Within-Pangenome Phylogeny Based on Structural Variants

by

Peng Xu

A Thesis submitted to the
University of Ottawa
in partial fulfillment of the requirements
for the degree of

Master of Science (MSc) in Statistics

Department of Mathematics and Statistics
Faculty of Science
University of Ottawa

Thesis Supervisor: David Sankoff

© Peng Xu, Ottawa, Canada, 2026

Acknowledgements

I would like to thank my parents for their support in enabling me to pursue my studies abroad. I am also grateful to the university for providing a supportive academic environment. I sincerely thank my supervisor, David Sankoff, for accepting me as his student and for his continuous guidance throughout this project. His patience, insight, and encouragement were essential at every stage of my research and played a central role in shaping both the direction and quality of this work. I also thank the science officer for timely assistance during my studies, and my senior, Zhong Haitian, for her constant help and support. She guided me through the early stages of my master's studies and helped me become familiar with academic life and daily arrangements.

Abstract

Structural variations, particularly chromosomal inversions, are a major source of genomic diversity, yet their potential for phylogenetic reconstruction within pangenomes has not been fully established. This study examines the phylogenetic signal in inversion presence and absence patterns under a Dollo style evolutionary framework that assumes complex mutations arise once and are not regained.

We analyzed two plant pangenomes, radish (*Raphanus sativus*) and cotton (*Gossypium* spp.). Large inversions were extracted from published resources and encoded as binary matrices. Using these matrices, we reconstructed phylogenies with the neighbour-joining method and compared them with published reference trees. The reference phylogeny for radish is based on sequence data, whereas the cotton reference tree was inferred from a genome-wide catalogue of structural variants. Tree similarity was assessed using three complementary measures: bipartition overlap, Maximum Agreement Subtree size, and co-phenetic correlation.

The inversion-based trees did not fully reproduce the reference topologies, but they retained consistent internal structure. MAST analyses revealed subsets of accessions whose relationships were stable across trees, and co-phenetic correlations were clearly higher than expected under randomization. Simulation results further showed that stability under noise depends strongly on local structure. In particular, the preservation of sister pairs played a central role in maintaining agreement as the perturbation level increased.

This work therefore contributes to a clearer understanding of what kinds of evolutionary information can and cannot be recovered from inversion data alone. The results indicate that inversions capture a meaningful, though incomplete, phylogenetic signal that is different from that provided by sequence data. Rather than viewing these differences as shortcomings, they reflect distinct evolutionary constraints acting on structural variation. Our analyses also show that split-based measures perform poorly on sparse SV data, while MAST and co-phenetic correlation provide more robust and interpretable assessments of tree similarity. Consequently, MAST and co-phenetic correlation offer a practical way to evaluate phylogenetic signal in sparse SV data and to distinguish evolutionary structure from technical noise.

Résumé

Les variations structurales, en particulier les inversions chromosomiques, constituent une source majeure de diversité génomique ; toutefois, leur potentiel pour la reconstruction phylogénétique au sein des pangénomes n'a pas encore été pleinement établi. Cette étude examine le signal phylogénétique contenu dans les patrons de présence et d'absence d'inversions dans un cadre évolutif de type Dollo, qui suppose que les mutations complexes apparaissent une seule fois et ne sont pas regagnées.

Nous avons analysé deux pangénomes végétaux, le radis (*Raphanus sativus*) et le coton (*Gossypium* spp.). Les grandes inversions ont été extraites de ressources publiées et codées sous forme de matrices binaires. À partir de ces matrices, nous avons reconstruit des phylogénies à l'aide de la méthode du neighbour-joining et les avons comparées à des arbres de référence publiés. La phylogénie de référence du radis repose sur des données de séquence, tandis que celle du coton a été inférée à partir d'un catalogue génomique complet de variations structurales. La similarité des arbres a été évaluée au moyen de trois mesures complémentaires : le chevauchement des bipartitions, la taille du sous-arbre d'accord maximal (MAST) et la corrélation cophenétique.

Les arbres fondés sur les inversions ne reproduisent pas entièrement les topologies de référence, mais ils conservent une structure interne cohérente. Les analyses MAST ont révélé des sous-ensembles d'accessions dont les relations demeurent stables d'un arbre à l'autre, et les corrélations cophenétiques étaient nettement supérieures à celles attendues sous randomisation. Les résultats des simulations montrent en outre que la stabilité face au bruit dépend fortement de la structure locale. En particulier, la préservation des paires sœurs joue un rôle central dans le maintien de la concordance lorsque le niveau de perturbation augmente.

Ce travail contribue ainsi à une meilleure compréhension des types d'information évolutive qui peuvent — ou non — être récupérés à partir des seules données d'inversions. Les résultats indiquent que les inversions capturent un signal phylogénétique significatif, bien qu'incomplet, distinct de celui fourni par les données de séquence. Plutôt que de considérer ces différences comme des insuffisances, elles reflètent des contraintes évolutives distinctes agissant sur la variation structurale. Nos analyses montrent également que les mesures fondées sur les partitions (splits) sont peu performantes sur des données de variations structurales clairsemées, tandis que le MAST et la

corrélation cophenétique offrent des évaluations plus robustes et interprétables de la similarité entre arbres. Par conséquent, le MAST et la corrélation cophenétique constituent une approche pratique pour évaluer le signal phylogénétique dans des données de variations structurales clairsemées et pour distinguer la structure évolutive du bruit technique.

Contents

Acknowledgements ii

Abstract iii

Résumé iv

1 Introduction 1

2 Methods 3

2.1 Tree Inference . . . 3

2.2 Tree comparison . . . 4

2.2.1 Approaches to comparison . . . 5

3 Data 10

3.1 Pangenomes . . . 10

3.1.1 Radish and Cotton . . . 11

3.2 Biological significance of inversions and structural variations . . . 12

4 Results 14

4.1 Topological comparisons . . . 14

4.1.1 Cotton Dataset . . . 14

4.2 Radish Dataset . . . 17

4.2.1 Bipartition Randomization Test . . . 20

4.3 Bootstrap Analysis . . . 22

4.4 MAST and cophenentic correlation analyses of the cotton and radish trees . . . 22

4.5 Noise Simulation . . . 27

5 Degrading Dollo 30

5.1 Simulation design . . . 30

5.1.1 Co-phenetic correlation across twenty groups . . . 34

5.1.2 Local structural explanation: preserved sister pairs 35

6 Conclusion **38**

A Radish Presence/Absence Matrix **40**

B Cotton Presence/Absence Matrix **45**

Code Availability **48**

List of Figures

4.1 neighbour-joining tree constructed from the inversion matrix in this study. 15
4.2 Reference tree published in Jin et al. (2023). 15
4.3 Published phylogenetic tree reported by Zhang et al. (2021). 17
4.4 neighbour-joining tree reconstructed from inversion data in this study. 18
4.5 Maximum agreement subtree for the radish dataset. 24
4.6 Maximum agreement subtree highlighted in the radish NJ phylogeny. 24
4.7 Maximum agreement subtree highlighted in the radish reference tree 25
4.8 Maximum agreement subtree for the cotton dataset. 25
4.9 Maximum agreement subtree highlighted in the cotton NJ phylogeny. 26
4.10 Maximum agreement subtree highlighted in the cotton reference tree. 26
4.11 Mean MAST size across noise levels in the cotton dataset. 28
4.12 Mean MAST size across noise levels in the radish dataset. 28

5.1 Example tree with 9.4% red tips (3 out of 32). 31
5.2 Example tree with 21.9% red tips (7 out of 32). 31
5.3 Example tree with 34.4% red tips (11 out of 32). 32
5.4 Example tree with 50.0% red tips (16 out of 32). 32
5.5 Mean MAST size under increasing tip flipping, 20 groups. 33
5.6 Mean co-phenetic correlation under increasing tip flipping, grouped into twenty bins by the initial proportion of ones. 35
5.7 Correlation between preserved sister pairs and mean MAST across twenty subgroups. 37

List of Tables

4.1 Topological comparison between the NJ tree reconstructed in this study and the reference tree reported in Jin et al. (2023). 16
4.2 Topological differences between the NJ tree and the published reference tree for the radish dataset. 19
4.3 Bipartitions shared between the original NJ tree and 100 randomized trees. 20
4.4 Co-phenetic correlations between NJ trees and reference trees. 27

A.1 Radish presence/absence matrix transposed 40

B.1 Cotton presence/absence matrix transposed 45

Chapter 1

Introduction

The research reported in this thesis evolved in two directions, the second one inspired by initial results from the first. The goal at the outset was to see if among the several constituent genomes of a pangenome, phylogenetic divergence at the level of structural variation mirrors that at the level of sequence mutation — the pangenome of a species consisting of all the genomes of all the variants of this species. This question is provoked by two competing tendencies. On the one hand, evolutionary rate heterogenity among structural variants, due in part to their likely functional concomitants, in contrast to the neutral mutations favoured by DNA sequence evolutionary protocols (like K_s or 4DTv), could distort inferred phylogenetic histories. Also structural variation among constituent genomes could easily be shared in cross-breeding populations, e.g., by introgression, whether or not they are the most closely related participants. These, or other departures from standard models of DNA change, could lead to lack of congruence of the sequence-level tree and the tree reflecting structural variation diversity. On the other hand, we know that at the higher level, that of distinct species genera, familes and orders, phylogenetic trees based on structural variants such as inversions or other rearrangements are often identical or similar to trees based on sequence divergence.

In investigating this question, we were also led to question whether the relatively short historical time interval separating members of a pangenome would lead to a restricted pattern of evolutionary change. For example, it might be that structural innovations might not have time to reverse themselves within a short interval, suggestive of a Dollo's law [1] regime, whereby the loss of a complex structure cannot be recovered. Although this principle is usually formulated in the context of organismal evolution, could phylogenetic inference in the pangenomic context be affected in such a case?

To address this question, we obtained data on two plant pangenomes. We compared the internal phylogeny based on structural variation with that based on sequence variation, the latter being reported by the authors of the original pangenome publication. We made use of a number of ways of comparing phylogenetic trees, while relying on a single well-known procedure for

tree construction, neighbour-joining, to ensure consistency among the test procedures. One of the comparative measures is the maximum agreement subtree (MAST), which is an essentially combinatorial criterion, relatively recently introduced. To tie this to more traditional concerns in numerical taxonomy, we also make use of co-phenetic correlation [2].We controlled its results with a more traditional, largely quantitative measure, the co-phenetic corrrelation coefficient.

Importantly, while there is some degree of resemblance between the sequence-based trees and the structural variation-based trees, there are many differences. This suggests that the sequence evolution and structural variant evolution may be following a different rhythm.

This led to the second topic, a simulation-based exploration of the Dollo hypothesis and how this is vulnerable to data that contain errors or other exceptions to this principle. In this chapter, we systematically document how phylogenies reflective of Dollo are gradually distorted as exceptions are allowed to proliferate.

In Chapter 5, we simulate pangenomic phylogenies satisfying Dollo's law, and systematically introduce errors to see how these affect the accuracy in the reconstruction of the original phylogenetic tree.

This leads to a synthesis of the results of Chapters 3 and 5. The main contributions that emerge from this thesis, are

- the disruption of phylogenetics congruence between sequence-based data and structural variant data among the constituent genomes within a single pangenome, due to a variety of factors tied to the the genomic similarity among these genomes — just a suggestion based on only two pangenomes, and

- the pattern of degradation of phylogenetic relationships among genomes generated originally by a Dollo process, as a function of the amount of noise or errors introduced in the data.

Chapter 2

Methods

2.1 Tree Inference

Building trees in biology has a long history, from the mid-eighteenth century classification by Carl Linnaeus [7] of 10,0000 biological species in a hierarchical taxonomy. The evolutionary interpretation of this concept, which we call "phylogeny", was formulated by Charles Darwin [8] a century later, who 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) duality was clearly reflected after another century in Sokal and Sneath's "phenetics" [9] versus Farris's "cladistics" [35] dispute preceding the modern statistical synthesis led by Felsenstein [37]. His use of likelihood methodology paved the way for today's Bayesian methodology [38]. In parallel to these developments in the 1960s and early 1970s, molecular phylogenies were initiated by Zuckerkandl and Pauling [10] and Dayhoff [11], using protein, and Sankoff and Cedergren for nucleic acids [12, 13].

Reconstructing the history of transmitted features in tree-like evolutionary systems also plays a role in the reconstruction of language families and sound changes in historical linguistics, from Swadesh's tree models [17, 18] to Bayesian phylogenetic analyses of language diversification [19, 20], textual criticism [21, 22], cultural evolution [15, 16] and the archaeology of stone tools and ceramic history [23, 24, 25, 26, 27].

Other subjects — bicycles [28], firearms [29] and aircraft [30] — 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 [31, 32, 33]. The historical development of technical standards — such as screw threads, railway gauges, electrical connectors, and communication protocols — often exhibits tree-like divergence [34].

There are many approaches to phylogenetic inference, from statistical agglomerative cluster

analysis like average link, single link and complete link [2] to the more combinatorial, like parsimony [35], quartet-based, triplet-based and bipartition-based [36], and to the modern maximum likelihood [37] and Bayesian approaches implemented in powerful software like Mr Bayes [38] and Beast [39]. We will be working with relatively small data sets, so that for consistency, we apply a single tree reconstruction method to generate phylogenies across different datasets. The neighbour-joining (NJ) algorithm [40] was selected for its efficiency, scalability, and widespread use. This method is a widely-used algorithm for building phylogenetic trees from a matrix of distances among species or varieties. There being no recognized probabilistic models for rearrangement phylogenetics, we do not explore maximum likelihood-based approaches, especially as neighbour-joining is known to closely approximate maximum likelihood in most contexts [41]. In common with other agglomerative methods it works by iteratively finding the pair of clusters whose connection results in the smallest possible increase in total branch length. These pairs are then joined step-by-step to produce unrooted trees designed to reflect the evolutionary relationships among the samples. NJ can generate results quickly without a major loss in topological accuracy. This balance between speed and reliability makes it a practical choice when we need to infer very many trees.

