive homologous regions (syntenic blocks) are mapped in different colours. Orange indicates the core genome; the other colours indicate accessory syntenic regions. From Lassalle and Didelot (2020) [12]

The second approach in this thesis is complementary to the first. It ignores gene order and rearrangements entirely and focuses on the loss and substitution of genes in a genome [13, 14]. This approach assumes Dollo's law [15] whereby a gene substitution cannot be reversed later in a lineage. The main question we ask is to what extent does gene content survive as a function of evolutionary rate in the lineages and speciation rate in the tree. Again we use phylogenetic validation to prune substitutions during a steinerizing reconstruction. We provide detailed answers through simulation.

While there is a considerable literature on formalisms for pangenomes and an entirely separate literature on ancestral genome reconstruction, *there is no previous work at all on the reconstruction of gene orders in ancestral pangenomes*, as natural a project it may seem.

1.1 Background

Pangenomes Pangenomes aim to represent all the variation found in the genomes of a set of related organisms — populations, species, genera — which we call the constituent genomes. There are two main approaches to the formal study of the gene complement of pangenomes; gene-centric as contrasted to the sequence-centric focus on the DNA sequence of genomes. One gene-centric approach is identification of the "core" genes (or ortholog group) present in all the constituent genomes, versus the "accessory" or "shell" genes, present in a sizable subset of the constituent genes and the "unique" or "cloud" genes, present in a single genome, as in Figures 1.1 and 1.2.

(The meanings of terms like accessible, cloud and dispensable vary in the literature.) The core may contain fewer than 10% of the pangenome genes, as in the case of some bacteria [16], from 30–70% for many plants and animals [17, 18, 19] or over 95% for humans [20].

In bacteria, these quantities must be adjusted in the case of "open pangenomes" such as that of Escherichia coli [21], where tens of thousands of rare variants accumulate at the same pace as new strains are sequenced.

The second gene-centric approach to the pangenome is that of "pangenome graphs". Here, genes (or ortholog groups) are represented as vertices. Adjacent genes in a chromosome of a constituent genome are connected by an edge, often a directed edge. The massive redundancies and conflicts inherent in the resulting raw structure are then reduced by various algorithms to acyclic or locally acyclic graphs. Many types of graph are used to represent the output of these algorithms, but most of these focus on sequences, where full analysis of the gene content in secondary or absent e.g., de Bruijn graphs [22] or cactus graphs [23]. A number of primarily gene-centric pangenome-graph algorithms and packages are, however, available [24, 25, 26]. The input to these procedures is a number, which may be quite small (e.g., less than 10) up to very large (e.g., many hundreds) of sequenced genomes, all of them strictly linear, the proposed constituents of a pangenome. The goal is to convert these data into a pangenome graph, which is generally not linear, but which compactly represents both the common portions of the individual constituent genomes as well as the diversity in sequence or gene order among them.

Phylogeny and ancestral genome reconstruction. Building trees in biology has a venerable history, most notably in the mid-eighteenth century comprehensive ouevre of Carl Linnaeus [27] categorizing all (around 10,0000 at the time) biological species in a nested hierarchical system, inherently a hierarchical taxonomy, though without any tree drawings. The evolutionary re-interpretation of this concept, which we call "phylogeny", was promulgated a century later by Charles Darwin [28] and included a tree drawing, as well as the impetus for paleontological trees tying branching patterns to geological strata. The taxonomy versus phylogeny (descriptive versus explanatory) tensions were clearly reflected after another century in Sokal and Sneath's "phenetics" [29] versus Farris's "cladistics" [30] dispute preceding the modern statistical synthesis led by Felsenstein [31]. His introduction of likelihood methodology paved the way for today's Bayesian methodology [32]. In parallel to these developments in the 1960s and early 1970s, molecular phylogenies were initiated by Zuckerkandl and Pauling [34] and Dayhoff [35], using protein, and Sankoff and Cedergren for nucleic acids [36, 38]. Wider applications of phylogenetics will be discussed in some detail in Chapter 3.2.

Although Chapter 2 will deal with gene-order, in that it will focus on how we can use the gene adjacencies in the data on present-day pangenomes to reconstruct

the gene adjacencies in the ancestor pangenome, it will diverge from previous work on single-ancestor genome reconstruction as in [39, 41, 42, 40, 43] in our lab and elsewhere [44, 45], designed to reconstruct entire single ancestral genomes. The previous work was all designed to produce a single ancestor genome, but we will see that this is antithetical to our goals. Despite much research relevant to the gene *content* of ancestral pangenomes [46] virtually no attention has been paid to their order.

Phylogenetic inference has two components: the "large phylogeny problem" and the "small phylogeny problem" [47]. The former is the inference of the "topology", the branching structure of the tree. The second is the inference of the DNA sequences, proteins or genomes inferred to be associated with each branching point. In my study of the nature of the ancestral pangenome, I will focus on the small problem and work with a fixed tree topology. This obviates the necessity of the single-, average- or complete-link agglomerative clustering [29], neighbour-joining [33], maximum likelihood or Bayesian phylogenetic methods, which are focused on the inferring the topology. Instead, we make use of a "phylogenetic validation" criterion to construct ancestors for a given tree.

Gene-substitution models Chapter 3 falls in the tradition of pre-genomic pure gene-substitution models such as that of Pagel [48, 49]. More recently, Treangen & Rocha [50] define gene replacement events distinct from gain/loss; they often act as swap events, preserving genome size.

Daubin & Ochman [51, 52] show that many bacterial "gene acquisitions" are actually replacements of pre-existing genes, not additive gains. They treat these as paired gain/loss events, not independent birth–death. There are a number of "constant genome size" models such as ours in comparative genomics. Some groups have worked with evolutionary models assuming a fixed number of genes, forcing gene "turnover" to occur as balanced replacement. Kunin & Ouzounis [53] worked on gene-repertoire dynamics under a fixed-size genome constraint. Gene change is handled as gain of non-homologous genes plus simultaneous loss of old genes, with single substitution events.

Another direct precedent for substituting non-homologous genes while keeping copy number constant is [54], describing processes where a genome trades one gene for another, maintaining a fixed gene count. Related work can be seen in [55, 56].

None of this work, however, deals with our substitution model where a completely new gene appears at each event, and the effect of substitution on ancestral inference in a phylogeny is the key question.

Our problem One topic almost never broached in the pangenome literature is the phylogeny of pangenomes. In the context of the phylogenetics of a number of species or genera each represented by a pangenome, why settle for simply reducing each pangenome to a linear, or at least locally acyclic, order and then proceed with

a traditional phylogenetic analysis of these linearized genomes? After all, it is not a new idea that an ancestral population may be more or less heterogeneous with respect to the genomes of individuals or groups. This is explicit in the modern recognition of incomplete lineage sorting [6], but it was understood earlier, such as in the description of species as clouds or quasispecies of more or less closely related individuals [7]. In this thesis, then, we develop a "small" phylogenetic analysis of pangenomes, where the inferred ancestors are also pangenomes.

We propose two approaches to attacking this problem and make progress suggestive of future lines of research.

Chapter 2

Ancestral pangenomes and their phylogenetic reconstruction

2.1 Summary of Approach

Looking past questions of gene content, we focus on structural variants of the genomes within a pangenome and seek to find a phylogeny where all the ancestral nodes, including the root, are also pangenomes. Representations of pangenomes generally search for compact structures that emphasize common regions or common duplications among the constituent genomes but necessarily sacrifice some other aspects of gene order. Since the gene order of a monoploid genome is basically just the set of all the gene adjacencies it is composed of, we will consider a pangenome as being made up all the adjacencies of genes appearing in at least one of its constituent genomes. Our key combinatorial tool, *phylogenetic validation*, does not involve optimization but is simply a filter that removes any adjacencies present in input (extant) pangenomes that are unlikely to have been present in any ancestor, inspired by Dollo's law of irreversible changes. In simulations, this tool turns out to be extraordinarily efficient in retrieving only adjacencies in the original ancestor.

2.2 Motivation

Pangenome graphs [1] represent a set of variant genomes of a species as a directed graph with some device for handling differences in gene content and gene orders among these variants. Differences in gene content are usually discussed statistically in terms of core versus non-core genes, while structural variants due to insertion, deletion and duplication (tandem or otherwise) are amenable to several kinds of graphical representation. However, in the models in this paper, for the purposes of focusing on the variability in gene order, we simply assume that gene content is identical across

all the genomes, and does not involve paralogy.

The strategy in previous approaches to comparing DAG or DG representations of gene order between two species has been to extract a linear order from each of the two graphs, doing as little violence as possible to the information contained in each of them, in such a way that these two linear orders are optimally similar in terms of rearrangement distance [2, 3].

In the context of the phylogenetics of a number of species each represented by a pangenome, it does not seem appropriate to simply reduce each pangenome to a linear order and then proceed with a traditional phylogenetic analysis of these linearized genomes. After all, it is not a new idea that an ancestral population may be more or less heterogeneous with respect to the genomes of individuals or groups. This is explicit in the modern recognition of incomplete lineage sorting [6], but it was understood earlier, such as in the description of species as clouds or quasispecies of more or less closely related individuals [7]. In this paper, then, we explore the notion of phylogenetic analysis of pangenomes, where the root ancestor and all the intermediate ancestors are also pangenomes. In this initial study, we model the pangenomes and their evolution in the simplest terms.

2.3 Definitions

Structure. A pangenome consists of a set of related genomes $G = \{g_1, ..., g_\gamma\}$, which are unichromosomal and linear and which all contain the same n genes. A gene x is denoted by the set of gene ends $\{x^h, x^t\}$, where h (heads) and t (tails) are assigned arbitrarily. The distinct identity of each genome g_i resides in its gene order, which is a set Adj_{g_i} or simply Adj_i, of $n + 1$ "adjacencies", ordered pairs containing two ends from two different genes (x, y), plus two terminal pairs representing the ends of the genome $(0, x^h)$ (or $(0, x^t)$) and $(0, y^h)$ (or $(0, y^t)$), such that all $2n$ gene ends are in exactly one pair of the $n + 1$ adjacencies in Adj_i. This set contains all information about the structure of a genome.

Evolution. Evolution of the pangenome proceeds by a number of inversions (reversals) affecting each of its constituent genomes independently.

An inversion in genome g_i replaces two adjacencies from Adj_i by two new pairs, with reversed order, as illustrated in Figure 2.1.

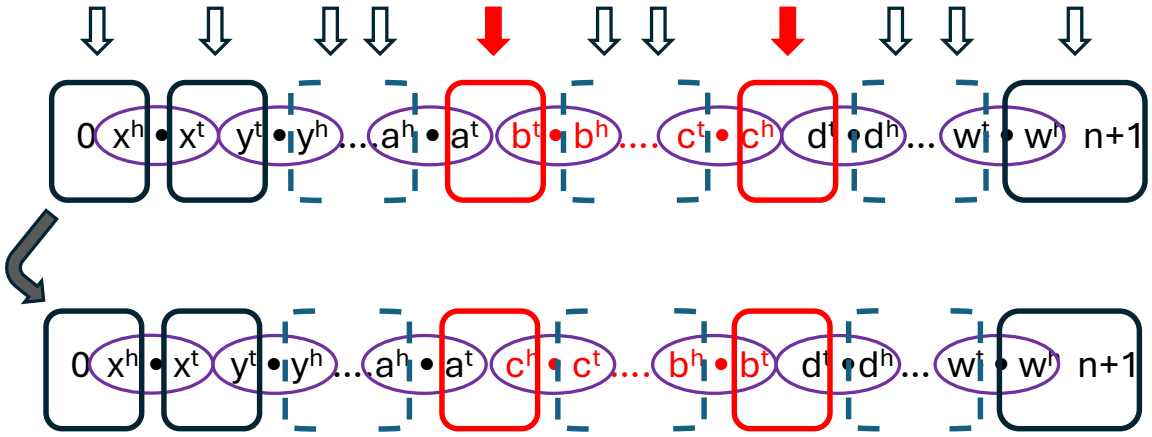

Figure 2.1: Inversion involving adjacencies between genes a and b and between c and d. Genes are enclosed in ovals, adjacencies in rectangles. Dashed incomplete rectangles contain just one member of an adjacency. Arrows indicate potential break points.

Phylogeny. To represent one instance of evolution, we contruct a phylogenetic tree including

- a root pangenome vertex,
- three descendant pangenome vertices connected to the root, each of which may represent a modern (extant) pangenome (terminal vertex — Degree 1) or an intermediate ancestor vertex (internal vertex — Degree 3),
- two descendant pangenome vertices connected to each intermediate ancestor vertex, where both of these can be either another intermediate ancestor pangenome or a terminal pangenome.

Although there is a natural temporal orientation — namely from the root towards the modern genomes, in simulating data — topologically, the tree is a binary branching tree, with the root having Degree 3, like all other non-terminal vertices.

Measurement. We write $pairs(g_i) = |Adj_i|$ and measure the similarity between two genomes g_i and g_j as $pairs(g_i \cap g_j) = |Adj_i \cap Adj_j|$ — and eventually between two pangenomes Y and Z, $pairs(Y \cap Z) = |(\cup Adj_{g_i \in Y}) \cap (\cup Adj_{g_j \in Z})|$ — as the number of adjacencies they contain in common. The count of adjacencies takes into account neither the relative order of the two genes in the genome nor the heads/tails identity of the gene ends involved.

2.4 Generating divergent modern pangenomes from an ancestral pangenome

2.4.1 Generation

The ancestor. The ancestral pangenome X is simulated by independently generating three genomes X_1, X_2, X_3 from the sequence $1, 2, 3, \ldots, 100$, using r random reversals of lengths sampled from a negative binomial with mean 10 and variance 400, appropriately truncated. We write $X_i \in X$ for $i = 1, 2, 3$, and the set of adjacencies in X is $Adj(X) = \cup_{i=1}^{3} Adj_i$. We explore parameter values $r = 5, 10, 20, 40, 80$ and 120.

The first generation. Three descendant genomes A_i, B_i, C_i are generated from each genome X_i in pangenome X using r random reversals of lengths sampled from the same negative binomial distribution as before. Then the descendant pangenomes are $A = \{A_1, A_2, A_3\}, B = \{B_1, B_2, B_3\}$ and $C = \{C_1, C_2, C_3\}$ and, for example, $Adj(A) = \cup_{i=1}^{3} Adj_{A_i}$.

Further branching. If a descendant pangenome D is itself to be considered an (intermediate) ancestor of two other pangenomes F and G, the three genomes in D each produce two further descendants, one which becomes part of F and the other part of G.

2.5 Inference

2.5.1 Phylogenetic validation and the pangenomic median

Our key reconstruction technique is based on Dollo's idea that, in certain biological contexts, phylogenetic characters, such as the adjacencies we study here, are gained only once and can never be regained if they are lost [15]. This is realized in an unrooted tree by the property that the set of vertices containing the character are connected. It is a necessary and sufficient condition, valid both for terminal vertices (or degree 1) and internal (ancestral) ones (degree 3 in an unrooted binary tree).

Formally, the connectedness condition for a character can be satisfied by a set of non-terminal nodes of a tree or, trivially, by a single terminal node. For phylogenetic reconstruction, however, we require that an adjacency be present in the "modern" genomes associated with at least two terminal vertices, so that by connectedness we can reconstruct that it must have been present in their most recent common ancestor. Otherwise, if it were present only in one terminal set, it could not be inferred as present in any of the non-terminal vertices.

In an unrooted binary branching tree, each non-terminal vertex subtends three subtrees, as in Figure 2.2. For an adjacency to be used in constructing a phylogenetic tree and the output sets, clearly each adjacency must be present in at least two of the three subtrees, as illustrated in Figure 1. More precisely, each adjacency must be present at least in one terminal vertex set in at least two of the three subtrees. We call these adjacencies "phylogenetically validated". The possibility that an adjacency

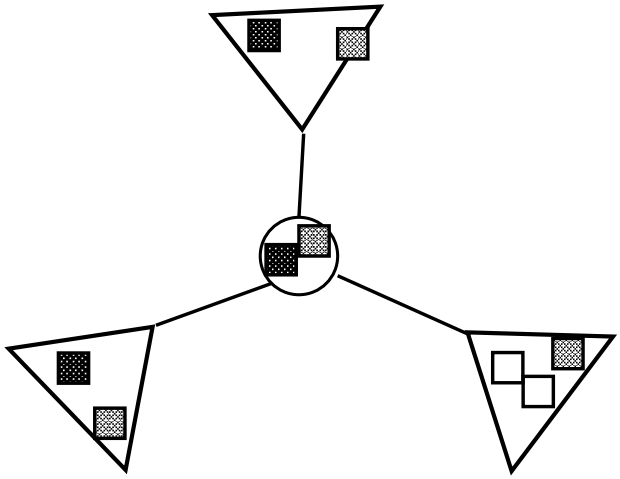

Figure 2.2: Necessary condition for adjacencies to appear at an internal vertex associated with an ancestral pangenome of a binary branching phylogenetic tree. Light shaded adjacency (small square) appears in all three trees (triangles) subtended by the internal vertex (circle). Dark shaded adjacency appears in only two of the trees. Unshaded adjacency appears in only one subtree so does not affect internal vertex. The shaded adjacencies are "phylogenetically validated" with respect to the internal vertex. The unshaded one is not validated. Adapted from [43]

originates twice or more over the phylogenetic time span, so that connectedness is not assured, is not zero, but very small, at least for small or moderate rates of evolution, so that errors in the validation process would be rare.

The median. The pairs in common between pangenomes A and B are $Adj_A \cap Adj_B$, between pangenome B and C are $Adj_B \cap Adj_C$ and between C and A are $Adj_C \cap Adj_A$. Then all the pairs in at least two of the three pangenomes are

$$X' = (Adj_A \cap Adj_B) \cup (Adj_B \cap Adj_C) \cup (Adj_C \cap Adj_A). \qquad (2.5.1)$$

Equation (2.5.1) is an expression of the phylogenetic validation criterion, excluding adjacencies that are in only one of the pangenomes A, B or C, as well as adjacencies that are in none of them.

Steinerization The phylogenies we are modelling and inferring are situated in historical time, with an original "root" vertex representing ancestor X at time zero and all edges directed away from this vertex.

In solving the small phylogeny problem through an iterative "steinerization" process — a generalization of Camin and Sokal's 1965 nearest-neighbour interchange concept [57], extended in 1994 to general metric spaces [58] and again in 1999 [59] — we first select any non-terminal vertex e.g., X, in a unrooted representation of the phylogeny. This subtends three disjoint subtrees. All gene pairs that occur in some modern pangenome in at least two of these three subtrees, as in Figure 2.2, are considered to form a first "candidate" set of pairs eligible to be in the ancestral pangenome reconstructed at X. At this point, vertex X is considered to be "processed".

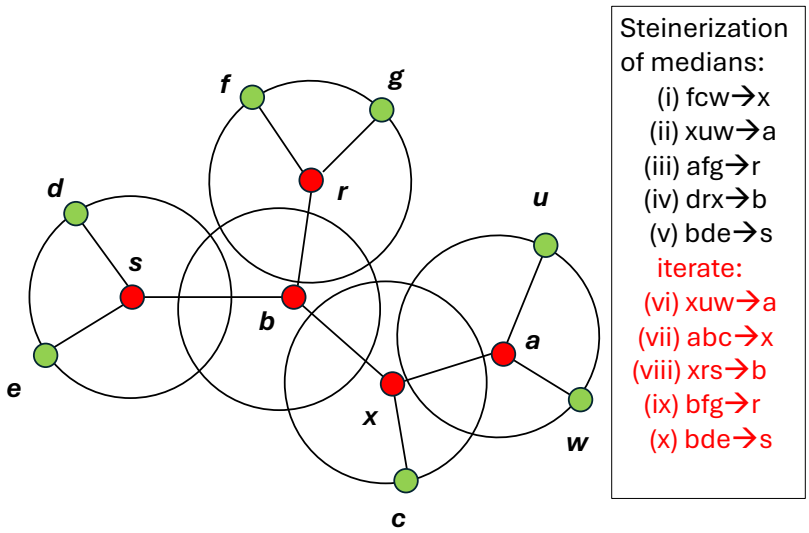

Figure 2.3: Calculating the ancestral pangenomes through steinerization based on the medians.

The next, and the following $n-2$ steps, are basically all the same: some non-terminal vertex Y is chosen that has not already been processed, but which is connected to at least one non-terminal vertex that has been. Then, instead of searching for candidate pairs in all the genomes in the three subtended trees, only pairs in the processed non-terminal vertices connected directly to Y are considered, plus any pairs in the modern pangenomes in the remaining subtended trees, if there are any.

The next cycle of $n-2$ steps begins again with X and continues with Y and the same sequence of non-terminal vertices visited in the first cycle. Now all non-terminal vertices will have been processed, so the candidate pairs at each ancestor vertex are drawn from the three connected vertices according to the two out of three rule. The cycles continue until convergence, defined in terms of the stability of the total number of candidate pairs in all the non-terminal vertices. These now constitute the sets of

pairs inferred to constitute the ancestral pangenomes.

2.6 Simulations

The steinerizing process calculates the ancestral (root and intermediate ancestors) pangenomes by iterating the median problem for all non-terminal vertices until convergence, which we illustrate in Figure 2.3 for seven terminal vertices and four non-terminal vertices.

For our simulations, however, we used the smaller tree in Figure 2.4 with only six terminal vertices and three non-terminal vertices. For $r = 5, 10, 20, 40, 80$ and 120, we generated genomes X_1, X_2 and X_3 from the sequence $1, 2, \ldots, 100$ using r inversions for each. We set the ancestor pangenome $X = (X_1, X_2, X_3)$ and generated the intermediate ancestors A, B and C as described in Section 2.4.1 above. From these ancestors, we then generated the "extant" pangenomes R, S, V, W, T, Z. With these simulated data, we could then reconstruct estimated intermediate ances-

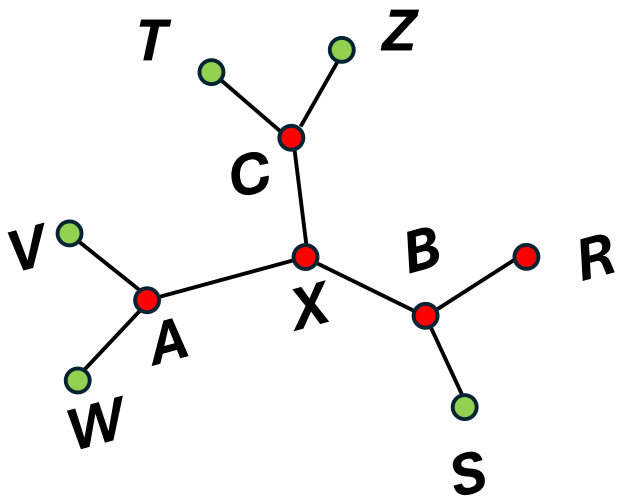

Figure 2.4: Phylogeny with ancestor pangenome X used in simulations.

tor pangenomes A', B', C' in terms of the adjacencies in the extant pangenomes that were filtered through the phylogenetic validation criterion.

Finally we used the intermediate ancestors to construct X'. The entire inference procedure was iterated as in Figure 2.3 until convergence. Each simulation was repeated 100 times and the mean numbers of adjacencies is reported in Table 2.1 and Figure 2.5. These results are remarkable in that for $r = 5$ and even $r = 10$, almost all the adjacencies in X are recovered in X', and few extraneous adjacencies manage

inversions	$\lvert Adj_X \cap Adj_{X'} \rvert$	$\lvert Adj_X \backslash Adj_{X'} \rvert$	$\lvert Adj_{X'} \backslash Adj_X \rvert$
5	125	2	4
10	137	14	8
20	121	69	12
40	44	195	9
80	3	276	3
120	0	290	1

Table 2.1: Results of the inference process. The parameter r measuring the rate of evolution ranges from 5 per time period to 120. The first two columns show that up to $r = 10$, almost all of the adjacencies in X are recovered in X'.

to make it into X'. On the other hand, it is clear that increasing the inversion rate to 40 or higher will defeat the method.