The matrices we will analyze are derived from tables of presence of absence of structural variants across a number of species. These are converted into species $\times$ species comparisons using Hamming distance applied to each table, and then to phylogenies using the NJ algorithm. Hamming distance, which is an L1 measure, is used in sequence comparison, as the sum of a position-by-position comparison is evolutionarily meaningful. A metric like Euclidean, on the other hand, would suggest that between a sequence and its descendant there might be intermediate objects which are not sequences.

2.2 Tree comparison

A graph-theoretical tree is a connected, acyclic graph. For our purposes, a phylogenetic tree is a binary branching tree, i.e., where all vertices have degree 3 (called internal or ancestor nodes) or degree 1 (called terminal vertices, leaves or tips), and where each of the $n < \infty$ tips has a distinct label. This kind of tree (or phylogeny) is used to represent the evolutionary relationships among different biological entities, be they genes, genomes, non-coding RNA, species, varieties, populations or other. The labeled leaf nodes represent the observed objects and the internal nodes denote ancestral objects. Positive numbers (branch lengths) may be associated with the edges of the graph, and can be interpreted as various types of distances, such as genetic distance or evolutionary time, thus providing a quantitative reference for the relationships among the studied objects, although this does not play a major role in this study. Sometimes one of the nodes, or one additional internal node of degree 2, called the root, is present and all the edges of the tree are

directed away from the root, symbolizing temporal progression from the ancestor to the present-day biological objects. Phylogenetic trees are not only a core tool in evolutionary biology but are also widely applied in areas such as crop improvement, pathogen tracking, and ecosystem studies, as well as areas outside of biology, as listed in the previous section.

In the first part of this thesis, we focus on the comparison of two or more phylogenetic trees on the same set of n labeled leaves. One tree is the "reference" trees extracted from the literature, which has been constructed by genome-sequence comparison. The other tree we construct based solely on structural variant data.

Trees reconstructed using different methods or data types may yield markedly different; trees based on different gene segments may reveal conflicting signals [46]; or in the analysis of ecologically associated groups such as hosts and their symbionts, it may be necessary to assess the similarity of their phylogenetic structures. To quantify that, various tree structure-based comparison methods have been introduced, such as co-phenetic correlation [2], bipartition analysis [42], the Robinson–Foulds distance [43], the Maximum Agreement Subtree (MAST) [4], triplet distance [44] and many others. Many of these methods can be computed in polynomial time and provide numerical measures of differences between phylogenetic trees.

2.2.1 Approaches to comparison

In this study, we make use of three methods having different emphases in handling topological variation: bipartition analysis [42], the Maximum Agreement Subtree (MAST) method [3] and co-phenetic correlation [2].

Bipartitions In the case of the binary branching trees that interest us, bipartition analysis operates by repeatedly partitioning the n leaves in a phylogenetic tree into $n - 3$ bipartitions, by deleting one of the $n - 3$ internal (non-terminal) edges from the tree. For each deletion, the n leaves in the two disjoint subtrees thus formed are assigned to two non disjoint subsets containing two or more leaves. The similarity between two trees is then quantified by counting the number of bipartitions they have in common. This approach is computationally straightforward, but it is sensitive to even minor topological rearrangements, as a single branch change can alter multiple bipartitions.

Co-phenetic Correlation It can be used to measure how accurately the pairwise distances between species in one tree match those in the other tree, acting as a goodness-of-fit metric. It calculates the Pearson correlation between the path lengths in each tree, from each pair of leaves to the node where they are joined.

Maximum Agreement Subtree MAST, identifies the largest set of leaves such that the subtree defined by this set is topologically identical in both trees, providing a conservative measure of shared phylogenetic signal. Compared with bipartition analysis, which emphasizes detailed branching differences, MAST is less sensitive to minor changes in the overall structure.

Building on the definition in [3], Finden and Gordon [4] extended the approach to what they termed the *largest common pruned tree*, which generalized the MAST idea to hierarchical classifications and dendrograms, and introduced practical algorithms for identifying such subtrees. They also explored related concepts such as truncation (focusing on low-level relationships), object categorization (identifying taxa consistently retained across maximum-size solutions), and regrafting (reattaching pruned branches while preserving consensus structure). These developments strengthened the methodological toolkit for structural tree comparison and influenced later computational implementations of MAST.

In [3] Gordon proved the MAST problem to be NP complete, meaning there is no known polynomial time algorithm that solves it efficiently for all input sizes. However, in our case, restricted to binary trees Aho et al.[80] found a polynomial time algorithm, which laid the groundwork for later algorithmic implementations. Later, Steel and Warnow [5] explored the mathematical properties of computing MAST among multiple trees and studied the increased complexity beyond pairwise comparisons.

The MAST framework is useful when handling incomplete or imperfect data, since it is based only the subset of taxa whose relationships are identically preserved in both trees, thereby minimizing the influence of unstable or spurious branches.

In recent years, several widely used software tools have implemented MAST algorithms, including `phangorn` in R (used in this study), `DendroPy` in Python, and `TreeCmp`. These tools have enabled efficient application of MAST to empirical datasets in various fields such as phylogenetic reconstruction, gene family evolution, and viral lineage analysis. In this study, MAST was implemented using the `phangorn` package, which allowed us to detect subtrees with consistent topologies across our NJ and reference trees, providing additional insights into the phylogenetic signal of structural variants.

Algorithm 1 Maximum Agreement Subtree Size (Rooted Trees)

procedure MAST(T_1, T_2)
  $L \leftarrow$ leafLabels(T_1) $\cap$ leafLabels(T_2)
  $T_1' \leftarrow$ restrictToLeaves(T_1, L)
  $T_2' \leftarrow$ restrictToLeaves(T_2, L)
  $r_1 \leftarrow$ root(T_1')
  $r_2 \leftarrow$ root(T_2')
  for all $u \in$ postorder(T_1') **do**
    for all $v \in$ postorder(T_2') **do**
      $DP[u, v] \leftarrow$ ComputeDP(u, v, DP)
    end for
  end for
  return $DP[r_1, r_2]$
end procedure

Algorithm 2 ComputeDP(u, v, DP)

if u is leaf **and** v is leaf **then**
  if label(u) = label(v) **then**
    return 1
  else
    return 0
  end if
end if
if u is leaf **then**
  return $\max_{w \in \text{leavesUnder}(v)} DP[u, w]$
end if
if v is leaf **then**
  return $\max_{w \in \text{leavesUnder}(u)} DP[w, v]$
end if
$A \leftarrow$ children(u)
$B \leftarrow$ children(v)
Construct weight matrix W of size $|A| \times |B|$
for $i = 1$ to $|A|$ **do**
  for $j = 1$ to $|B|$ **do**
    $W[i, j] \leftarrow DP[A_i, B_j]$
  end for
end for
$M \leftarrow$ MaxWeightBipartiteMatching(W)
return weight(M)

We finish this methdological chapter with a technical summary of the procedures to be applied to the structural variant × accession matrices we derived from pangenome data.

For both datasets, phylogenetic trees were reconstructed using the neighbour-joining method.

Reference trees from the original publications were imported and compared with the neighbour-joining trees reconstructed in this study. Bipartitions were extracted from reference trees and NJ trees, and the number of shared bipartitions was counted to assess topological similarity.

This pipeline was applied consistently to both cotton and radish datasets, which enabled direct comparison of phylogenetic stability and robustness across species. Before presenting the results of these analyses, we first consider why structural variations, particularly inversions, are of evolutionary and biological importance.

Let n denote the number of sampled accessions and m denote the number of structural variation events. For each dataset, we constructed a binary character matrix

$$X = (x_{ij}) \in \{0, 1\}^{n \times m},$$

where $x_{ij} = 1$ if inversion j is present in accession i, and $x_{ij} = 0$ otherwise.

For the radish dataset, $n = 11$ and $m = 141$. For the cotton dataset, $n = 11$ and $m = 81$. Rows correspond to accessions and columns correspond to distinct inversion events. Both matrices are sparse, containing a high proportion of zero entries, reflecting the limited overlap of inversion events across accessions.

Distance computation Phylogenetic reconstruction was based on Hamming distances. For any pair of accessions i and k, the distance is defined as

$$d_{ik} = \sum_{j=1}^{m} |x_{ij} - x_{kj}|.$$

This defines a metric on the binary vector space $\{0, 1\}^m$ and measures the total number of inversion state differences between two genomes.

Tree reconstruction Given the distance matrix $D = (d_{ik})$, an unrooted tree T was reconstructed using the neighbour joining algorithm, which iteratively joins taxa to minimize a balanced minimum evolution criterion derived from pairwise distances.

Bipartition comparison For a tree T with leaf set $\mathcal{L}$, each internal edge induces a bipartition of $\mathcal{L}$. Let $\mathcal{B}(T)$ denote the set of non-trivial bipartitions (internal splits) of T, excluding splits that separate a single leaf from the remaining taxa, since such splits are shared by all trees on the same leaf set.

Topological similarity between two trees T_1 and T_2 was quantified as

$$|\mathcal{B}(T_1) \cap \mathcal{B}(T_2)|,$$

that is, the number of identical internal splits shared between the trees.

Maximum Agreement Subtree Because exact bipartition matching is sensitive to minor topological perturbations, we additionally employed the Maximum Agreement Subtree method as a coarser but more robust measure of shared structure.

Given two trees T_1 and T_2 on leaf set $\mathcal{L}$, the Maximum Agreement Subtree is defined as the largest subset

$$S \subseteq \mathcal{L}$$

such that the induced subtrees $T_1|_S$ and $T_2|_S$ are identical.

The statistic $|S|$ was used to quantify the maximal shared topological structure between trees.

Noise simulation To assess robustness under increasing levels of artificial error, controlled noise was introduced into the binary matrix. Let

$$Z = \{(i, j) : x_{ij} = 0\}$$

denote the set of zero entries in X.

At each noise level k, k elements of Z were selected uniformly at random and converted to one, simulating false positive inversion calls. A new tree was reconstructed from the perturbed matrix, and MAST size relative to the original tree was recorded.

Noise levels increased in increments of 30 modified entries per replicate.

Bipartition Randomization Test To evaluate whether the bipartitions in our NJ trees reflect phylogenetic structure rather than noise, we carried out a randomization test. For each dataset, we generated 100 randomized matrices by shuffling the positions of ones within each row while preserving row sums. We then reconstructed a neighbour-joining tree from each randomized matrix and counted the number of bipartitions shared with the original tree.We then performed a bipartition randomization test by shuffling the positions of ones within each row while preserving row sums, but not column wise frequencies, to generate one hundred randomized matrices. neighbour-joining trees were reconstructed for each replicate, and the number of bipartitions shared with the original tree was counted.

Chapter 3

Data

3.1 Pangenomes

One objective of this study is to investigate the phylogenetic relevance of structural variations, with a particular emphasis on chromosomal inversions, focusing on the pangenome scale. We searched for datasets containing a sufficient number of inversion events distributed across the genome, where each inversion was identified by chromosomal breakpoints, allowing the construction of presence/absence binary matrices based on inversion events. We also required the availability of previously constructed reference phylogenetic trees, based on sequence data in radish and on structural variation profiles in cotton.

Pangenomes have now been published for a wide range of flowering plants, with the majority focused on agronomically important crops [45, 47, 48, 49]. These include

- rice (*Oryza sativa*) [51, 52, 53],
- maize (*Zea mays*) [54],
- bread wheat (*Triticum aestivum*) [55, 56],
- soybean (*Glycine max* and its wild relative *G. soja*)[57, 58],
- tomato (*Solanum lycopersicum*) [59, 60, 61],
- oilseed rape or canola (*Brassica napus*) [62],
- cabbage and relatives (*B. oleracea*) [50, 63, 64],
- the model plant thale cress (*Arabidopsis thaliana*) [65, 66, 67],
- cucumber (*Cucumis sativus*) [68],

- foxtail millet (*Setaria viridis*/*S. italica*) [69],
- potato (*Solanum tuberosum*) [70],
- strawberry (*Fragaria* spp.) [71],
- tea plant (*Camellia sinensis*) [72, 73],
- pepper (*Capsicum* spp.) [74],
- barley (*Hordeum vulgare*) [75],
- banana (*Musa* spp.) [76],
- radish (*Raphanus sativus* L.) [78]
- cotton (*Gossypium* spp.) [79],

3.1.1 Radish and Cotton

Two crop species were identified as suitable sources for this study: radish (*Raphanus sativus* L.) [78] and cotton (*Gossypium* spp.) [79], both of which provide publicly available pangenome resources. These datasets encompass multiple subspecies, diverse geographical origins, and distinct evolutionary stages, and include detailed and experimentally validated annotations of SVs, particularly inversions. Both also provide pre-computed DNA sequence-based phylogenetic trees on one hand, and summary SV statistics among accessions on the other hnd, making them especially suitable for the goals of this study.

In the study by Zhang et al. (2021) [78], a pangenomic dataset was constructed for the genus *Raphanus* by systematically sampling representative accessions from domesticated, wild, and weedy types. Eleven genomes were *de novo* assembled using PacBio long-read sequencing with Canu [84], and three of them were further scaffolded with Hi-C using Juicer [85]. The selected accessions included seven domesticated varieties, such as cherry belle radish (*R. sativus* var. *radicula*), black radish (*R. sativus* var. *niger*), and rat's tail radish (*R. sativus* var. *caudatus*), alongside two wild subspecies (*R. raphanistrum* ssp. *landra* and ssp. *raphanistrum*), one cultivated-wild hybrid, and one invasive weedy type.