The power of the phylogenetic validity filter is clear from Table 2.1, when we see the hundreds of adjacencies that are filtered away either in the reconstruction of A', B' and C', or in the final reconstruction of X'.

inversions	union of $(R, S, T, V, W, Z) \backslash X'$	union of $(A, B, C) \backslash X'$
5	229	79
10	424	155
20	780	318
40	1293	600
80	1630	810
120	1691	855

Table 2.2: Adjacencies in extant pangenomes and in intermediate ancestors that are filtered out by the phylogenetic validity criterion.

2.7 Nonlinearity

Genes in more than two adjacencies give rise to branchings that disrupt the linearity of the reconstruction. Figure 2.6 shows, for various numbers of inversions in the generation scheme, the final set of genes assembled according to the adjacencies as-

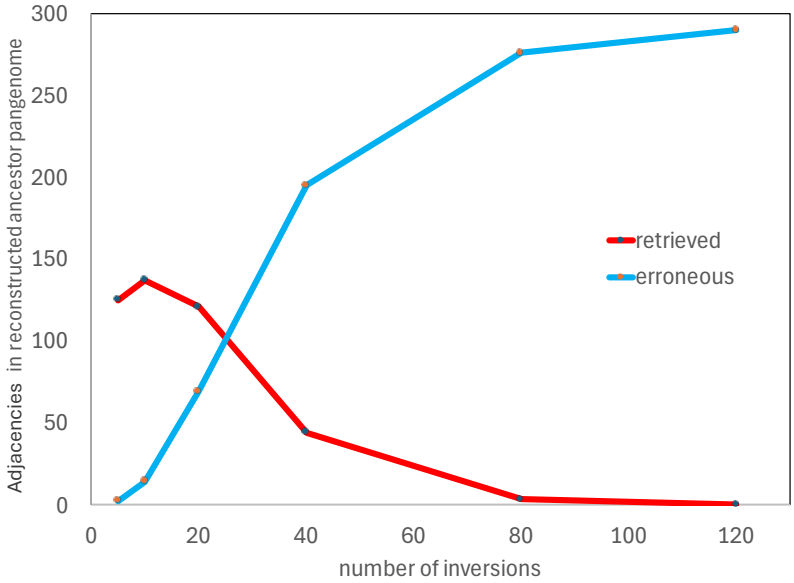

Figure 2.5: Effect of evolutionary rate on number of correctly reconstructed adjacencies. With more than 30 inversions per evolutionary period, the number of accurately retrieved adjacencies drops sharply, while the number of simulated adjacencies not present in the ancestral pangenome increases.

sembled for the ancestor. These reveal multiple branching points and knotting of the structure.

Future work may focus on using the branching points in this kind of assembly as hints toward teasing out the disjoint linear genomes forming the ancestral pangenome.

2.8 Large phylogeny

In the pangenomic context, there is a major difference with other phylogenetic problems in the small phylogeny context: namely, the use of phylogenetic validation instead of some optimization criterion. In the large phylogeny case, however, there is little difference in the basic intractability of the problem, necessitating exhaustive approaches, heuristics and the like. Figure 2.7 shows the case of six terminal vertices and only 105 different possible phylogenies. In the present study, however, we simply evaluated one additional tree using the same data.

The results of using the second tree to reconstruct the ancestor at the origin of the data were equivocal: slightly fewer original adjacencies recovered and also slightly more extraneous adjacencies in X'. The potential of the method for large phylogenies awaits further investigation.

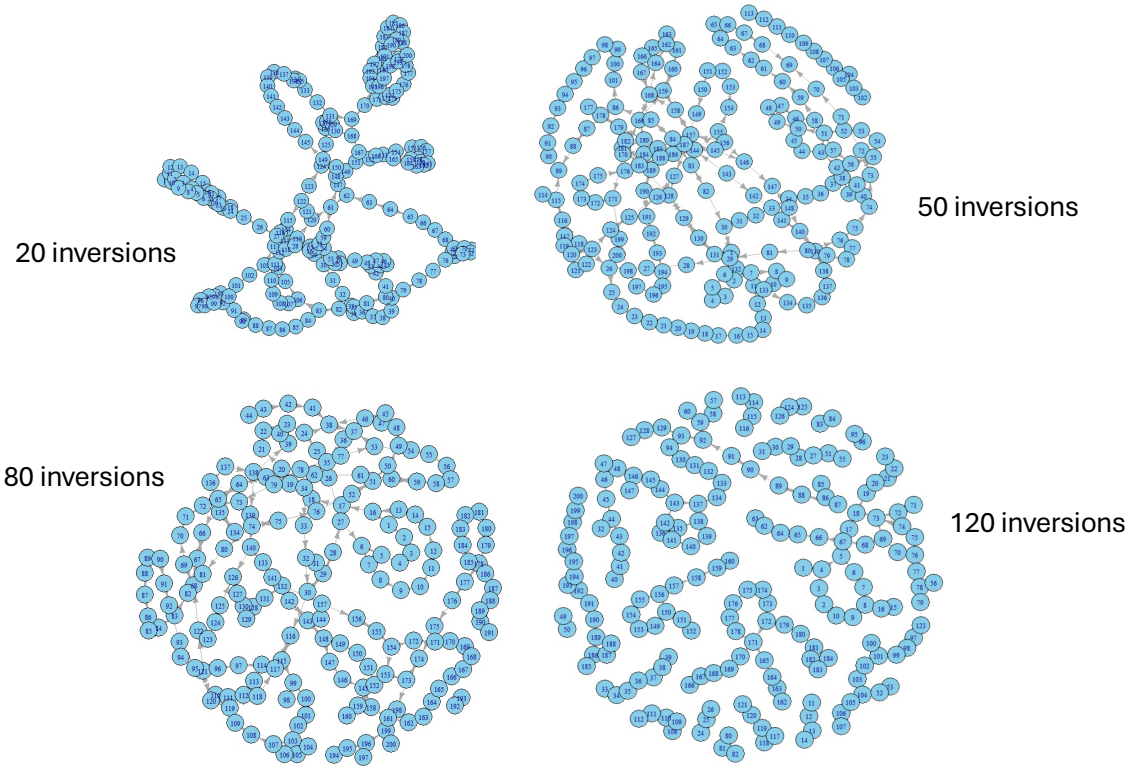

Figure 2.6: Inferred adjacency sets, as assembled from data generated by increasing numbers of inversions. Each gene occurs exactly once in each case, as is the convention in combinatorial work on reconstruction [39, 41, 42, 40, 43]. Genes in more than two adjacencies give rise to branchings that disrupt the linearity of the reconstruction.

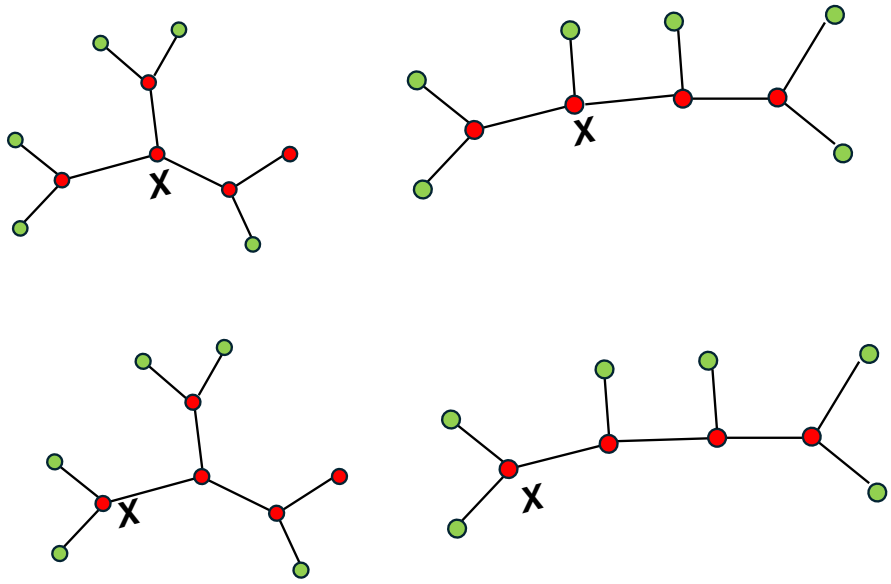

Figure 2.7: Tree shape and root position with six terminal vertices. The tree on top left was used for generation of the data and reconstruction. The same data was used for reconstruction of the tree on top right.

The computational steps in these analyses were hard-wired and required manually entering not only parameter values, not a large task, but also tree structures, a complex and time-consuming challenge. In contrast to molecular phylogeny, there is no reliable software available to handle gene-order phylogeny, and it was beyond the scope of the present project to construct a new general program for the large phylogeny problem for gene order and adjacencies. This could, however, provide part of a foundation for a long-term investigation.

2.9 Discussion

The most striking result from this work is the power of the phylogenetic validation criterion based on Dollo's principle to weed out the massive amounts of recently generated adjacency data to preserve the original gene order information in the original pangenome. One of the referees of the published version of the paper remarked "The paper is nice and presents a proof of concept of the possible usage of phylogenetic validation to filter adjacencies."

Dollo's law has receded in importance with the emergence of molecular phylogenetics and the modern understanding of the mechanics of evolutions [10]; it is largely confined to the study of curious changes in complex organs and is not pertinent to the analysis of protein or DNA evolution, where the amino acid residues or nucleotides

often undergo reversible mutations. The role of Dollo — or our interpretation of it in phylogenetic validation — is, however, key in tracing breakpoint introduction, which is only occasionally reversed.

Our model is extremely simple. Moreover, no suitable data exist to our knowledge for even a more relaxed and parameterized model. Nevertheless, we submit that we have showed proof of principle for a new approach to ancestral pangenome reconstruction, which is itself a new objective.

One of the most promising outcomes from this work is illustrated in Figure 2.6, where each 3-degree vertex violates the key tenet of the ancestral gene-order reconstruction literature as enunciated in [11]; namely, linearity of the reconstruction. How to peel these 3-degree vertices apart to reveal linear constituent genomes is a challenge for further work on pangenome reconstruction.

Chapter 3

A substitution model under Dollo

3.1 A Gene Substitution Model

In this chapter, we propose and analyze a probabilistic tree model to simulate the inheritance and replacement of discrete objects (such as genes or markers) along a hierarchical phylogenetic tree. Each object at the root node may be replaced by a novel object, as a way of implementing the mutation reversal prohibition embodied in Dollo's law [15], with a given probability as it is transmitted down each branch, modeling evolutionary turnover or innovation. In the context of pangenomics, this process models the evolution of the "cloud" or "shell" [13, 14, 12] portions of the gene complement of a pangenome. These involve genes that are present in only one or a few of the genomes making up a pangenome.

We provide theoretical analyses for the expected overlap and retention of ancestral objects across descendant nodes. Our work further investigates the reconstruction of ancestral object sets via the steinerization procedure, in which each internal node's objects are defined as those present in at least two out of three neighbouring nodes. Through simulation experiments, we compare the reconstructed sets with the true ancestral objects and quantify the effects of replacement probability, tree depth, and random convergence. The results demonstrate both the accuracy and limitations of theoretical formulas, highlighting the role of coincidental overlap among newly introduced objects and the challenges of ancestral reconstruction in highly dynamic systems. Our code framework enables flexible exploration of these processes, providing a foundation for more realistic modeling of pangenome and evolutionary scenarios.

3.2 Introduction

Reconstructing the history of transmitted features in tree-like evolutionary systems is a foundational task in fields ranging from evolutionary genomics to cultural evolu-

tion [60, 61]. Applications include the reconstruction of language families and sound changes in historical linguistics, from Swadesh's tree models [62, 63] to Bayesian phylogenetic analyses of language diversification [64, 65], textual criticism [66, 67] and the archaeology of stone tools and ceramic history [68, 69, 70, 71, 72].

Other subjects — bicycles [73], firearms [74] and aircraft [75] — have likewise been analyzed as evolving lineages in which design constraints, incremental innovations and functional trade-offs generate tree-like diversification patterns, as well as the reconstruction of musical-instrument lineages and divergence histories [76, 77, 78]. The historical development of technical standards — such as screw threads, railway gauges, electrical connectors, and communication protocols — often exhibits tree-like divergence [79].

In phylogenetic analyses, we are often given only the observed states of present-day entities (the leaves of a tree) and must infer the past — the composition of internal nodes or the root — given a stochastic process of loss and replacement along each branch. This "inverse" problem is crucial for tasks such as ancestral genome reconstruction, tracking the spread of information and studying innovation in languages or technology.

However, as features (genes, objects, or information units) are lost or replaced as we move down the tree, it becomes increasingly difficult to infer the true ancestral state. Standard approaches may rely on direct inheritance, but in realistic scenarios, replacement or mutation events are frequent, and majority rules may fail. Thus, simulation models combined with algorithmic reconstruction methods (such as steinerization) are essential to quantify what can actually be recovered and what information is irrevocably lost.

In this work, we present a flexible simulation-based study of object transmission and replacement in a fixed, multi-level tree, paired with a recursive intersection-based reconstruction (steinerization) of internal node states. We systematically analyze how the amount of retained and reconstructable ancestral information depends on the per-branch replacement rate and provide practical guidance for interpreting results of such reconstructions in real-world data.

3.3 Model Design

To parallel the experiments in the previous chapter, we focus here on the effects of evolutionary rates in improving or degrading the recovery of the ancestral pangenome. Here we model the rate effect in terms of a replacement probability p, while this effect was controlled in the previous chapter by the inversion rate. As with the latter, the loss of signal is amplified as the process proceeds along the branches of a phylogeny. Our experiments were carried out on pangenomes whose constituents had gene content $n = 200$. First, however, we experimented with obscuring much of the phylogenetic

structure by setting a smaller gene content, $n = 10$, allowing for greater overall variation than with $n = 200$, where average behaviour predominates.

3.3.1 Tree Structure

Our simulation operates on a fixed, rooted tree with five levels, as in Figure 3.1:

- **Root (A)**: The ancestor, containing n objects (e.g., $n = 10, n = 200$).
- **Level 1**: Three children (B, C, D) branching from A.
- **Level 2**: Each Level 1 node has two children (E, F from B; G, H from C; I, J from D).
- **Level 3**: Each Level 2 node has two children (K1–K12).
- **Level 4 (Leaves)**: Each K node has two leaves (L1–L24), for a total of 24 leaves.

This balanced design provides a controlled context for transmission, replacement, and recursive reconstruction, while being large enough to capture extensive variability.

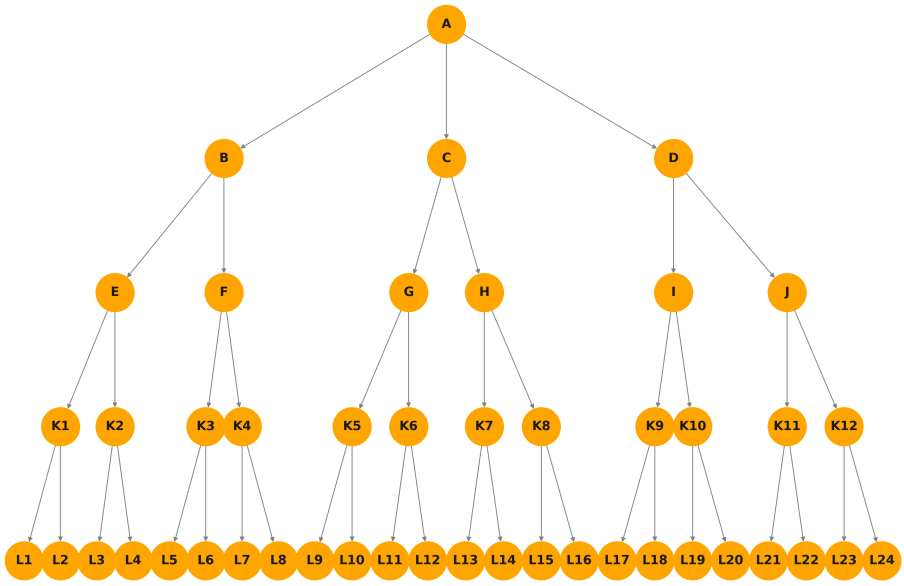

Figure 3.1: Simulation tree structure

3.3.2 Object Replacement Mechanism

Each node, except the root, initially inherits all objects from its parent. Then, for each object, with probability p, it is replaced by a new, unique object not present anywhere else in the tree. This models random irreversible innovation or mutation per feature per step (Dollo's law [15]).

For low values of p, most objects are inherited unchanged, mimicking high conservation. They are transmitted across several generations before being replaced. At high p, most are replaced, mimicking rapid loss or innovation.

3.3.3 Theoretical retention

For a path of h steps from A, the probability that an object from A survives in a given descendant is $(1-p)^h$. The expected count of such objects is $n(1-p)^h$. Overlap between two descendant nodes at different depths can be computed similarly, as explained below. This "expected retention" is our baseline for comparison against simulation and reconstruction.

3.4 Theoretical Analysis

3.4.1 Theoretical Analysis: 2-out-of-3 Overlap for Three Nodes

Suppose we have three nodes X, Y, and Z in the tree, where X is an ancestor of both Y and Z (but Y and Z are not ancestors of each other), with path lengths from X to Y and Z given by h_1 and h_2 respectively. The replacement probability per step is p, and n is the number of objects in X.

For any object originating in X:

- The probability it survives to Y is $(1-p)^{h_1}$.
- The probability it survives to Z is $(1-p)^{h_2}$.
- The probability it survives to both Y and Z is $(1-p)^{h_1+h_2}$ (since replacements on each path are independent).

We seek the expected number of such objects found in at least two out of the three nodes X, Y, Z as illustrated in Figure 3.2. Using inclusion-exclusion, the probability that a given object is present in at least two of X, Y, Z is:

$$P_{2/3} = (1-p)^{h_1} + (1-p)^{h_2} - (1-p)^{h_1+h_2}. \tag{3.4.1}$$

Thus, the expected number of objects in at least two of these nodes is:

$$E_{2/3}(X,Y,Z) = n\left[(1-p)^{h_1} + (1-p)^{h_2} - (1-p)^{h_1+h_2}\right]. \tag{3.4.2}$$

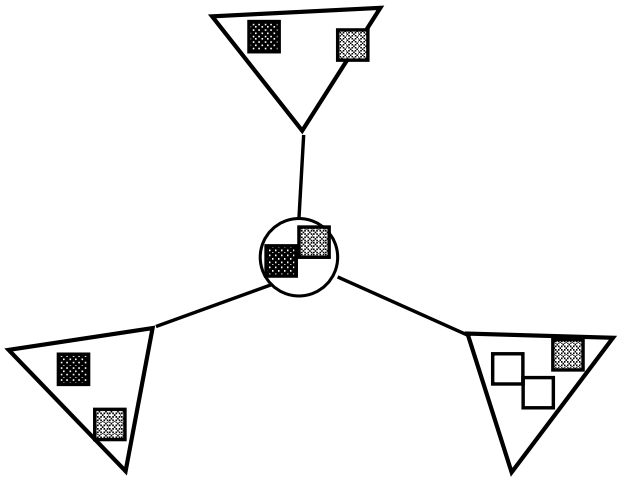

Figure 3.2: (Reproduced from Chapter 2) Necessary condition for adjacencies to appear at an internal vertex associated with an ancestral pangenome of a binary branching phylogenetic tree. Light shaded adjacency (small square) appears in all three trees (triangles) subtended by the internal vertex (circle). Dark shaded adjacency appears in only two of the trees. Unshaded adjacency appears in only one subtree so does not affect internal vertex. The shaded adjacencies are "phylogenetically validated" with respect to the internal vertex. The unshaded one is not validated. Adapted from [43]

Special case: When $h_1 = h_2 = h$, this reduces to

$$E_{2/3} = n\left[2(1-p)^h - (1-p)^{2h}\right]. \tag{3.4.3}$$

3.5 Simulation, reconstruction and visualization

3.5.1 Data Generation Algorithm

Our R simulation function operates as follows:

1. Initialize node A with $n = 200$ or $n = 10$ unique objects.
2. Recursively create all child nodes, applying the object-replacement rule at every branching step, with replacement probabilities $p = 0.05, 0.10, 0.20, 0.30, 0.50$.
3. Ensure each replaced object receives a unique identifier, maintaining global uniqueness.

Each node's object set and tree position is recorded for further analysis.

3.5.2 Computing Overlap and Similarity

For each simulation:

- Calculate, for any node X, the number of objects it shares with the true ancestor A ($|\text{objects}(X) \cap \text{objects}(A)|$).
- Compute pairwise overlaps for all node pairs, summarizing as a similarity matrix.
- Visualize this matrix as a heatmap to depict which nodes are most similar.

3.5.3 Visualization Techniques

We use the following R functions:

- `igraph` to draw the explicit tree structure. This R package provides tools for creating and visualizing graphs and networks, making it suitable for rendering hierarchical or phylogenetic trees.
- `pheatmap` for similarity matrices. This R package generates heatmaps with optional clustering, allowing us to visualize pairwise similarity or distance between nodes or genomes.
- `ggplot2` for plotting retention and overlap curves as functions of depth or replacement probability. It is a widely used data visualization package in R.

We also implemented the following custom R functions to simulate object inheritance and reconstruct ancestral nodes:

- `simulate_custom_tree_objects` to simulate a hierarchical tree of objects with probabilistic replacement. Each node receives a copy of its parent's object set, with each object having probability P of being replaced by a new unique object. This version builds a generic binary tree to arbitrary depth.
- `simulate_named_tree_objects_with_map` is an enhanced version that assigns standard node labels (e.g., A, B, C, D, E, ..., L24) and returns both the object tree and a node name mapping. It facilitates downstream reconstruction and analysis.
- `get_nodes_at_level` retrieves all node names at a specified tree depth. It is used for selecting valid triplets of nodes across different layers.
- `is_ancestor` determines if one node is an ancestor of another by testing prefix containment (e.g., B is an ancestor of B-1).

- `overlap2of3` computes the set of objects that appear in at least two of three input vectors. This function is the core of all steinerization procedures.
- `steiner_one_round` performs one full round of steinerization. It reconstructs nodes K1–K12, E–J, B–D, and A based on the current object sets at the lower nodes using `overlap2of3`.
- `run_steinerization_fully_correct` iteratively applies `steiner_one_round` and tracks the reconstructed node `A` across rounds until convergence.
- `run_steinerization_true_window` implements a variant of steinerization in which the leaf nodes `L1--L24` remain unchanged. This "fixed-window" version allows us to isolate the propagation behavior through internal levels.
- `print_steiner_round` and `update_leaf_from_K` are helper functions for step-by-step steinerization. The former prints object sizes at each level; the latter updates the leaves based on reconstructed K nodes.

3.6 Similarity analysis and tree reconstruction

Effect of Object Count on Similarity Matrix Stability

To investigate the impact of the number of objects per node on similarity patterns among nodes, we simulated tree structures with the same replacement probability ($P = 0.2$), but with different object counts: $n = 200$ and $n = 10$.

Effect of Object Set Size on Similarity Matrix Stability

To investigate how the number of objects per node affects the stability of the similarity matrix and the recoverability of the tree structure, we compare two scenarios: one with $n = 200$ objects per node and another with $n = 10$.

In the first case, when each node contains $n = 200$ objects (see Figure 3.3), the resulting similarity heatmap is much clearer and structurally meaningful. Several high-similarity blocks align well with subtrees of the original simulated tree, particularly among sibling nodes. This indicates that with a sufficiently large number of objects, the shared object proportion between nodes becomes stable and can accurately reflect the hierarchical structure of the tree.