Comprehensive SV detection revealed a broad spectrum of events, including insertions, deletions, duplications, inversions, and translocations. Between 20,000 and 30,000 medium-sized SVs and thousands of large presence/absence variations (PAVs) were reported per genome. Inversions were particularly abundant and varied widely among accessions: in some genomes, up to 134–193 inversions were identified, ranging from 2.7 Mb to 21.1 Mb in size, with 949 large inversions

exceeding 50 kb. Wild accessions generally exhibited more numerous and larger inversions than domesticated ones, indicating more active chromosomal rearrangements in wild populations. For the purpose of this thesis, we utilized the inversion presence and absence information provided in the supplementary tables of Zhang et al. (2021) and converted these entries directly into a binary matrix for the eleven radish accessions to facilitate a phylogenetic analysis.

Similarly, the pangenome dataset for cotton described by Jin et al. (2023) [79] includes sixteen accessions spanning different evolutionary lineages, including Asian and American domesticated cottons as well as multiple wild species. These represent various stages of domestication and agricultural adaptation. Genome assemblies were generated using Illumina short-read sequencing with SOAPdenovo [86], 10X Genomics linked reads assembled by SuperNova [87], and chromosome-scale scaffolding based on Hi-C data using HiC-Pro [88]. The dataset provides detailed SV annotations and genome-wide comparisons among accessions, making it an ideal reference for investigating the phylogenetic informativeness of structural variations.

Inversions represented an important part of the structural variation catalogue in the cotton dataset. On average, 349 inversions were identified per genome (ranging from 113 to 542), and across all assemblies a non-redundant set of 2,236 inversions was obtained [79], ninety-eight of which were larger than 1 Mb and extended up to 32.4 Mb. Large-scale inversions of this kind are known to suppress recombination and may have significant evolutionary consequences. In the present study, we focused on the large inversions that are summarized in the pan-inversion map in Figure 1A of Jin et al. (2023). We manually digitized this figure and recorded, for each megabase-scale inversion, whether it was present in each of the eleven accessions included in our analysis. This procedure produced a binary presence and absence matrix of large inversions, which forms the basis for tree reconstruction.

3.2 Biological significance of inversions and structural variations

The large number of inversions identified in both the radish and cotton datasets raises a central question for this study: how do these events contribute to evolutionary patterns and to traits of agronomic value. Structural variation refers to genomic changes larger than 50 bp, including insertions, deletions, duplications, translocations, and inversions. Because such variants can alter regulatory context or gene dosage, their effects on gene expression, chromatin organization, and phenotype are often more extensive than those of single nucleotide polymorphisms.

An inversion arises when a chromosomal segment breaks, rotates, and rejoins in the opposite orientation. Although the total amount of DNA remains unchanged, this rearrangement can modify

regulatory relationships between genes and their control elements and can strongly affect local recombination patterns. As a result, inversions may influence phenotypic traits and contribute to longer term evolutionary dynamics.

In natural populations, recombination suppression within inverted regions can maintain combinations of alleles that function well in specific environments. Such regions may behave as so called supergenes, preserving adaptive haplotypes and, in some cases, contributing to partial reproductive isolation. In crop species, the same mechanism can stabilize sets of loci that affect key agronomic traits. Whereas a single nucleotide polymorphism typically affects one site, a structural variant can span a much larger genomic interval, increasing its potential impact on gene regulation and trait variation.

In *Gossypium hirsutum*, Jin et al. (2023)[79] reported several inversions associated with important phenotypes. For example, one inversion on chromosome A07 shifts the positions of regulatory elements, resulting in altered gene expression patterns that influence fiber length, yield, and flowering time. This inversion appears to have been favored during domestication and remains of interest for modern breeding programs.

In *Raphanus sativus*, Zhang et al. (2021)[78] observed that inversions are particularly frequent in wild accessions and often occur in genomic regions characterized by restricted gene flow. Some of these inversions show associations with root morphology, suggesting that they may influence developmental programs and have contributed to radish diversification.

Inversions often occur in variable or non reference regions, and their breakpoints are difficult to define when relying on a single reference genome. To address this limitation, both Jin et al. (2023)[79] and Zhang et al. (2021)[78] adopted a pangenome framework that integrates assemblies from multiple individuals and distinguishes core regions shared by all accessions from variable regions present only in some. By enabling direct comparison across assemblies, the pangenome approach captures genomic diversity that a single reference cannot represent and facilitates more accurate identification of inversion boundaries. This framework was therefore essential for detecting large numbers of inversions, comparing their distributions across samples, and relating specific events to traits or lineage divergence.

Despite their biological relevance, inversions remain challenging to incorporate into phylogenetic reconstruction. Unlike nucleotide sequences, structural variants cannot be aligned and are usually represented as binary presence or absence matrices. Such matrices often contain many missing entries and detection inconsistencies, particularly in repetitive or low coverage regions. Differences in event size and uncertainty in annotation further complicate distance estimation, leading to unstable tree topologies and reduced comparability between inferred trees.

Chapter 4

Results

4.1 Topological comparisons

4.1.1 Cotton Dataset

For the cotton dataset, inversion data were converted into a binary matrix in which rows correspond to accessions and columns correspond to specific inversion events. A value of 1 indicates the presence of a given inversion in an accession, and a value of 0 indicates its absence.

Allotetraploid cotton was selected as a case study because Jin et al. (2023)[79] provide high quality genome assemblies and a published reference phylogeny for this group. In that study, a neighbour-joining tree was constructed from the complete structural variation dataset for twelve assemblies using the full SV genotype matrix rather than only large inversions (Figure S31). In the present work, we reconstructed a neighbour-joining tree from the binary matrix of large inversions and compared it with the published SV based reference tree. Figure 4.1 shows the NJ tree reconstructed in this study, and Figure 4.2 shows the published reference tree.

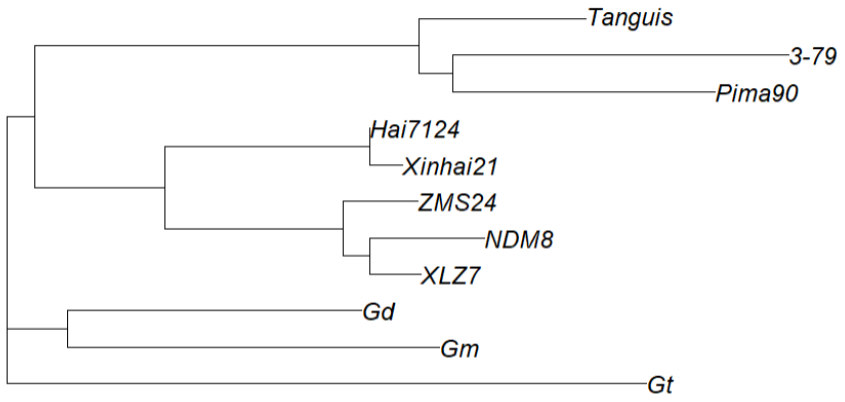

Figure 4.1: neighbour-joining tree constructed from the inversion matrix in this study.

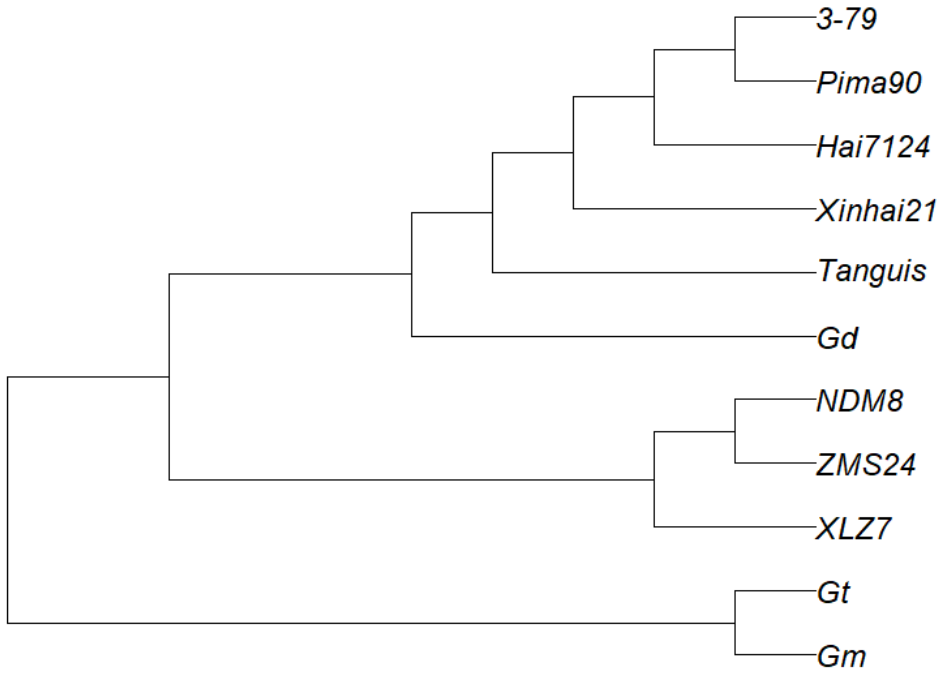

Figure 4.2: Reference tree published in Jin et al. (2023).

Although both trees were constructed using the neighbour-joining method, their topologies differ in several aspects. It is important to note that inversion based phylogenies are not necessarily expected to recapitulate the same signal as sequence based or genome wide reference phylogenies. Inversion events can be shaped by functional constraints and selection, whereas reference trees derived from large collections of variants or neutral sequence sites are intended to reflect broader genealogical history. From this perspective, topological discrepancies are neither surprising nor a criticism of the inversion based tree, but indicate that the two data types may capture complementary

aspects of evolution. Table 4.1 summarizes key discrepancies, including the placement of *Gd*, *Gm*, and *Gt*, and relationships involving Tanguis and ZMS24.

Table 4.1: Topological comparison between the NJ tree reconstructed in this study and the reference tree reported in Jin et al. (2023).

Comparison	**This Study**	**Reference Tree**
Gd position	Close to the upland cotton clade	Separate lineage
Gm and *Gt* ordering	*Gt* diverges before *Gm*	*Gm* diverges before *Gt*
Tanguis position	Groups with Hai7124 and Xinhai21	Diverges earlier
ZMS24 relationship	Closer to NDM8	Closer to XLZ7
Overall placement of *Gd*, *Gm*, and *Gt*	*Gd* close to the upland cotton clade	*Gd* forms an independent lineage

Bipartition analysis showed that the NJ tree constructed from the inversion matrix contained eight bipartitions, whereas the reference tree contained nine. Only two bipartitions were shared: one grouping NDM8, XLZ7, and ZMS24, and another grouping 3-79 and Pima90. This limited overlap indicates substantial topological divergence. Several factors likely contribute to this pattern. First, the reference tree is based on the full structural variation catalogue, whereas the present analysis uses only a manually digitized subset of large inversions and therefore captures only part of the information in the original SV genotype matrix. Second, Jin et al. (2023)[79] detected structural variants using SyRI with stringent size and support thresholds, merged events across samples using SURVIVOR, and genotyped them using the Giraffe mapping strategy. In contrast, the inversion matrix used here was reconstructed from published figures without reproducing the original calling and validation pipeline. Differences in variant detection, filtering, and representation are therefore expected to introduce inconsistencies in inferred presence or absence patterns and to affect tree topology.

A further difference lies in the size and composition of the datasets used to build the trees. The large inversion matrix analyzed here is relatively small and sparse compared with the full SV genotype matrix of Jin et al. (2023), and it excludes other classes of structural variation such as insertions and deletions that contribute to overall genomic divergence. neighbour-joining trees inferred from a limited number of binary characters can be sensitive to small changes in the input matrix, so even a small number of misclassified or missing inversions may alter inferred relationships. The reference SV based tree integrates signal from many more events distributed across the genome and therefore reflects a broader structural variation history than the inversion only tree, which focuses on a narrow subset of large rearrangements.

4.2 Radish Dataset

To assess whether an inversion based tree captures meaningful phylogenetic structure in the radish dataset, we compared the neighbour-joining tree reconstructed in this study with the reference phylogeny reported by Zhang et al. (2021)[78]. As for cotton, the NJ tree was inferred from a binary matrix that recorded the presence or absence of inversions in each accession and was derived from the inversion map in Figure 2B of the original study. In contrast, the published reference tree for radish was constructed from sequence data and does not rely on structural variation. Zhang et al. (2021)[78] inferred this phylogeny using a maximum likelihood approach based on fourfold degenerate site transversions from 4,464 single copy orthologous genes.