In contrast, when each node contains only $n = 10$ objects (see Figure 3.4), the similarity heatmap appears noisy and less interpretable. Due to the small object set, random loss or replacement of just a few objects can cause large fluctuations in pairwise similarity scores. As a result, nodes that should be similar may appear

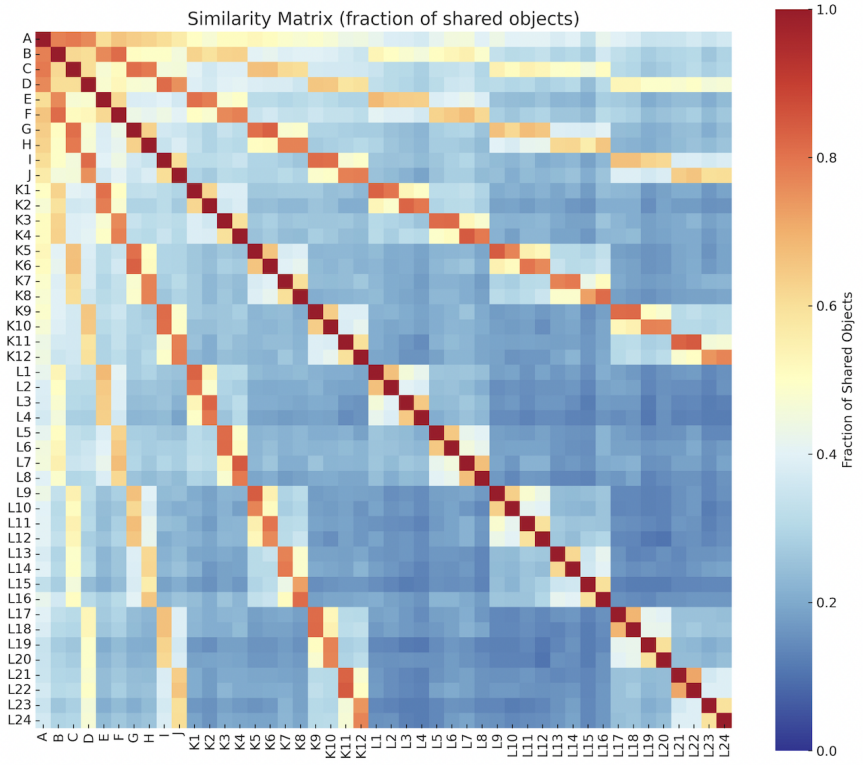

Figure 3.3: Heatmap of the similarity matrix when each node contains $n = 200$ objects. The increased number of objects per node stabilizes the similarity scores and reveals distinct similarity blocks corresponding to the tree hierarchy, especially among sibling nodes. This matrix preserves the structure of the simulated tree.

dissimilar, and some distant nodes may coincidentally show high similarity, obscuring the underlying tree topology.

In summary, using a larger number of objects per node (e.g., $n = 200$) helps stabilize similarity estimates and improves the accuracy of tree reconstruction methods such as neighbour-joining. On the other hand, when the object set is small (e.g., $n = 10$), the similarity matrix is more susceptible to randomness and fails to reflect the true tree structure.

We further investigated the effect of object set size per node on the accuracy of tree reconstruction using the neighbour-joining algorithm. Two similarity trees were generated from simulated node data using pairwise Jaccard similarity and plotted via the neighbour-joining method. The only difference between the two scenarios is the number of objects per node: one with $n = 200$, the other with $n = 10$.

The tree reconstructed from nodes with $n = 200$ objects (Figure 3.5) exhibits a

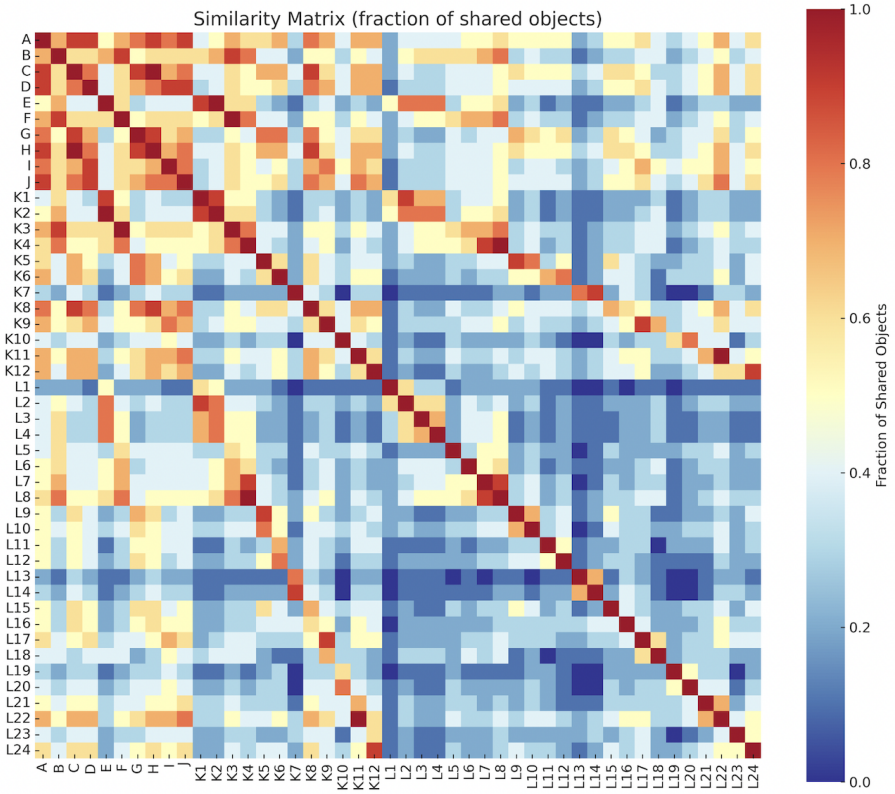

Figure 3.4: Heatmap of the similarity matrix when each node contains $n = 10$ objects. The similarity between each pair of nodes is calculated as the fraction of shared objects. Due to the small number of objects per node, the similarity scores are highly variable, resulting in an unstable and noisy matrix that fails to clearly reflect the tree structure.

topology that closely mirrors the original hierarchical structure. Sibling leaf nodes are correctly grouped under their simulated parents (e.g., L_{23} and L_{24} under K_{12}), and the major branches of the tree are preserved. This illustrates that larger object sets improve similarity estimation, leading to a more accurate and robust tree structure.

In contrast, when $n = 10$, the reconstructed tree (Figure 3.6) shows substantial topological deviations from the original simulated tree. Multiple internal nodes are misplaced, and leaf nodes originating from the same parent are often grouped under incorrect clusters. This is due to the high variance in similarity scores caused by the small number of objects, making the pairwise similarity matrix unstable and unreliable.

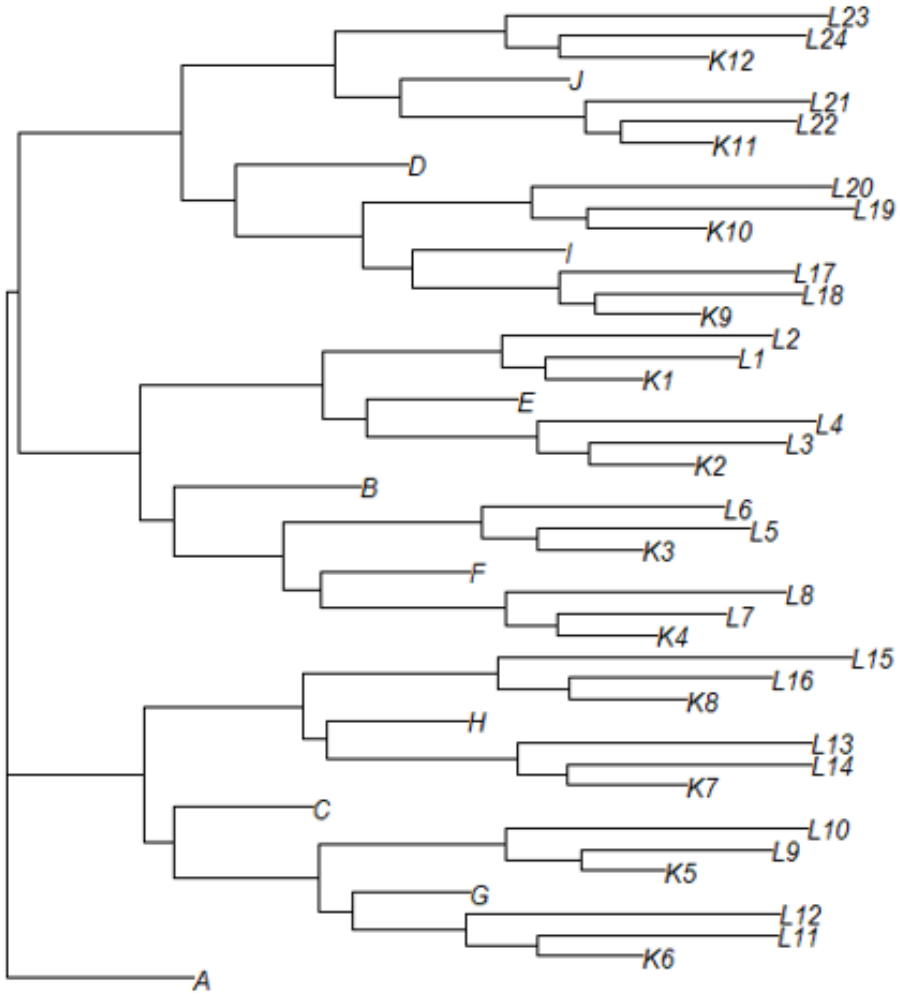

Figure 3.5: Similarity tree reconstructed using neighbour-joining with $n = 200$ objects per node. The structure more closely resembles the original tree, reflecting improved stability and clustering accuracy with larger object sets.

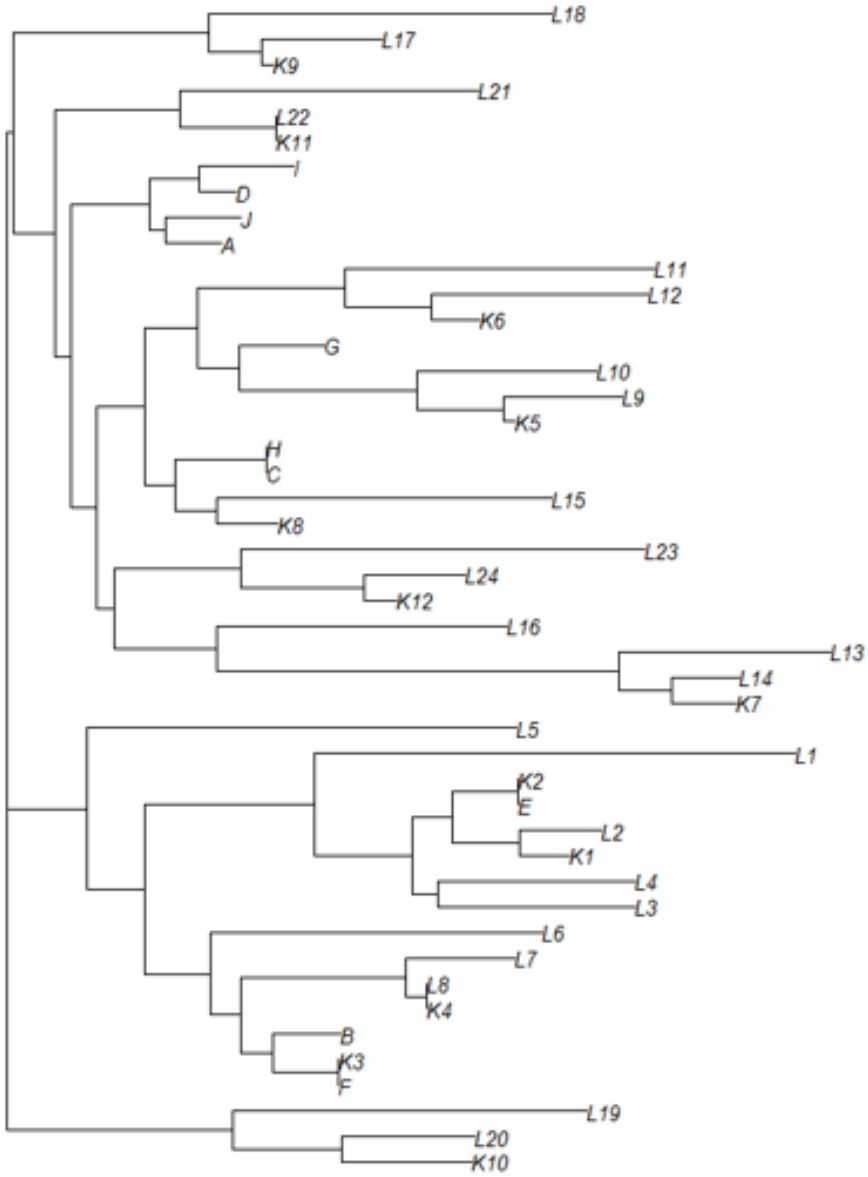

Figure 3.6: Similarity tree reconstructed using neighbour-joining with $n = 10$ objects per node. The tree exhibits substantial topological deviation from the simulated tree due to high noise in the similarity estimates.

3.7 Comparison of Simulation and Theory: Overlap Decay with Tree Distance

Figure 3.7 compares the observed (simulation) and predicted (theoretical) number of shared objects between pairs of nodes as a function of their tree distance (h), for various object replacement probabilities (p). Each colour denotes a distinct p value, with solid lines representing simulation data and dashed lines representing theoretical predictions.

- **Overlap decay:** For all values of p, the overlap decreases rapidly as h increases. This illustrates how, as the distance between nodes grows, random replacements along each path erode the retention of original objects.
- **Impact of p:** Lower values of p (e.g., $p = 0.1$) result in higher overlaps, since objects are less likely to be replaced at each step. Larger p values accelerate the loss of overlap.
- **Simulation vs. theory:** Theoretical predictions closely follow the simulation results.

Overall, the agreement validates the theoretical model for shallow trees and small p, while highlighting the emergence of new-object overlaps in deeper or more dynamic evolutionary scenarios.

3.8 Recursive steinerization reconstruction

3.8.1 The inverse problem

In the usual biological systematics context, only the leaves (L1–L24) could be observed. Can we reconstruct the object set of internal nodes (especially the root A) from just the leaves? This inverse problem is fundamental to evolutionary biology and related fields.

3.8.2 Steinerization heuristic

Steinerization is a recursive algorithm that reconstructs the internal node's object set as those present in at least two out of its three connected neighbours (children or sliding window group). At each round:

1. For each internal node, compute the "at least 2 of 3" intersection from its descendants or neighbour sets.
2. Iteratively move up the tree, using newly reconstructed nodes as input for the next round.

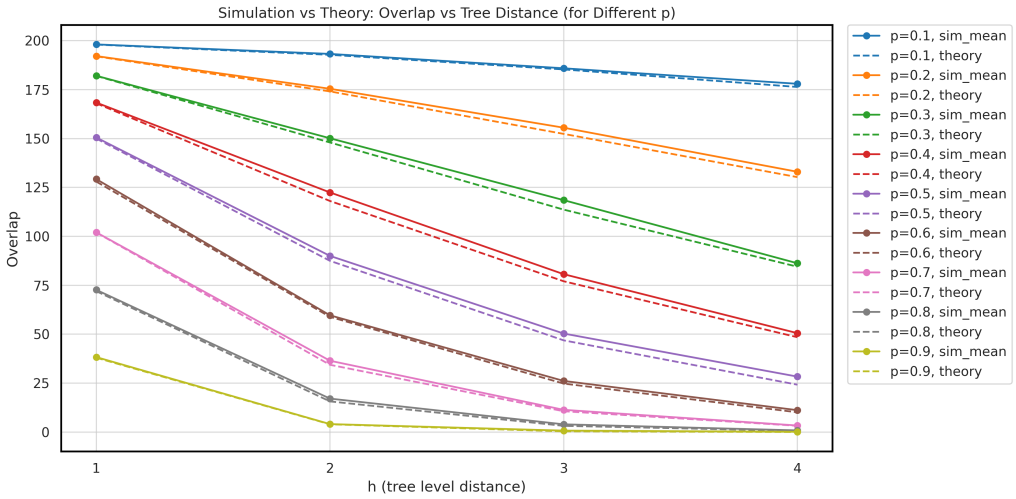

Figure 3.7: Simulation vs. theory: Number of shared objects (overlap) as a function of tree distance h, for various object replacement probabilities p. Solid lines represent simulation results, and dashed lines represent theoretical expectations computed using equation (3.4.2).

3.8.3 Full Procedure

Initialization: The tree is simulated with a depth of 4 layers, with nodes labeled as follows:

- Leaf nodes: L1 to L24 (bottom layer)
- Internal nodes: K1 to K12, E to J, B to D
- Root: A

Full Procedure with Iterative Sliding Window

We divide the reconstruction process into two main phases:

Phase I: Initial Upward Reconstruction (Rounds 1–4)

- **Round 1 (Leaf → K nodes):**
  - Each K_i is computed from three consecutive leaf nodes:

$$K_1 = \texttt{overlap2of3}(L_1, L_2, L_3)$$
$$K_2 = \texttt{overlap2of3}(L_3, L_4, L_5)$$
$$\vdots$$
$$K_{12} = \texttt{overlap2of3}(L_{23}, L_{24}, L_1)$$

- **Round 2 (K → E-J nodes):**
  - Each intermediate node E to J is reconstructed:

$$\begin{aligned} E &= \texttt{overlap2of3}(K_1, K_2, K_3) \\ F &= \texttt{overlap2of3}(K_3, K_4, K_5) \\ G &= \texttt{overlap2of3}(K_5, K_6, K_7) \\ H &= \texttt{overlap2of3}(K_7, K_8, K_9) \\ I &= \texttt{overlap2of3}(K_9, K_{10}, K_{11}) \\ J &= \texttt{overlap2of3}(K_{11}, K_{12}, K_1) \end{aligned}$$

- **Round 3 (E-J → B, C, D):**
  - Higher-level nodes reconstructed:

$$\begin{aligned} B &= \texttt{overlap2of3}(E, F, G) \\ C &= \texttt{overlap2of3}(G, H, I) \\ D &= \texttt{overlap2of3}(I, J, E) \end{aligned}$$

- **Round 4 (B, C, D → A):**
  - The first reconstructed version of root node A:

$$A^{(1)} = \texttt{overlap2of3}(B, C, D)$$

Phase II: Iterative Refinement (Round 5+)

- **Round 5 onward: Sliding window update using previous round's internal nodes.**
- Example updates:
  - Update K using: two leaf nodes + previous E:

$$K_1^{(2)} = \texttt{overlap2of3}(L_1, L_2, E^{(1)})$$

  - Update E using: two new K and previous B:

$$E^{(2)} = \texttt{overlap2of3}(K_1^{(2)}, K_2^{(2)}, B^{(1)})$$

  - Update B using: new E, F, and old A:

$$B^{(2)} = \texttt{overlap2of3}(E^{(2)}, F^{(2)}, A^{(1)})$$

– Update root:

$$A^{(2)} = \texttt{overlap2of3}(B^{(2)}, C^{(2)}, D^{(2)})$$

- **Convergence criterion:** Process stops when

$$A^{(t)} = A^{(t-1)}$$

This recursive process continues until the reconstructed A node stabilizes; i.e., when $A^{(t)} = A^{(t-1)}$.

3.9 Experimental Setup

3.9.1 Parameters

- **Number of objects per node**: 200
- **Replacement probabilities tested**: 0.05, 0.10, 0.20, 0.30, 0.50
- **Tree topology**: As described above

3.9.2 Metrics Collected

For each simulation and each value of p, we record:

- The number of original ancestor (A) objects present in selected nodes (L1, K1, E, B, etc.).
- The size of the reconstructed A node after steinerization, and its overlap with true A.
- The standard deviation across replicates for each value.

3.9.3 Data Presentation

Table 3.1: Summary of object counts for different p values

p	A_{true}	A_{steiner}	A_{overlap}
0.05	200	189	189
0.10	200	163	163
0.20	200	78	78
0.30	200	15	15
0.50	200	3	3

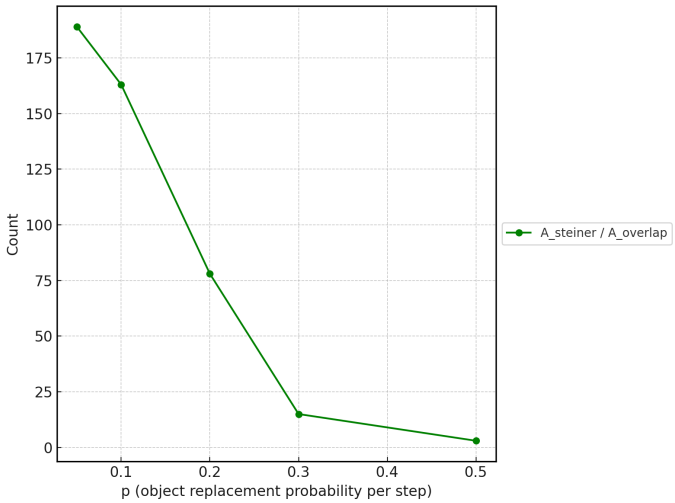

Figure 3.8: Reconstruction accuracy as a function of object replacement probability. The plot shows the number of objects in the reconstructed A node after steinerization (A_{steiner}) and their overlap with the original ancestral node (A_{overlap}), averaged over multiple simulations. As the replacement probability p increases, the number of correctly recovered ancestral objects drops sharply. This highlights the detrimental effect of high replacement rates on reconstruction accuracy.

Table 3.1 is comparable to Table 2.1 in the previous chapter, showing how the degree of recovery of objects in the ancestral pangenome depends on the evolutionary rate.

3.9.4 Analysis.

Table 3.1 and Figure 3.8 illustrate the effect of the replacement probability p on the ability to reconstruct the ancestral state using steinerization. When p is low, the reconstructed A node retains almost all the original objects, and the overlap is nearly perfect. However, as p increases, both the number of reconstructed objects and their overlap with the true ancestor drop precipitously. For high p (e.g., $p = 0.5$), nearly all memory of the original ancestor is lost, and almost no ancestral objects are recovered. These results demonstrate that the efficacy of content-based ancestral reconstruction depends critically on the rate of object turnover in the system.

As shown in Table 3.2 and Figure 3.9, the overlap with the original root node A systematically decreases as the object replacement probability p increases. For low values of p, most original objects are retained even in distant nodes. As p increases, the expected overlap falls rapidly due to the cumulative effect of object loss at each step.

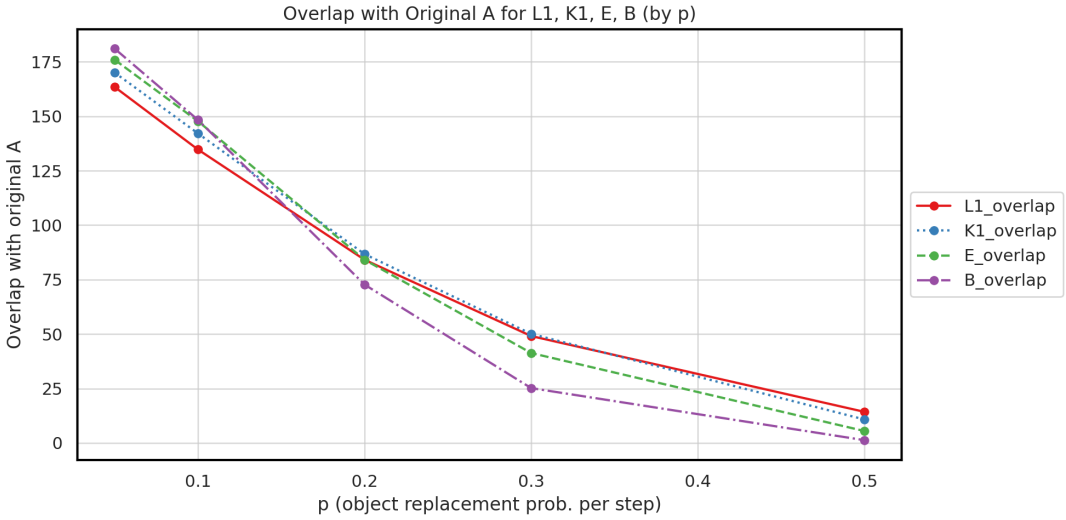

Figure 3.9: Overlap with original A for L1, K1, E, B as a function of object replacement probability p.