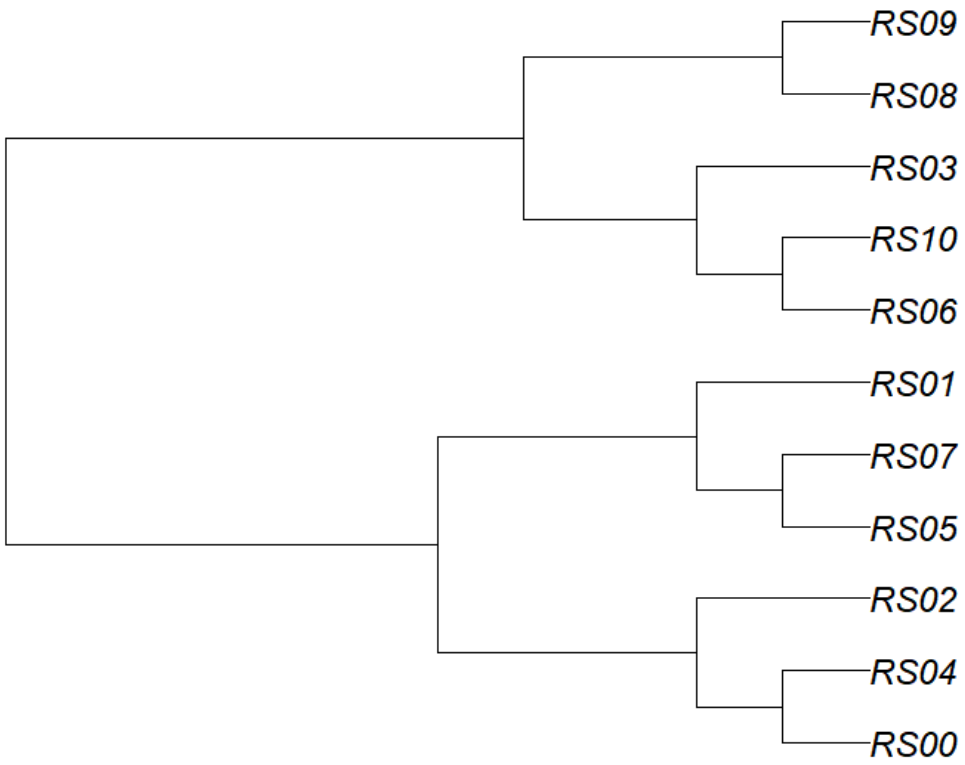

Figure 4.3: Published phylogenetic tree reported by Zhang et al. (2021).

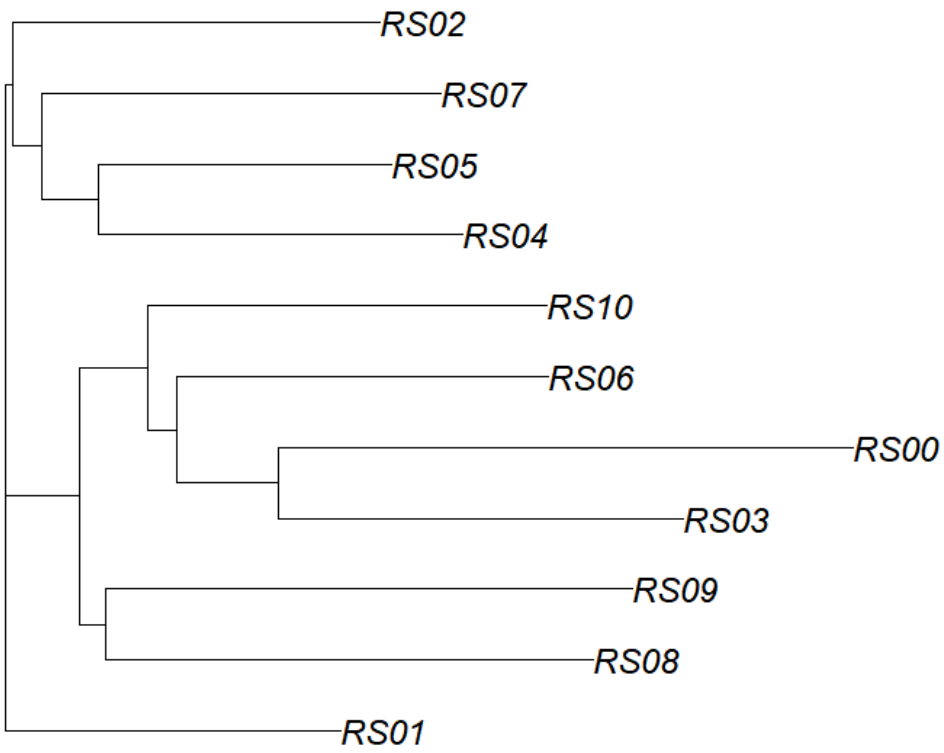

Figure 4.4: neighbour-joining tree reconstructed from inversion data in this study.

The main topological differences between the two trees are summarized in Table 4.2. Most discrepancies involved RS00 to RS05, whereas RS06 to RS10 showed higher agreement.

Table 4.2: Topological differences between the NJ tree and the published reference tree for the radish dataset.

Sample	**This Study**	**Published Tree**
RS00	Grouped with RS03	Grouped with RS04
RS01	Grouped with RS08	Grouped with RS06
RS02	Grouped with RS04, RS05, and RS07	Grouped with RS04
RS03	Grouped with RS00	Grouped with RS10
RS04	Grouped with RS05, then RS07	Grouped with RS02, then RS05 and RS07
RS05	Grouped with RS04, then RS07	Grouped with RS07
RS06	Grouped with RS10	Grouped with RS10
RS07	Grouped with RS05	Grouped with RS05
RS08	Grouped with RS09, then RS03	Grouped with RS09
RS09	Grouped with RS08, then RS03	Grouped with RS08
RS10	Grouped with RS06, then RS03	Grouped with RS06, then RS03

In the radish dataset, the NJ tree reconstructed in this study produced eight bipartitions, whereas the published tree contained nine. Only one bipartition, grouping RS08 and RS09, was shared between the two trees. This limited overlap indicates substantial topological divergence, which likely reflects differences in both input data and tree construction strategies. As discussed above for cotton, such discrepancies are not unexpected, because inversion based trees and sequence based reference phylogenies are influenced by different evolutionary signals.

The inversion matrix used in this study recorded only the presence or absence of inversions and did not incorporate inversion size, genomic position, or locus specific evolutionary relevance. In the original study, Zhang et al. (2021)[78] reported that only two inversions showed a pattern perfectly matching the phylogenetic tree, suggesting that inversion presence or absence alone may provide insufficient resolution for accurate phylogenetic inference in this dataset.

Differences in genome assembly quality may also have contributed to the observed discrepancies. For example, RS01 had a contig N50 of 1.89 Mb, the lowest among the accessions, which may reduce the sensitivity and consistency of structural variant detection and lead to misleading phylogenetic placements.

In contrast to the inversion based NJ tree, the published reference phylogeny was inferred using a maximum likelihood approach applied to 4,464 single copy orthologs using fourfold degenerate site transversions. Sequence based data of this type may provide stronger resolution of evolutionary relationships, particularly among closely related accessions.

4.2.1 Bipartition Randomization Test

To evaluate whether the bipartitions in our NJ trees reflect phylogenetic structure rather than noise, we carried out a randomization test for both the radish and cotton datasets. For each dataset, we generated 100 randomized matrices by shuffling the positions of ones within each row while preserving row sums. We then reconstructed a neighbour-joining tree from each randomized matrix and counted the number of bipartitions shared with the original tree.

In the radish dataset, the randomized trees shared on average only 0.09 bipartitions with the original tree, with a maximum of one. This indicates that bipartition overlap under randomized input is extremely limited. The cotton dataset produced slightly higher values, with a mean of 0.17 and a maximum of two shared bipartitions. Even so, overall overlap remained low. A small number of splits, such as the grouping of 3-79 and Pima90, appeared repeatedly, suggesting that these specific patterns may reflect genuine biological signal.

Table 4.3: Bipartitions shared between the original NJ tree and 100 randomized trees.

Dataset	Mean	SD	Min	Max
Radish	0.09	0.288	0	1
Cotton	0.17	0.403	0	2

As summarized in Table 4.3, these results indicate that the bipartitions observed in the inversion based trees are unlikely to arise purely as random artifacts. Although inversion based trees are not expected to reproduce published reference topologies exactly, a small subset of internal splits recurs across randomizations, suggesting that the inversion matrix contains non random structure that is consistent with genuine evolutionary signal.

The comparison between datasets further supports this interpretation. The cotton dataset produced a clearer signal in this test, with a standard deviation of 0.40, indicating that a subset of randomized cotton trees recovered multiple bipartitions found in the original tree. In contrast, the radish dataset yielded a weaker signal, with a standard deviation of 0.29, indicating that randomized trees rarely approximated the original bipartition set.

Differences in sample quality likely contribute to this pattern. The cotton inversion matrix contains fewer events, yet the corresponding signal appears more consistent. In contrast, assembly quality varies substantially among radish accessions. For example, RS01 has the lowest assembly continuity, with a contig N50 of 1.89 Mb (Zhang et al., 2021, Supplementary Table 1). Such fragmentation can reduce the accuracy and consistency of inversion detection and annotation. When variant calls are incomplete or inconsistent, phylogenetic signal weakens and uncertainty increases in sample placement within the inferred tree. Consistent with this interpretation, Zhang et al. (2021)[78] reported that only a small fraction of inversion events were congruent with the

expected phylogeny, suggesting that inversion presence or absence carries limited evolutionary information in this dataset.

The randomization results nonetheless suggest that the inferred trees are not entirely random, because some bipartitions recur more often than expected under permutation. At the same time, the analysis highlights limitations of bipartition based comparisons for structural variation data. Presence or absence matrices derived from inversions are often sparse and unevenly annotated and can be highly sensitive to small perturbations in the input. Even minor topological differences can therefore lead to complete bipartition mismatches despite broad similarity in overall structure.

Taken together, these issues indicate that bipartition overlap is poorly suited for evaluating tree similarity in the context of structural variation. Methods that tolerate limited topological disagreement provide a more appropriate alternative. The Maximum Agreement Subtree approach focuses on the largest set of taxa that preserve consistent relationships across trees rather than requiring exact split matching. Under conditions of noisy or incomplete structural variant annotation, MAST provides a more stable assessment of phylogenetic structure and helps separate robust relationships from those driven primarily by uncertainty.

4.3 Bootstrap Analysis

To evaluate the robustness of the inversion based trees, we conducted a bootstrap analysis. For each dataset, 10000 bootstrap matrices were generated by resampling the presence or absence matrix with replacement, and a Neighbour-Joining tree was reconstructed for each replicate. The Maximum Agreement Subtree method was then used to quantify the agreement between bootstrap replicate trees and the published reference topology using MAST. The major goal is to see whether the trees we inferred for each pangenome had a better MAST score than the levels suggested by the bootstrap. First, we examine the extent of the MAST values for each pangenome.

The radish dataset showed high topological variability. MAST sizes ranged from 3 to 10, with a mean of 6.04. On average, only about half of the taxa retained consistent relationships when the input matrix was perturbed. This instability is consistent with challenges previously reported for this dataset, including uneven genome assembly quality. For example, RS01 has a contig N50 of 1.89 Mb, which may limit accurate inversion detection and contribute to uncertainty in its phylogenetic placement. Additional noise in structural variant annotation may further weaken the phylogenetic signal.

The cotton dataset showed stronger topological consistency across bootstrap replicates. MAST sizes ranged from 4 to 10, with a mean of 6.56. Although the number of taxa is the same, the cotton trees varied less across resampled datasets. This increased stability may reflect higher assembly quality, more reliable inversion annotations, or the effect of curation applied during dataset construction.

Overall, the bootstrap results indicate some difference between the two datasets: inversion based relationships are less stable in radish than in cotton. While neither dataset fully reproduces the established reference topology, both contain subsets of relationships that are reproducible under resampling. Considered alongside the bipartition and MAST comparisons reported above, these results support the use of multiple complementary metrics to evaluate phylogenetic signal in structural variation data.

Both pangenomes fell near the centre of the bootstrap distributions, suggesting that we cannot conclude that the inferred tree was a better reflection of the data than could expected from the distribution. As discussed above, however, we can derive some information from this analysis.

4.4 MAST and cophenentic correlation analyses of the cotton and radish trees

.

The bipartition and bootstrap analyses above indicate that inversion based trees can capture

reproducible phylogenetic patterns but remain sensitive to noise and annotation inconsistencies. To quantify structural similarity between trees more directly and to identify relationships that persist despite such variation, we applied the Maximum Agreement Subtree method. This approach identifies the largest subset of taxa for which the topological relationships are identical between two trees, providing a conservative measure of shared phylogenetic signal.

For each dataset, we compared the neighbour-joining tree reconstructed from the inversion presence or absence matrix with the corresponding reference tree reported in the original study. Comparisons were restricted to shared taxa, and only topology was evaluated, not branch lengths. This analysis complements earlier bipartition counts and bootstrap summaries by emphasizing relationships that remain topologically stable across trees derived from different data sources.

To clarify the computational procedure underlying this comparison, the following algorithm summarizes the steps used to compute the observed MAST size.

Algorithm: MAST-based tree agreement

Input: Inversion-based tree T_{inv} and reference tree T_{ref}

1. Identify the set L of taxa shared by T_{inv} and T_{ref}.
2. Restrict both trees to the shared taxa.
3. Compute the maximum agreement subtree between the restricted trees.
4. Record the number of taxa retained in the agreement subtree as the observed MAST size.

Figure 4.5 shows the agreement subtree for the radish dataset. The MAST retains seven samples: RS01, RS02, RS03, RS04, RS08, RS09, and RS10, indicating that these accessions share a consistent topology between the inversion based NJ tree and the published reference tree.

Notably, the clustering of RS08 and RS09 is preserved in the MAST, consistent with their shared bipartition and suggesting a relatively stable relationship between these two accessions. In addition, the retained relationships involving RS03 and RS10 and the relationship between RS02 and RS04 indicate that some internal structure is reproducible across both trees. It is also notable that RS01 is retained in the agreement subtree despite having the lowest assembly continuity, with a contig N50 of 1.89 Mb, suggesting that its placement is not entirely driven by assembly fragmentation.