Table 3.2: Average overlap between selected nodes and root.

p	L1_overlap	K1_overlap	E_overlap	B_overlap
0.05	163.4	170.0	175.9	181.0
0.10	134.7	142.1	147.7	148.3
0.20	84.1	86.7	84.1	72.7
0.30	49.1	50.1	41.3	25.2
0.50	14.3	10.8	5.5	1.3

In this analysis, we compared the overlap between the true root objects (A) and various reconstructed or descendant nodes (L1, K1, E, B) after applying the steinerization procedure to simulated tree-structured data. Somewhat unexpectedly, we observed that the reconstructed node B — despite being the direct child of A — sometimes exhibits a *lower* overlap with A than more distant nodes (such as K1 or L1), particularly as the object replacement probability p increases.

This counterintuitive phenomenon is explained by the global merging strategy of the steinerization algorithm. During reconstruction, the B node is not merely inheriting objects directly from its parent (A); instead, it combines noisy or highly replaced information from both its own descendants and sibling branches. When the replacement probability p is high, this merging process dilutes the direct signal from A, reducing the observed overlap.

In contrast, the overlap between A and single-path descendant nodes (such as L1 or K1) depends only on the direct retention probability along a single lineage. This follows the expected theoretical decay of overlap:

$$\mathbb{E}[\text{overlap}(A, \text{descendant at depth } h)] = n(1-p)^h, \tag{3.9.1}$$

where n is the initial number of objects in A.

Overall, these results highlight that global reconciliation methods, while effective for recovering deep ancestral content, may introduce additional "noise" at shallow reconstructed nodes, especially when object turnover is high. Therefore, overlaps for reconstructed ancestral nodes should be interpreted with caution, as they reflect a blend of multiple evolutionary paths rather than simple direct inheritance.

3.10 Results

3.10.1 Raw Retention

Simulation results match the theoretical expectation that ancestor object retention decays exponentially with tree depth and with increasing replacement probability.

3.10.2 Steinerization performance

- At low p, reconstructed A nodes via steinerization capture a large fraction of the original A's objects.
- As p increases, the reconstructed A's size and overlap drop sharply.
- Interestingly, the non-monotonicity appears: occasionally, a deeper node (e.g., L node) shows higher overlap with A than an intermediate node (e.g., B or K1). This is due to stochastic effects and the combinatorics of the sliding window/intersection rule.

3.10.3 Visualization

- **Overlap vs. p**: Overlap curves clearly show decreasing trends with increasing p, and differences among node types.
- **Heatmaps**: Visualize block-diagonal structure reflecting tree hierarchy.
- **Simulation vs. Theory**: Simulation averages match theory at low p, but reconstruction diverges at higher p.

3.11 Discussion

The work in this chapter was motivated by a different kind of question about the evolution of pangenomes. In the previous chapter, we focused on the core of the pangenome. In that context, gene content is relatively constant across the constituent genomes making up the pangenome, but the order of the genes was subject to rearrangements. The core gene content remained constant throughout the pangenome, and the shell genes were not considered. In the present chapter, it is the core that is put aside in order to concentrate on processes that affect the shell genes.

A clear tendency that emerged from this work is that, at least for the phylogenetic configuration we studied, there is a range of values of the rate parameter p that allow for the recovery of most of the genes in the shell genome. Admittedly, the constituent genomes were some orders of magnitude shorter than real genomes, and the tree structure we used was neither deep nor highly ramified. Nevertheless, our simulations showed that enough of a signal was transmitted from the ancestor to the genomes at the leaf of a tree that the ancestor shell could be largely reconstructed via our validation criterion.

The gene-replacement model we adopted was largely dictated by the fact that accessory genes do not come from a limited fixed pool. Even if a new gene has some module or substructure in common with a core gene, there are thousands of core genes, which means that, for all intents and purposes, the genes that are dropped have no functional or historical connection with the new additions to the shell. The fact that we used a one-to-one replacement mechanism was only to keep the number of genes constant over several algorithmic steps, not for any biological motive.

In our concentration on rates of replacement, we only briefly explored other aspects of the phylogenetic context: size of the genome ($n = 10$ instead of $n = 200$) and reconstruction accuracy. We basically used a single tree as a testing ground. Both of these tentative efforts would warrant more investigation in a longer-term project. The current model assumes independent replacement and a fixed tree. Real systems may have correlated replacement, convergent evolution or variable topology. Extending this framework to more realistic scenarios would be a worthwhile project.

In sum, we have presented a simulation and recursive reconstruction framework for studying how much of an ancestral object set can be retained and recovered in a tree with stepwise stochastic replacement. Our results clarify the promise and limits of reconstruction using intersection-based (steinerization) rules and hopefully provide a foundation for more advanced ancestral inference methods.

Chapter 4

Conclusions and further work

Lacking any previous work on reconstructing ancestral pangenomes, we have introduced two experimental approaches, one exploring a model of structural variation in the core of the pangenome, the other a model of loss and substitution pertinent to the accessory genes of the pangenome: the shell and the cloud.

The goal of reconstructing a pangenome is hard to achieve in the context of current methodologies that embody, even in the context of gene duplication and paralogy, the fact that a genome must have a linear, non-branching structure. With the technical exception of circular DNA, mostly in prokaryotes, this linearity is fundamental. It is enunciated most clearly in the little-known or cited article of Tannier et al. [11] in which the major criterion advocated for any reconstruction method is the minimization of branching structure caused, for example, by conflicting adjacencies. In contrast, the production of conflicting adjacencies in our simulations provides a key to inferring several distinct constituent genomes in the ancestral pangenome. In future work, this could involve calculating the multiplicity of reconstructed adjacencies, both compatible and conflicting, where large runs of high-multiplicity adjacencies would indicate segments of the core common to all the individual constituent genomes.

The most striking result from this work is the power of the phylogenetic validation criterion based on Dollo's principle to weed out the massive amounts of recently generated adjacency data to preserve the original gene order information in the original pangenome. Dollo's law is rarely invoked these days, but its extension to our concept of phylogenetic validation is also implicit in much evolutionary research at the level of complex molecular structures.

Our model is simple. No suitable data exists to our knowledge for even a more relaxed and parameterized model. Nevertheless, I submit that I have shown proof of principle for a new approach to ancestral pangenome reconstruction, which is itself a new objective.

The modelling in Chapter 3 does not explicitly involve multi-genome pangenomes as input, but the fixed-size genomes under the Dollo regime targets the evolution of the

accessory genes in pangenomes. The calculation and simulations of rates should help understand the frequency distributions of genes in the shell and cloud proportions of the non-core gene complement. Part of the innovation in this chapter is that instead of focusing on given extant genomes, it is concerned with how fast ancestral genes disappear in descendants, which suggests new ways of interpreting the fate of shell genes in ancestral pangenomes. Thus we have presented a simulation and recursive reconstruction framework for studying how much of an ancestral object set can be retained and recovered in a tree with per-step stochastic replacement. Our results clarify the limits of reconstruction using intersection-based (steinerization) rules and hopefully provide a foundation for more advanced ancestral inference methods.

There are important limitations to this work. Most of these are due to the small scale of our exploratory models. The current model in Chapter 3 assumes independent replacement and a fixed tree. More realistic systems may allow correlated replacement, convergent evolution, variable rates or variable topology. Extending this framework to these more ramified scenarios is a likely direction for further work.

The exclusive use of adjacency data in Chapter 2 assumes that the orthology groups (identity of gene correspondences) are given — genes labelled with the same number in different species are treated as if they are identical, whereas in reality, the nucleotide sequences may differ substantially. In practice, the construction of orthology groups across several genomes is a non-trivial task. In our focus on gene-order structure in genomes and pangenomes, we assume this task has already been accomplished.

Dollo's law is simply that evolutionary changes are irreversible and is simplistic in the light of modern knowledge. In my research, however, it is simply a way of understanding the principle of phylogenetic validation. The two models in Chapters 2 and 3 both rely on this principle of phylogenetic validation. Both make use of it as part of the steinerization protocol of recursively improving tree structure. Phylogenetic validation and steinerization will likely remain important aspects of ancestral-pangenome inference.

The context of gene substitution involving a pool of tens of thousands of different genes means Dollo and phylogenetic validation are acceptable assumptions in our model in Chapter 3. The phylogenetic validation principle, however, could be weakened to allow for a low rate of reverse mutation, in a future larger-scale model, though this would involve a much more complicated inferential apparatus.

Our model in Chapter 2 assumes that no gene has multiple copies in a single genome. This is well-justified in the single-genome ancestral reconstruction [39, 40, 41, 42, 43] tradition, based on the assumption that genome doubling and extensive paralogy is a feature of extant genomes and their recent ancestors. In the ancestral pangenome reconstruction context relaxing this would add another difficulty, interfering with the multiplicities solution proposed above.

The constant genome size assumption in Chapter 3 allows straightforward com-

putation of the joint probabilities of gene occurrence in two or more ancestors. I dealt exclusively with closed pangenomes. I did not enter the topic of determining whether pangenomes are open (involving no limit to the number of accessory or cloud genes, up with hundreds or thousands of genes, as with many bacterial species) or closed.

The two projects in this thesis represent initial approaches to the problem of ancestral pangenome gene-order reconstruction, including the inference about the multiplicity of the constituent genomes. These were only the first steps; in the first project, while we succeeded in recovering most of the adjacencies present in our given ancestor pangenome, the nature and even the number of constituent genome could not be ascertained, but at least their existence was signaled by the number of genes with more than two adjacencies. In the second project, the best we could do was to demonstrate that Dollo's law, in our reformulation as phylogenetic validation, succeeded in recovering a good number of the original genomic elements; even though multiple replacement events in a lineage would result the the extinction of a gene, the branching structure of the phylogeny makes it likely that it will survive in at least one lineage.

These modest contributions are suggestive of one promising approach; in both projects, we did not take into account how often an adjacency or a gene was recovered, but this emerges as a likely strategy. If several items are recovered at the same high multiplicity m, this suggests that they all are part of the core genome of m constituent genomes. Hopefully this insight will inspire further advances to the challlenge of reconstructing an ancestral pangenome.

Bibliography

[1] Eizenga JM, Novak AM, Sibbesen JA, Heumos S, Ghaffaari A, Hickey G, et al. Pangenome graphs. *Annu Rev Genom Hum Genet*. 2020;21:139–62.

[2] Zheng C, Lenert A, Sankoff D. Reversal distance for partially ordered genomes. *Bioinformatics*. 2005;21:502–8.

[3] Zheng C, Sankoff D. Genome rearrangements with partially ordered chromosomes. In: *International Computing and Combinatorics Conference. Lect Notes in Comp Sci*. 2005; 3595: 52–62.

[4] Tettelin H, Medini D. *The Pangenome: Diversity, Dynamics and Evolution of Genomes*. Cham: Springer; 2020.

[5] Zhou X, Sankoff D. Ancestral pangenomes and their phylogenetic reconstruction. *Lect Notes Comput Sci*. 2026;15666:141–9.

[6] Maddison WP, Knowles L, Lacey T. Inferring phylogeny despite incomplete lineage sorting. *Syst Biol*. 2006;55:21–30.

[7] Eigen M, Schuster P. A principle of natural self-organization. *Naturwissenschaften*. 1977;64:541–65.

[8] Brockhurst MA, Harrison E, Hall JP, Richards T, McNally A, MacLean C. The ecology and evolution of pangenomes. *Curr Biol*. 2019;29:R1094–103.

[9] Wikipedia. Pan-genome. Available from: `https://en.wikipedia.org/w/index.php?title=Pan-genome&oldid=1314668163`. Accessed November 30, 2025.

[10] Kathryn R. Elmer KR, Clobert J, Dollo's law of irreversibility in the post-genomic age. *Trends in Ecology & Evolution*. 2025; 40: 136–146.

[11] Tannier E, Bazin A, Davín A, Guéguen L, Bérard S, Chauve C. Ancestral genome organization as a diagnosis tool for phylogenomics. In: *Phylogenetics in the Genomic Era*. 2020. p. 2–5.

[12] Lassalle F, Didelot X. Bacterial microevolution and the pangenome. In: *The Pangenome*. Cham: Springer; 2020.

[13] Cummins EA, Hall RJ, McInerney JO, McNally A. Prokaryote pangenomes are dynamic entities. *Curr Opin Microbiol*. 2022;66:73–8.

[14] Hyun JC, Palsson BO. Reconstruction of the last bacterial common ancestor from 183 pangenomes reveals a versatile ancient core genome. *Genome Biol*. 2023;24:183.

[15] Dollo L. Les lois de l'évolution. *Bull Soc Belge Geol Paleontol Hydrol*. 1893;7:164–166.

[16] Tantoso E, Eisenhaber B, Kirsch M, Eisenhaber F, et al. To kill or to be killed: pangenome analysis of *Escherichia coli* strains reveals a tailocin specific for pandemic ST131. *BMC Biol*. 2022;20:146.

[17] Qin P, Lu H, Du H, Wang H, Chen W, Chen Z, et al. Pan-genome analysis of 33 genetically diverse rice accessions reveals hidden genomic variations. *Cell*. 2021;184:3542–58.

[18] Gerdol M, Moreira R, Cruz F, et al. Massive gene presence-absence variation shapes an open pan-genome in the Mediterranean mussel. *Genome Biol*. 2020;21:275.

[19] Gong Y, Li Y, Liu X, et al. A review of the pangenome: how it affects understanding of genomic variation, selection and breeding in domestic animals. *J Anim Sci Biotechnol*. 2023;14:73.

[20] Liao WW, Asri M, Ebler J, et al. A draft human pangenome reference. *Nature*. 2023;617:312–24.

[21] Chauhan SM, Ardalani O, Hyun JC, Monk JM, Phaneuf PV, Palsson BO. Decomposition of the pangenome matrix reveals structure in gene distribution in the *Escherichia coli* species. *mSphere*. 2025;10:e00532–24.

[22] Depuydt L, Renders L, Abeel T, et al. Pan-genome de Bruijn graph using the bidirectional FM-index. *BMC Bioinformatics*. 2023;24:400.

[23] Hickey G, Monlong J, Ebler J, Novak AM, Eizenga JM, Gao Y, et al. Pangenome graph construction from genome alignments with Minigraph-Cactus. *Nat Biotechnol*. 2024;42:663–73.

[24] Page AJ, Cummins CA, Hunt M, Wong VK, Reuter S, Holden MT, et al. Roary: rapid large-scale prokaryote pan genome analysis. *Bioinformatics*. 2015;31:3691–3.

[25] Tonkin-Hill G, MacAlasdair N, Ruis C, et al. Producing polished prokaryotic pangenomes with the Panaroo pipeline. *Genome Biol.* 2020;21:180.

[26] Li H, Marin M, Farhat MR. Exploring gene content with pangene graphs. *Bioinformatics.* 2024;40:btae456.

[27] Linnaeus Carl *Systema Naturae per Regna Tria Naturae, Secundum Classes, Ordines, Genera, Species, cum Characteribus, Differentiis, Synonymis, Locis.* 10th ed. Vol. I: Regnum Animale. Holmiae (Stockholm): Laurentii Salvii; 1758.

[28] Darwin Charles *On the Origin of Species by Means of Natural Selection.* London: John Murray; 1859.

[29] Sokal RR, Sneath PHA. *Principles of Numerical Taxonomy.* San Francisco: W. H. Freeman; 1963.

[30] Farris JS. Methods for computing Wagner trees. *Syst Zool.* 1970;19(1):83–92.

[31] Felsenstein J. *Inferring Phylogenies.* Sunderland (MA): Sinauer Associates; 2004.

[32] Huelsenbeck JP, Ronquist F. MRBAYES: Bayesian inference of phylogenetic trees. *Bioinformatics.* 2001;17(8):754–755.

[33] Saitou N, Nei M. The neighbor-joining method: a new method for reconstructing phylogenetic trees. *Mol Biol Evol.* 1987;4(4):406–425.

[34] Zuckerkandl E, Pauling L. Molecular disease, evolution, and genic heterogeneity. In: Kasha M, Pullman B, editors. *Horizons in Biochemistry.* New York: Academic Press; 1962. p. 189–225.

[35] Dayhoff MO. Computer analysis of protein evolution. *Sci Am.* 1969;221(1):86–95.

[36] Sankoff D, Cedergren RJ. Simultaneous comparison of three or more sequences related by a tree. *Nature New Biol.* 1973;245(147):232–235.

[37] Sankoff D. Minimal mutation trees of sequences. *SIAM J Appl Math.* 1975;28(1):35–42.

[38] Sankoff D. Minimal mutation trees of sequences. *SIAM J Appl Math.* 1975;28(1):35–42.

[39] Xu Q, Jin L, Zheng C, Leebens-Mack JH, Sankoff D. RACCROCHE: ancestral flowering plant chromosomes and gene orders based on generalized adjacencies and chromosomal gene co-occurrences. *Lect Notes Comput Sci.* 2021;12686:97–115.

[40] Xu Q, Jin L, Zheng C, Zhang X, Leebens-Mack JH, Sankoff D. From comparative gene content and gene order to ancestral contigs, chromosomes and karyotypes. *Sci Rep*. 2023;13:6095.

[41] Xu Q, Jin L, Leebens-Mack JH, Sankoff D. Validation of automated chromosome recovery in the reconstruction of ancestral gene order. *Algorithms*. 2021;14:160.

[42] Xu Q, Zhang X, Zhang Y, Zheng C, Leebens-Mack JH, Jin L, Sankoff D. The monoploid chromosome complement of reconstructed ancestral genomes in a phylogeny. *J Bioinform Comput Biol*. 2021;19:6.

[43] Xu Q, Sankoff D. Gene order phylogeny via ancestral genome reconstruction under Dollo. *Lect Notes Bioinform*. 2023;13883:100–11.

[44] Muffato M, Louis A, Nguyen NTT, et al. Reconstruction of hundreds of reference ancestral genomes across the eukaryotic kingdom. *Nat Ecol Evol*. 2023;7:355–66.

[45] Perrin A, Varré J-S, Blanquart S, Ouangraoua A. Procars: progressive reconstruction of ancestral gene orders. *BMC Genomics*. 2015;16(Suppl 5):6.

[46] Loegler V, Friedrich A, Schacherer J. Dynamics of genome evolution in the era of pangenome analysis, *Cell Genomics*. 2025: 101067.

[47] Sankoff D, Rousseau P. Locating the vertices of a Steiner tree in an arbitrary metric space. *Math Program*. 1975;9(1):240–246.

[48] Pagel M. Detecting correlated evolution on phylogenies. *Proc R Soc B*. 1994;255:37–45.

[49] Pagel M. The maximum likelihood approach to reconstructing ancestral character states of discrete characters. *Syst Biol*. 1999;48:612–22.

[50] Treangen TJ, Rocha EPC. Horizontal transfer, not duplication, drives the expansion of protein families in prokaryotes. *Nat Rev Microbiol*. 2011;9:605–10.

[51] Daubin V, Moran NA, Ochman H. Phylogenetics and the cohesion of bacterial genotypes. *Genome Res*. 2003;13:1704–11.

[52] Daubin V, Ochman H. Start-up genes link genome evolution to ecology. *PLoS Biol*. 2004;2:e306.

[53] Kunin V, Ouzounis CA. The balance of driving forces during genome evolution in prokaryotes. *Genome Res*. 2003;13:1589–99.

[54] Collins RE, Higgs PG. Testing the infinitely many genes model for evolution of the bacterial core genome and pangenome. *Mol Biol Evol*. 2012;29:3413–25.

[55] Csűrös, M., Miklós, I. (2006). A probabilistic model for gene content evolution with duplication, loss, and horizontal transfer. Research in Computational Molecular Biology. RECOMB 2006. Lect Notes Comp Sci, Springer, Berlin, Heidelberg. 2006; 3909: 206–20.

[56] Rasmussen MD, Kellis M. Accurate gene-tree reconstruction by learning gene- and species-specific substitution rates. *Genome Res*. 2007;17:1932–42.

[57] Camin JH, Sokal RR. A method for deducing branching sequences in phylogeny. *Evolution*. 1965;19:311–26.

[58] Sankoff D, Abel Y, Hein J. A tree · a window · a hill: generalization of nearest-neighbor interchange in phylogenetic optimization. *J Classification*. 1994;11:209–32.

[59] Huson D, Nettles S, Warnow T. Disk-covering: a fast-converging method for phylogenetic tree reconstruction. *J Comput Biol*. 1999;6:369–86.

[60] Stubbersfield, J.,Tehrani, J. Expect the Unexpected? Testing for Minimally Counterintuitive (MCI) Bias in the Transmission of Contemporary Legends: A Computational Phylogenetic Approach: A Computational Phylogenetic Approach. *Social Science Computer Review*. 2012; 31: 90–102.

[61] da Silva SG, Tehrani JJ. Comparative phylogenetic analyses uncover the ancient roots of Indo-European folktales. *R Soc Open Sci.*. 1; 2016; 3: 150645.

[62] Swadesh M. Salish internal relationships. *Int J Am Linguist*. 1950;16:157–167.

[63] Swadesh M. Lexico-statistic dating of prehistoric ethnic contacts. *Proc Am Philos Soc*. 1952;96:452–463.

[64] Gray RD, Atkinson QD. Language-tree divergence times support the Anatolian theory of Indo-European origin. *Nature*. 2003;426:435–439.

[65] Gray RD, Atkinson QD, Greenhill SJ. Language evolution and human history. *Proc Natl Acad Sci USA*. 2009;106:16014–16019.

[66] Tanselle GT. Textual criticism and scholarly editing. *Studies in Bibliography*. 1995;48:1–56.

[67] Warnow T, Ringe D, Taylor A. Reconstructing the evolutionary history of natural languages. *Proc Natl Acad Sci USA*. 2006;103:8762–8767.

[68] Sackett JR. Approaches to style in lithic archaeology. *J Anthropol Archaeol*. 1982;1:59–112.

[69] Binford LR. Archaeology as anthropology. *Am Antiquity*. 1962;28:217–225.

[70] Shennan S. *Genes, Memes and Human History*. London: Thames & Hudson; 2002.

[71] O'Brien MJ, Lyman RL, Collard M. Cladistics is useful for reconstructing archaeological phylogenies. *J Archaeol Sci*. 2001;28:111–125.

[72] O'Brien MJ, Lyman RL. *Cladistics and Archaeology*. Salt Lake City: University of Utah Press; 2003.

[73] Lake MW, Venti J. Quantitative Analysis of Macroevolutionary Patterning in Technological Evolution: Bicycle Design from 1800 to 2000. In *Pattern and Process in Cultural Evolution*, Shennan SJ, ed; 2009:146–161.

[74] Bean JR. Pistol Phylogeny. https://surplused.com/index.php/2021/02/08/pistol-phylogeny-part-1-introduction/accessed February 14, 2026.

[75] Philippe Tatry, Justine Fesquet, Pascal Tassy, Gaetan Sciacco, Francis Duranthon, et al.. Cladistics applied to Aerospace. 9th European Conference for Aeronautics and Space Science (EUCASS). *HAL-Open Science*. hal-03909298.

[76] Baily J. Music structure and human movement. *Ethnomusicology*. 1985;29:237–259.

[77] Tëmkin I, Eldredge N. Phylogenetics and material cultural evolution. *Curr Anthropol*. 2007;48:146–153.

[78] Brown S, Savage PE, Ko AM. Music evolution, cultural transmission, and lineage diversification. *Ann N Y Acad Sci*. 2013;1289:54–64.

[79] Basalla G. *The Evolution of Technology*. Cambridge: Cambridge University Press; 1988.