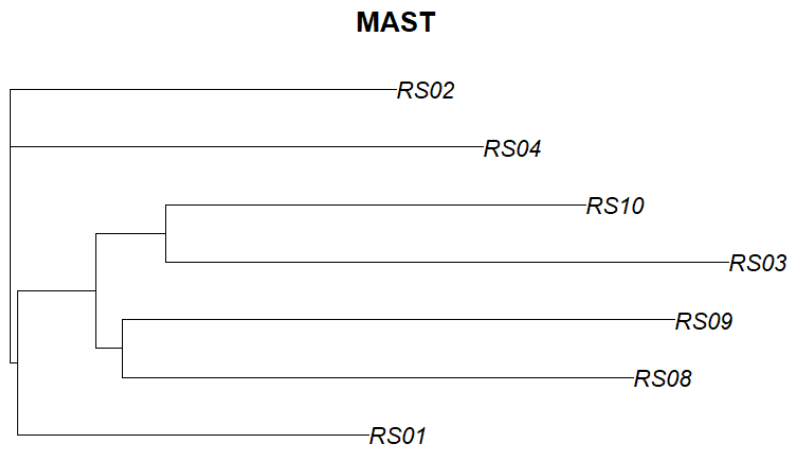

Figure 4.5: Maximum agreement subtree for the radish dataset.

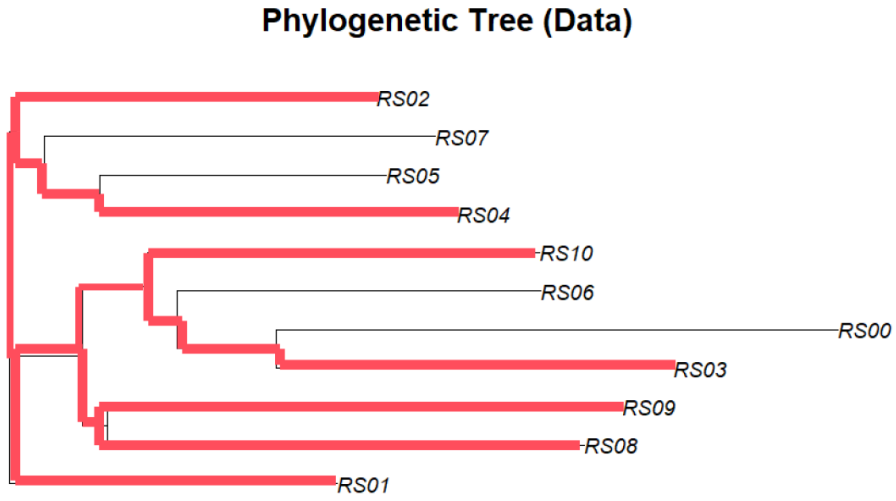

Figure 4.6: Maximum agreement subtree highlighted in the radish NJ phylogeny.

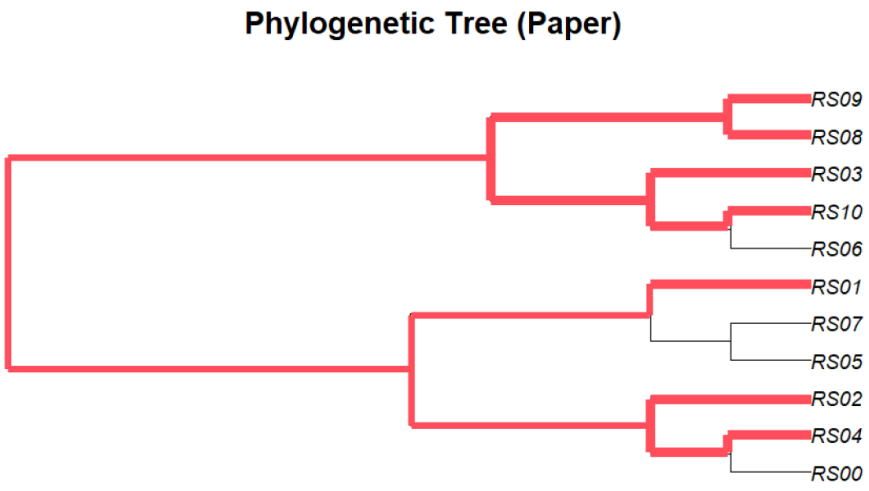

Figure 4.7: Maximum agreement subtree highlighted in the radish reference tree

Figure 4.8 presents the MAST result for the cotton dataset. The agreement subtree retains seven accessions: Gm, Gt, Hai7124, Xinhai21, XLZ7, Tanguis and ZMS24.

Key patterns include the stable clustering of Hai7124 with Xinhai21 and of XLZ7 with ZMS24. The retention of Gm and Gt further indicates that their separation is consistent across both trees.

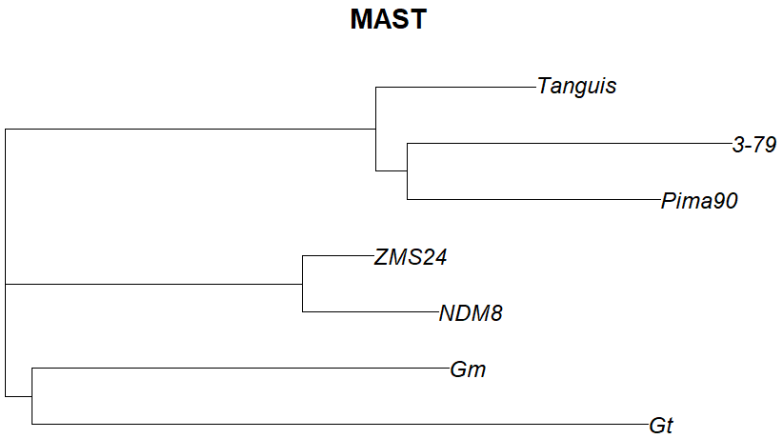

Figure 4.8: Maximum agreement subtree for the cotton dataset.

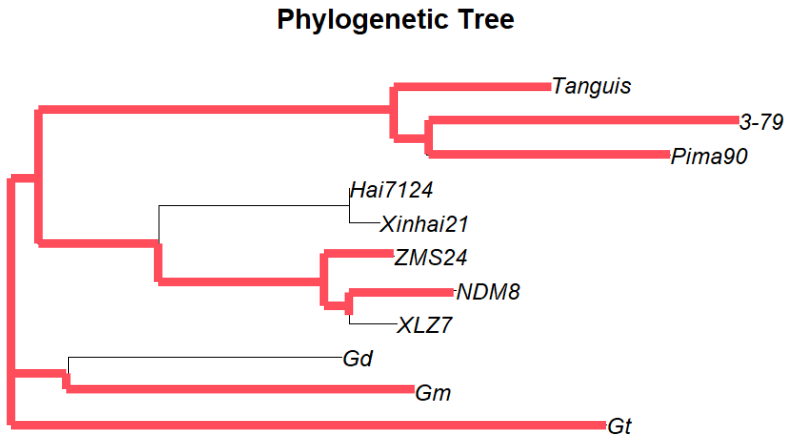

Figure 4.9: Maximum agreement subtree highlighted in the cotton NJ phylogeny.

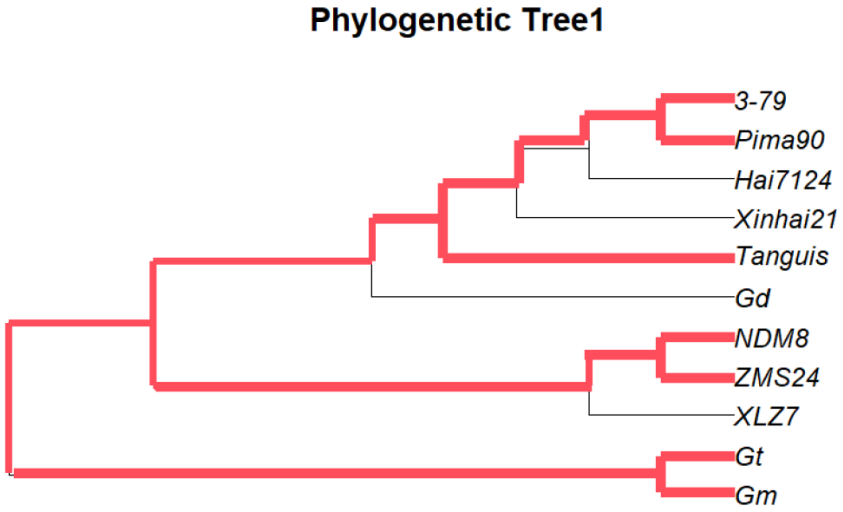

Figure 4.10: Maximum agreement subtree highlighted in the cotton reference tree.

To further assess overall structural similarity, we also calculated co-phenetic correlations between the NJ trees reconstructed in this study and the published reference trees. This measure summarizes the degree to which pairwise distance patterns are preserved between two trees and complements the MAST results by capturing similarity at a broader structural level. There are $n(n-1)/2$ non-diagonal entries in a matrix used for calculating the correlation between two variables with n matched observations. This holds true in comparing the matrix of two trees that are

both produced by neighbour-joining. When one of the trees is an ultrametric, as in the reference genomes for the pangenomes we study, there are only $n - 1$ different leaf-to leaf distances, so that the correlation calculation will involve repeating each term in the ultrametric with several in the neighbour-joining tree. This should not affect the orrelation, although it may impact its significane level. The correlations were 0.68 for radish and 0.67 for cotton, indicating moderate agreement in global distance structure even though several finer scale branching patterns differ.

Table 4.4: Co-phenetic correlations between NJ trees and reference trees.

Dataset	Co-phenetic correlation
Radish	0.68
Cotton	0.67

We note that with both pangenomes the MAST trees both included seven genomes slightly more than half the total of eleven genomes. We will return to this observation in Chapter 5.1.

4.5 Noise Simulation

To systematically evaluate the impact of noise on phylogenetic trees reconstructed from structural variation data, we conducted a controlled noise simulation. Noise was introduced by randomly converting a specified number of zeros to ones in the original binary presence or absence matrix, thereby simulating false positive inversion calls. Noise levels were increased in steps of 30 converted entries. At each noise level, we generated 10 replicate matrices, reconstructed neighbour-joining trees, and compared each replicate tree with the original tree using the Maximum Agreement Subtree method. This design provides a structured assessment of how increasing false positive rates affect topological consistency and therefore informs the robustness of SV based phylogenetic inference.

In the cotton dataset, the baseline MAST size was 11, indicating complete topological agreement between the original and replicate trees. Converting 30 entries reduced the mean MAST size to 9.3. As noise increased, agreement continued to decline and then stabilized at approximately 6.0 once more than 420 zeros had been converted. This pattern suggests that the phylogenetic structure inferred from cotton inversion data is initially robust but becomes increasingly sensitive to error as noise accumulates. The plateau at higher noise levels further indicates that a subset of relationships remains detectable even under substantial perturbation.

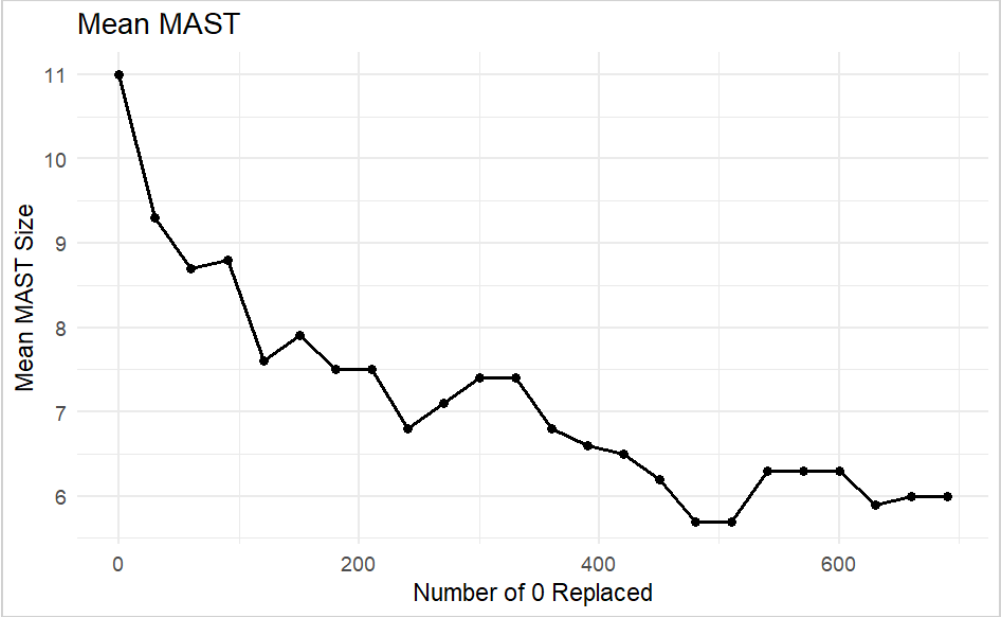

Figure 4.11: Mean MAST size across noise levels in the cotton dataset.

In the radish dataset, the baseline MAST size was also 11. However, topological consistency decreased more sharply after noise was introduced, with the mean MAST size dropping to 8.8 after 30 entries were converted. Further increases in noise led to a more gradual decline, with the mean MAST size stabilizing around 6.0 after approximately 240 to 300 conversions.

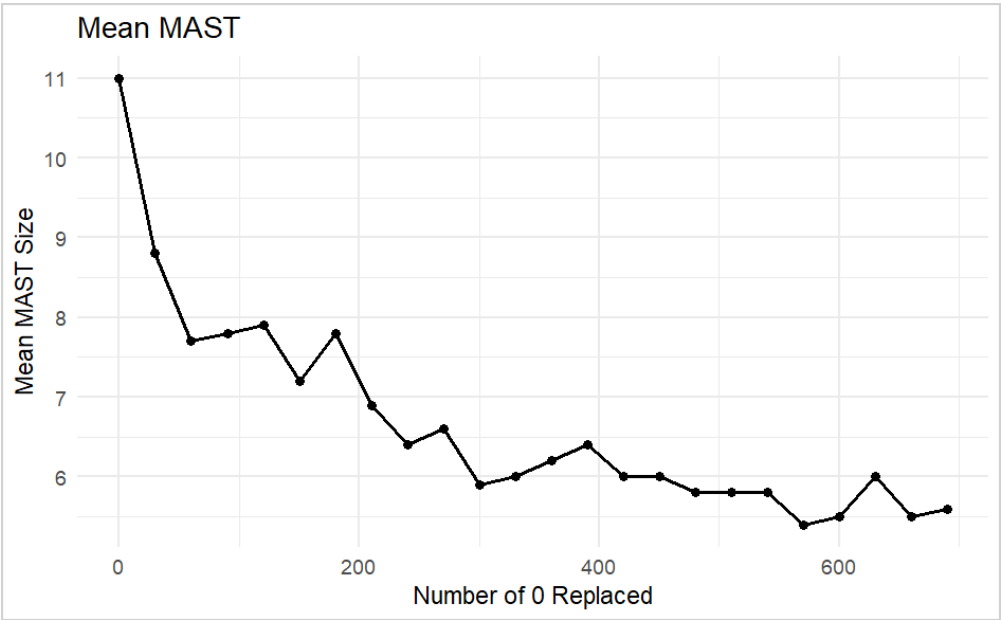

Figure 4.12: Mean MAST size across noise levels in the radish dataset.

Together, these results highlight differences in noise tolerance between datasets and underscore

the importance of data quality for SV based phylogenetic reconstruction. The simulation provides empirical evidence for how phylogenetic signal degrades under artificial perturbation, while also indicating that a subset of relationships persists across noise levels and therefore represents comparatively robust signal.

Chapter 5

Degrading Dollo

5.1 Simulation design

In the previous chapter, we saw that although the MAST values comparing the structural variation tree and the reference tree was somewhat better than random, the difference was not great and remained within the confidence interval determined by the resampling distribution of the variant scores. As we have stressed, there are a number of explanations for this, one of which was a Dollo's law constraining reversed mutations. Without insisting on this versus the many other biology-based explanations, we carried out a simulation experiment to see how departures from strict Dollo evolution would degrade the topological structure of a phylogeny.

To better understand how changes in tip, or leaf, states influence phylogenetic structure, we generated simulations based on the `rtree` routine for random binary branching trees. Inherent in this tree is a specific distance matrix. For each of 100 replicates, we generated an rooted phylogenetic tree with 32 tips. All tips and internal nodes were initially assigned the state 0. One internal node was then randomly selected and assigned the state 1, which was propagated to all descendant nodes. This procedure mimics the emergence of a signal in an ancestral lineage that is inherited by all downstream taxa. This design follows a Dollo-style assumption: a state 0 at a node is generally inherited by descendants, but if it is lost in node x it becomes a 1, as do all of the descendants x, and cannot revert to the 0 state.

Because the selected node can occur anywhere within the tree, the proportion of tips labeled as 1 varied across replicates. In most cases, this proportion remained below 25%, which is expected because most internal nodes in a randomly generated tree subtend relatively few descendants.

The following figures are examples with varying proportion of tips labelled 1:

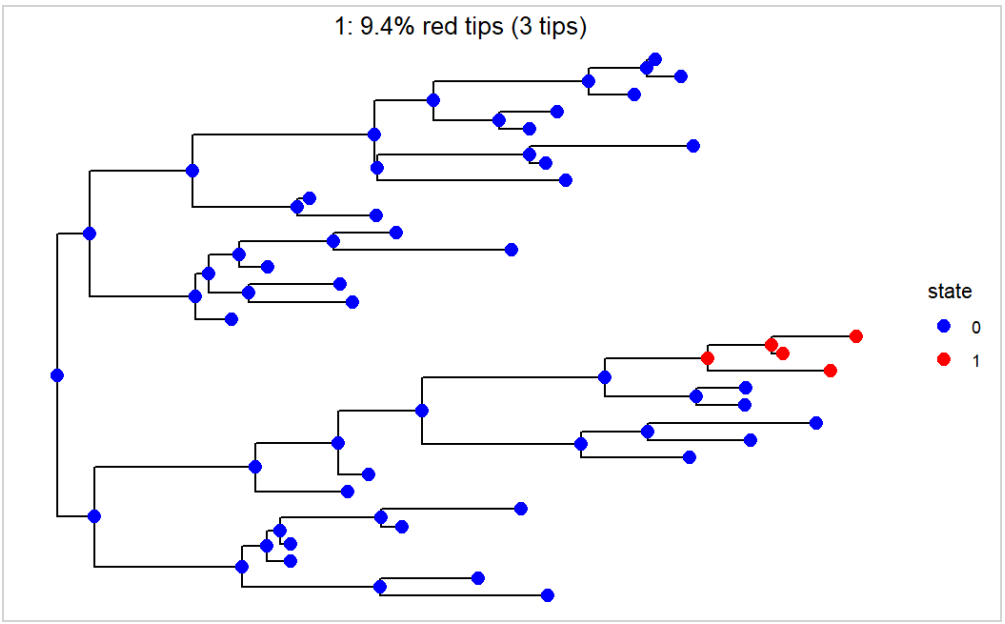

Figure 5.1: Example tree with 9.4% red tips (3 out of 32).

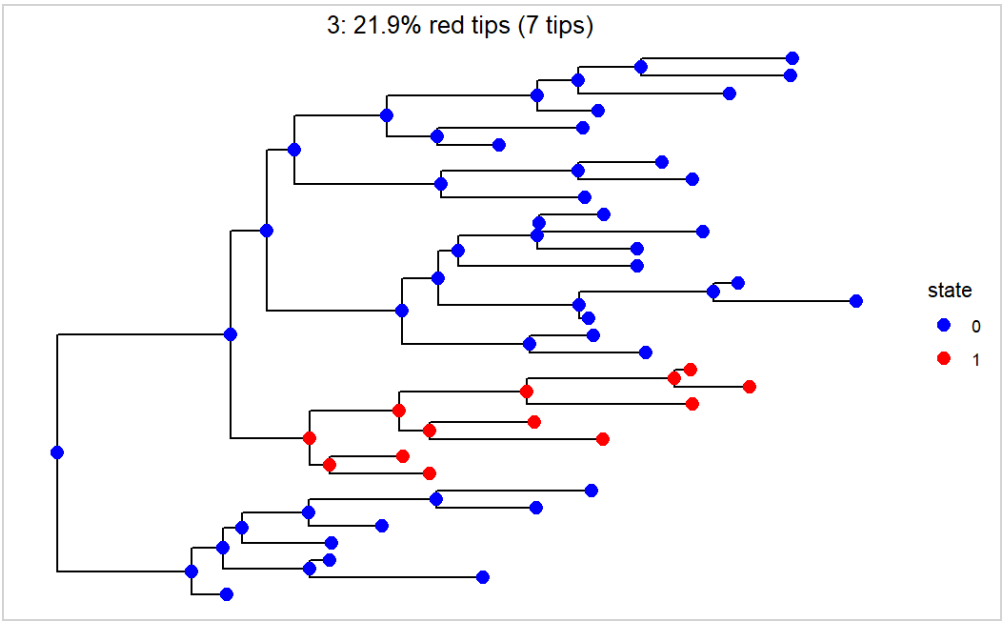

Figure 5.2: Example tree with 21.9% red tips (7 out of 32).

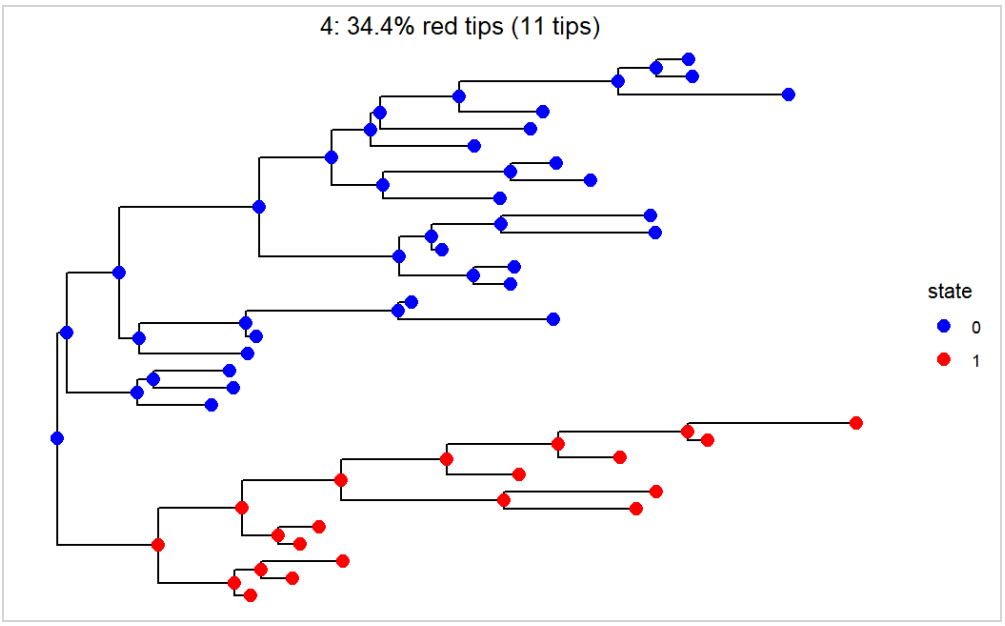

Figure 5.3: Example tree with 34.4% red tips (11 out of 32).

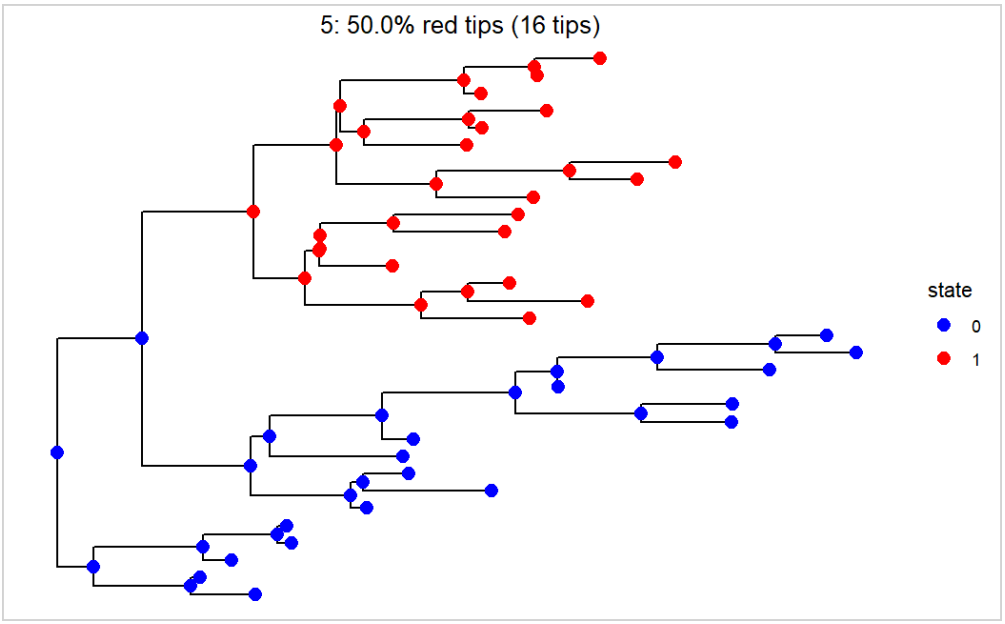

Figure 5.4: Example tree with 50.0% red tips (16 out of 32).

To simulate increasing disturbance, we progressively flipped tip states from their original values. For each integer from 1 to 32, we randomly selected that number of tips and reversed their states. Each flipping level was repeated 15 times to reduce sampling variability. For each replicate, a new neighbour-joining tree was reconstructed from the perturbed tip states, and its similarity to the original tree was quantified using the Maximum Agreement Subtree method. We recorded the

mean MAST size across the 15 repetitions as the representative value for that disturbance level. This procedure was repeated for all 100 trees. Finally, each tree was assigned to one of twenty groups based on the initial proportion of tips labeled as 1,

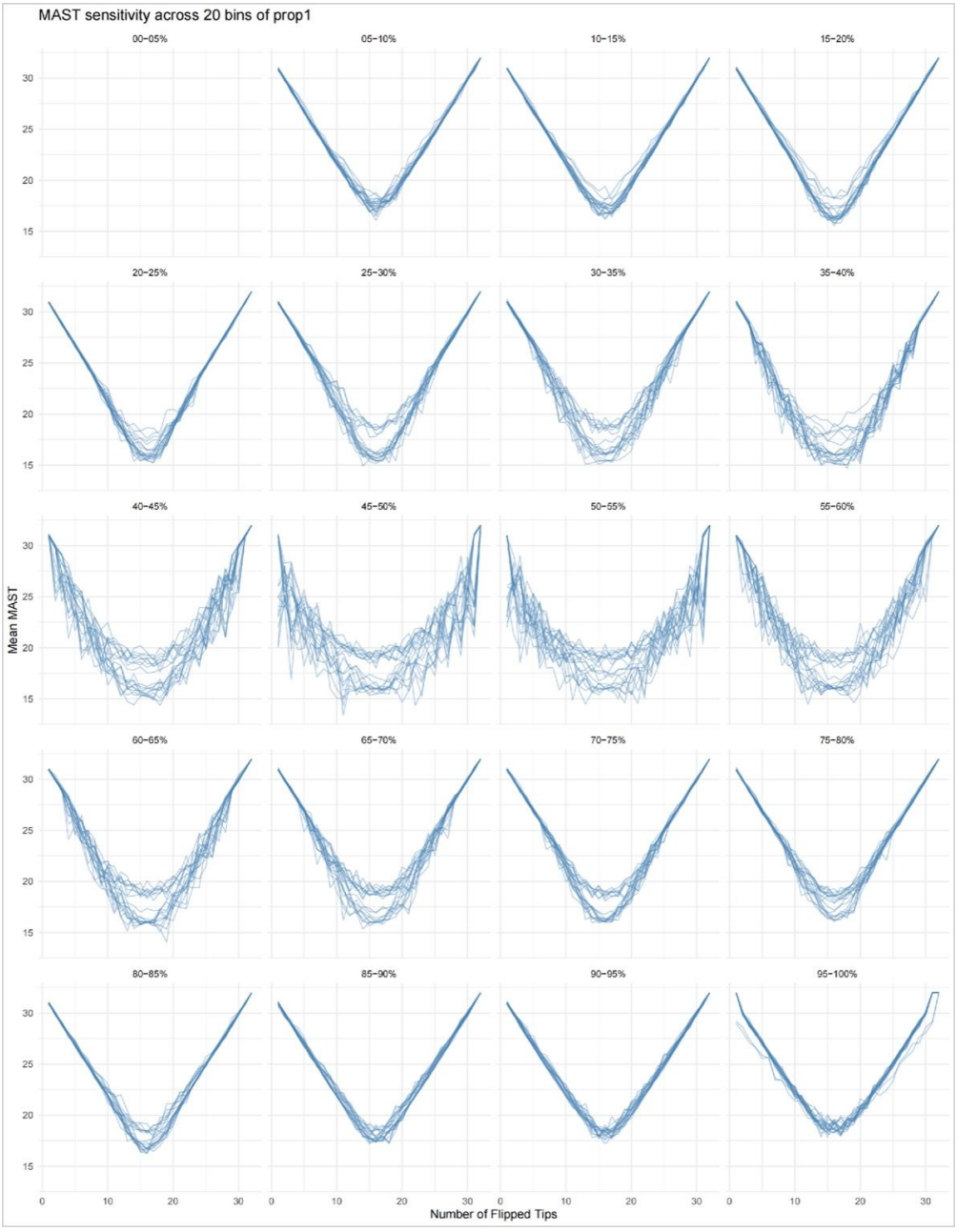

Figure 5.5: Mean MAST size under increasing tip flipping, 20 groups.

Across all groups, MAST size initially decreased as more tips were flipped, indicating increasing disagreement between the perturbed tree and the original (Figure 5.5). After reaching a minimum, typically near half of the tips, MAST size increased again. When all tip states were reversed, the

resulting binary pattern became the complement of the original, which can yield a reconstructed tree that partially realigns with the original topology. This symmetry arises because Hamming distance is invariant under global bit complementation, that is, $d(x, y) = d(\bar{x}, \bar{y})$. When all tip states are flipped, pairwise distances remain unchanged, and the reconstructed topology therefore realigns with the original tree.

Although the overall curve shape was similar across groups, sensitivity to perturbation differed. The extreme groups showed smoother and more stable trajectories. In contrast, the intermediate groups showed greater variability and stronger fluctuations, particularly at low and intermediate flipping levels. This suggests that trees with a more balanced initial distribution of ones and zeros may be more sensitive to small changes, whereas extreme distributions are more resilient under the same disturbance. One explanation is that extreme state distributions provide less discriminating information. When most tips share the same state, the signal becomes more redundant and local perturbations have a smaller effect on overall tree structure, leading to smoother changes in MAST size as noise increases.

In Chapter 4.4 we observed that MAST scores for the two pangenomes were slightly higher than half the number of genomes. In Chapter4.5, we observed that the introduction of noise sufficient to obscure the relationship between two trees resulted in MAST half the number of genomes. In this chapter, we similarly observe the minimum MAST is half the total number of leaves of the tree, athough there are three times the number of genomes in the simulated pangenomes as in the empirically derived ones. This, however, cannot be taken as a generaal result since it is known that in the limit, the smallest MAST score is about the square root of the number of genomes [89].

5.1.1 Co-phenetic correlation across twenty groups

We also analyzed how data perturbation affected co-phenetic correlation. For each simulated dataset, we computed co-phenetic distance matrices for the original and perturbed neighbour-joining trees and calculated the Pearson correlation between them. At each perturbation level, values were averaged across 15 replicates.

Grouped by the initial proportion of ones, the resulting correlation curves showed a clear U shaped pattern (Figure 5.6). Correlations decreased as noise increased and reached very low values, and in some cases slightly negative values, at intermediate perturbation levels. When most tips were flipped, correlations increased again. This response is consistent with the behavior observed for MAST, with both metrics showing their greatest instability at intermediate perturbation levels.

The co-phenetic correlation curves varied more smoothly than those obtained from MAST. Co-phenetic correlation aggregates information across all pairwise distances in the tree, whereas MAST changes in a stepwise manner. Even small topological differences can remove entire clades

from the agreement subtree, leading to abrupt changes in MAST size. As a result, MAST can react strongly to minor perturbations, whereas co-phenetic correlation tends to change more gradually. Despite these differences, both metrics capture the same overall pattern of topological stability.

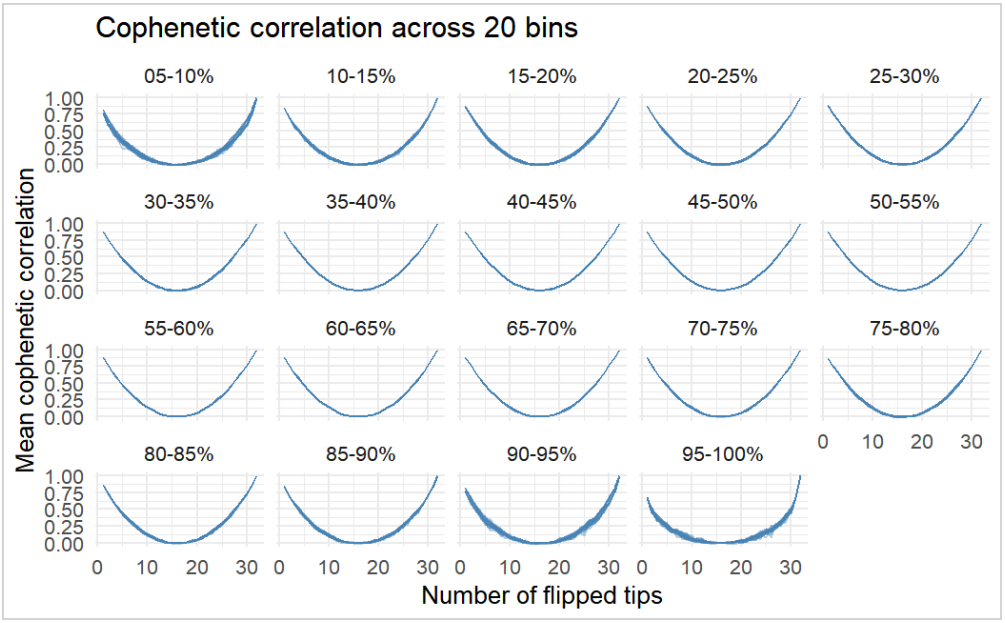

Figure 5.6: Mean co-phenetic correlation under increasing tip flipping, grouped into twenty bins by the initial proportion of ones.

5.1.2 Local structural explanation: preserved sister pairs

The analysis across twenty groups revealed substantial variation in tree responses to tip perturbation. It also suggests that factors beyond tip state proportions, such as the preservation of specific local topological features, may explain differences in stability. To explore this possibility, we evaluated several candidate factors, including state overlap, whether flipping occurred near terminal or internal branches, and the size of original subtrees. One factor that proved particularly informative was the preservation of sister pairs.

We first examined this relationship in the 40 to 45 percent group. Sister pairs refer to pairs of tips that directly descend from the same internal node. We counted how many sister pairs in the original tree remained intact in the reconstructed NJ tree after tip flipping. For each tree, we treated the number of preserved sister pairs as a local structural stability index and examined its association with MAST size. When k, the number of flipped tips, equaled 15, the correlation coefficient between mean MAST and mean preserved sister pairs across twenty trees reached 0.81. This result indicates that, within this group, sister pair retention plays a major role in maintaining overall tree structure.

We then extended the analysis to all twenty subgroups. For each subgroup and each flip level from 1 to 32, we computed the correlation between mean MAST and the mean number of preserved sister pairs. In many subgroups, particularly those with balanced state proportions between 40 and 60 percent, the correlation remained above 0.7. This pattern suggests that sister pair integrity is an important determinant of robustness under perturbation.

In contrast, in extreme subgroups such as 0 to 5 percent and 95 to 100 percent, correlations were weak. One explanation is that when tip states are nearly uniform, perturbations affect the tree more evenly, and disruption of any single pair has limited impact on overall topology.

By contrast, trees in the intermediate proportion ranges tend to contain more locally informative structure. Flipping a small number of critical sister pairs can disrupt symmetry and lead to sharp drops in MAST. Whether key sister pairs were disrupted during flipping often explains differing behaviors within the same subgroup. For example, in the 45 to 50 percent subgroup, some trees had sister pairs concentrated within particular clades. When those pairs remained unchanged, MAST curves were comparatively stable. In other trees, where flipping disrupted multiple key sister pairs, MAST sizes declined rapidly.

Overall, these results indicate that sister pairs, as a local structural unit, provide a useful indicator for understanding tree stability under disturbance. Their retention helps explain fluctuations in MAST across trees and groups and provides a practical structural criterion for evaluating the reliability of trees reconstructed from structural variation data.

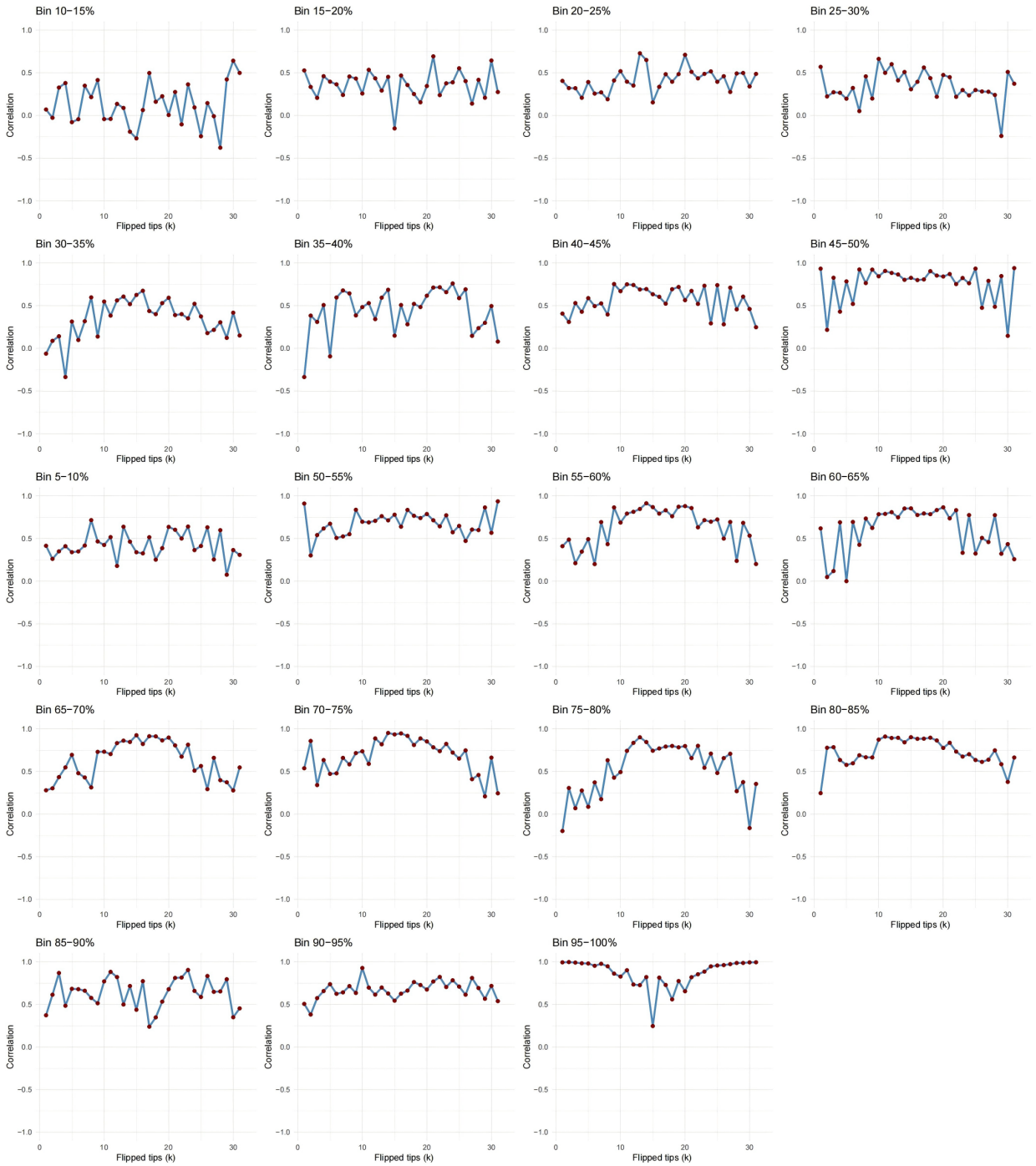

Figure 5.7: Correlation between preserved sister pairs and mean MAST across twenty subgroups.

Chapter 6

Conclusion

This thesis examined within-pangenome phylogeny based on structural variations. Using binary presence or absence matrices of inversions from two datasets, radish and cotton, we reconstructed neighbour-joining trees and compared them with published reference phylogenies using two complementary measures: Maximum Agreement Subtree size and co-phenetic correlation.

Starting from published pangenome resources, we extracted inversion and other structural-variation data from supplementary tables and figures and converted them into binary matrices. For radish, we used inversion calls and presence information reported in the supplementary material of Zhang et al [79], whereas for cotton we digitized the large inversions shown in the pan inversion map of Jin et al [78]. From these matrices, we reconstructed neighbour joining trees and compared them with the reference trees reported in the original studies, which were based on sequence data. This first set of analyses addressed the extent to which structural variant and inversion-based trees resemble the corresponding reference phylogenies.

Across both species, our structural variant-based trees shared some reproducible internal structure with the reference trees, although agreement was far from complete. Maximum agreement subtrees retained a subset of accessions whose relationships were stable across methods, and co-phenetic correlations between our trees and reference trees were higher than values obtained under matrix randomization. But only a small number of bipartitions were shared and several internal relationships differed.

Alongside these empirical comparisons, we evaluated how best to measure agreement between trees reconstructed from sparse presence or absence data. Bipartition overlap treats internal splits as exact matches and can react strongly to small local changes in topology. In our setting, minor rearrangements can reduce shared splits to zero even when overall structure remains similar. The maximum agreement subtree provides a more stable alternative by identifying the largest subset of taxa that preserves consistent relationships across trees and by effectively setting aside unstable tips. Co-phenetic correlation provides a complementary view based on all pairwise tip distances

and summarizes similarity in global clustering even when individual splits differ.

The simulation studies placed these observations into a broader context. We generated random trees under a simple Dollo style scenario, assigned binary states along the tree, and then progressively flipped tip states away from the original configuration. Across replicates, mean MAST size followed a characteristic pattern: it declined as perturbations accumulated, reached a minimum at intermediate noise levels, and then increased again when most tips had been reversed. This U shaped response reflects the symmetry between the original and fully flipped configurations and shows that substantial state changes do not necessarily remove all similarity between inferred trees. Grouping simulated trees by the initial proportion of ones further showed that balanced starting configurations tend to be more sensitive to perturbation than configurations in which most tips share the same state.

Several limitations of this study should be acknowledged. Inversion and structural variant information was extracted from published figures and tables rather than obtained by calling variants directly from raw sequencing data. In cotton, we considered only large inversions and did not include smaller rearrangements. In radish, inversion detection sensitivity likely varies among accessions because assembly quality is uneven. We also focused on a single reconstruction strategy based on Hamming distances and neighbour-joining. Alternative distance measures and model based approaches for binary data may yield different topologies and could change the degree of agreement with reference trees.

Despite these limitations, our work provides a systematic view of what structural variants can reveal about relationships within pangenomes.

Future work could extend this study in two closely related directions. First, the same framework could be applied some of the additional pangenome datasets listed in Chapter 3 to test whether the patterns observed here also hold in other species.

Second, tree comparison could be refined beyond the use of maximum agreement subtree size alone. While MAST provides a simple summary of shared topology, it reduces agreement to a single value and does not capture how similarity may be distributed across multiple smaller substructures.

Appendix A

Radish Presence/Absence Matrix

The original table was arranged horizontally, but here it is transposed, displayed vertically for easier formatting.

Table A.1: Radish presence/absence matrix transposed

	RS01	RS02	RS03	RS04	RS05	RS06	RS07	RS08	RS09	RS10	RS00
1	0	0	0	0	1	0	0	0	1	0	1
2	0	0	0	0	0	1	0	1	0	0	1
3	0	0	1	0	0	0	0	0	0	1	1
4	1	0	1	0	1	1	1	1	1	1	1
5	1	1	1	1	1	1	1	0	1	1	1
6	1	1	1	1	1	1	1	0	1	1	1
7	0	0	1	0	1	0	1	0	0	0	1
8	1	1	1	1	1	0	1	1	1	1	1
9	1	1	0	0	1	1	1	0	1	0	1
10	1	1	1	1	1	1	1	1	0	1	1
11	0	0	1	1	1	1	1	0	0	1	1
12	0	1	0	1	1	1	0	1	1	0	1
13	0	0	0	0	0	0	1	0	0	1	1
14	1	1	1	1	1	1	1	0	1	1	1
15	1	1	1	0	1	1	0	0	1	1	1
16	1	1	0	1	0	1	1	1	1	1	1
17	1	1	1	1	1	1	1	0	1	1	1

Continued on next page

Table A.1: Radish presence/absence matrix transposed

	RS01	RS02	RS03	RS04	RS05	RS06	RS07	RS08	RS09	RS10	RS00
18	1	0	1	1	1	1	1	1	1	0	1
19	1	1	1	0	0	1	0	1	0	0	1
20	1	0	1	0	1	1	0	1	0	0	1
21	1	1	1	0	1	1	1	1	1	1	1
22	0	0	0	0	0	0	1	0	0	1	1
23	0	0	0	1	1	0	0	0	0	0	1
24	1	0	0	1	1	0	1	1	1	0	1
25	0	0	0	0	0	0	0	1	1	0	1
26	0	0	0	0	0	0	1	1	0	0	1
27	0	0	0	1	1	0	0	0	0	0	1
28	0	0	0	1	0	0	1	1	0	0	1
29	0	0	1	0	0	0	1	0	1	0	1
30	1	0	1	0	0	0	0	0	0	1	1
31	0	0	1	0	0	1	0	0	1	0	1
32	1	0	0	1	1	0	1	0	0	0	1
33	0	0	1	0	1	0	0	1	1	0	1
34	0	0	0	0	1	0	1	0	1	0	1
35	1	1	1	1	1	1	1	1	0	1	1
36	1	0	0	0	0	0	0	1	0	1	1
37	1	1	1	1	1	1	1	1	1	0	1
38	0	1	0	0	0	1	0	0	1	1	1
39	1	1	1	1	1	1	1	1	0	1	1
40	1	0	1	1	1	1	1	1	1	1	1
41	1	1	1	1	0	0	1	1	1	1	1
42	1	1	0	1	1	0	1	1	1	0	1
43	0	0	1	0	0	1	0	1	1	1	1
44	0	0	1	0	0	1	0	1	1	0	1
45	1	1	0	1	1	0	0	0	0	0	1
46	0	0	1	0	0	1	0	1	0	1	1
47	0	0	1	0	0	1	0	1	1	1	1
48	0	1	0	1	1	0	1	0	0	0	1

Continued on next page

Table A.1: Radish presence/absence matrix transposed

	RS01	RS02	RS03	RS04	RS05	RS06	RS07	RS08	RS09	RS10	RS00
49	0	0	1	0	0	1	0	0	0	1	1
50	0	0	0	1	0	0	1	0	0	0	1
51	1	0	0	1	1	0	1	0	0	0	1
52	1	1	0	1	1	1	0	0	1	1	1
53	0	1	1	1	1	1	0	1	1	1	1
54	0	0	0	1	0	1	0	0	0	0	1
55	0	0	1	0	0	1	0	0	0	0	1
56	0	0	0	0	0	1	0	1	1	0	1
57	1	0	0	0	0	0	0	0	0	1	1
58	0	0	0	0	0	0	0	1	0	1	1
59	0	0	1	0	0	1	1	0	1	0	1
60	0	0	1	1	1	0	0	0	0	0	1
61	0	0	1	0	0	0	1	0	1	1	1
62	0	0	1	0	0	0	0	0	0	0	1
63	0	0	1	1	0	0	0	0	0	1	1
64	1	1	1	1	1	1	1	0	1	1	1
65	0	0	1	1	0	0	0	0	0	0	1
66	0	0	0	0	0	1	0	0	1	0	1
67	0	1	0	0	0	1	0	0	1	1	1
68	1	1	1	0	1	0	1	1	1	1	1
69	1	1	1	1	1	1	1	1	0	1	1
70	0	1	1	1	1	1	1	1	0	1	1
71	0	0	0	1	0	0	1	0	0	0	1
72	0	0	0	0	1	1	0	0	0	0	1
73	0	0	1	1	0	0	0	0	0	1	1
74	0	1	1	0	0	1	1	1	0	1	1
75	1	1	0	1	0	0	1	1	0	0	1
76	1	1	0	1	0	0	1	1	0	0	1
77	0	0	0	0	0	0	0	1	0	1	1
78	1	1	1	1	0	1	1	1	0	1	1
79	1	1	1	1	1	1	1	0	0	1	1

Continued on next page

Table A.1: Radish presence/absence matrix transposed

	RS01	RS02	RS03	RS04	RS05	RS06	RS07	RS08	RS09	RS10	RS00
80	1	1	1	1	1	1	1	0	0	0	1
81	1	0	1	0	0	1	0	0	0	0	1
82	1	1	1	1	1	1	1	1	0	1	1
83	1	0	1	1	1	1	1	1	1	1	1
84	0	0	1	0	1	0	0	0	0	0	1
85	1	1	1	1	1	1	0	1	1	1	1
86	1	1	1	0	1	1	1	1	1	1	1
87	0	1	1	0	1	0	1	0	1	1	1
88	0	1	0	0	0	1	0	0	1	1	1
89	0	0	0	0	0	1	0	0	0	1	1
90	0	0	1	0	0	1	1	0	0	0	1
91	0	0	0	1	1	0	0	1	0	0	1
92	0	0	0	1	0	0	0	1	0	0	1
93	0	0	1	0	0	0	0	0	0	1	1
94	0	0	0	0	0	1	0	0	1	0	1
95	1	1	0	0	0	0	0	0	1	0	1
96	1	0	1	1	1	1	0	1	0	1	1
97	0	0	1	0	0	0	0	0	1	0	1
98	1	0	0	0	1	0	1	0	0	0	1
99	1	0	0	0	0	0	0	1	1	0	1
100	1	1	0	1	1	1	1	1	1	1	1
101	0	0	1	0	0	0	1	0	0	1	1
102	1	0	1	1	1	1	0	1	1	1	1
103	1	1	0	1	1	1	1	1	1	1	1
104	0	1	0	0	0	1	0	0	0	0	1
105	1	0	1	0	0	0	1	0	1	1	1
106	0	1	0	1	0	0	0	1	0	0	1
107	0	0	0	1	0	1	1	0	0	1	1
108	1	1	1	0	1	1	1	1	1	1	1
109	0	1	1	1	1	0	1	0	1	0	1
110	0	0	1	0	0	1	0	0	0	0	1

Continued on next page

Table A.1: Radish presence/absence matrix transposed

	RS01	RS02	RS03	RS04	RS05	RS06	RS07	RS08	RS09	RS10	RS00
111	0	0	1	0	0	1	1	0	0	0	1
112	0	1	0	0	0	0	0	0	1	0	1
113	0	1	0	0	1	1	0	1	0	0	1
114	0	0	1	1	1	1	1	1	1	1	1
115	0	0	0	0	0	0	0	1	1	0	1
116	0	0	0	0	1	0	0	0	0	1	1
117	0	0	0	0	0	0	0	1	0	1	1
118	1	1	1	0	1	1	1	0	0	0	1
119	0	0	1	0	0	0	0	1	1	0	1
120	0	1	1	0	1	1	1	0	1	0	1
121	0	0	1	0	0	0	0	0	1	0	1
122	0	0	1	0	0	0	1	0	0	0	1
123	0	0	1	1	0	1	1	1	1	1	1
124	0	1	1	1	1	0	0	1	0	0	1
125	1	0	1	1	1	1	0	1	0	1	1
126	1	1	0	0	0	0	0	0	1	0	1
127	1	1	1	0	1	1	1	1	1	1	1
128	1	0	1	0	0	0	1	1	0	1	1
129	0	1	1	1	1	1	1	1	1	1	1
130	1	1	1	1	1	1	1	1	0	1	1
131	0	0	1	0	0	1	0	0	0	0	1
132	1	1	1	1	1	1	1	1	0	1	1
133	0	1	1	1	1	1	1	0	0	1	1
134	0	1	1	0	1	0	0	0	0	0	1
135	1	1	0	1	1	1	1	1	1	1	1
136	1	0	0	1	1	1	1	1	1	1	1
137	1	1	1	1	1	1	1	0	1	1	1
138	1	0	1	1	1	0	0	0	0	1	1
139	1	0	0	0	0	0	0	1	0	1	1
140	1	1	0	1	1	1	1	1	1	1	1
141	0	1	0	1	0	1	1	0	0	1	1

Appendix B

Cotton Presence/Absence Matrix

Table B.1: Cotton presence/absence matrix transposed

	Gt	Gm	Gd	XLZ7	ZMS24	NDM8	Tanguis	Xinhai21	Hai7124	Pima90	3-79
1	1	0	0	0	0	0	0	0	0	0	0
2	0	1	1	0	0	0	1	1	1	0	0
3	0	0	0	0	0	0	1	0	0	0	0
4	0	0	0	0	0	0	0	0	0	0	1
5	1	0	0	0	0	0	0	0	0	0	0
6	0	1	0	0	0	0	0	0	0	0	0
7	0	0	1	0	0	0	1	0	0	1	1
8	1	1	1	0	0	0	1	1	1	1	1
9	0	0	0	1	1	1	0	0	0	0	0
10	0	0	0	0	0	0	0	0	0	1	0
11	1	0	1	0	0	0	1	1	1	0	0
12	0	1	0	0	0	0	0	0	0	0	0
13	0	0	0	0	0	0	0	0	0	1	0
14	1	0	1	0	0	0	1	0	0	1	1
15	0	0	0	0	0	0	0	0	0	0	1
16	0	0	0	0	0	0	1	0	0	1	1
17	0	0	1	0	0	0	0	0	0	0	1
18	0	1	0	0	0	0	0	0	0	0	0
19	1	0	0	0	0	0	0	0	0	0	0

Continued on next page

Table B.1: Cotton presence/absence matrix transposed

	Gt	Gm	Gd	XLZ7	ZMS24	NDM8	Tanguis	Xinhai21	Hai7124	Pima90	3-79
20	1	0	0	1	0	1	0	0	0	0	0
21	0	1	0	0	0	0	0	0	0	0	0
22	0	0	0	0	0	0	1	1	1	1	0
23	1	0	0	0	0	0	0	0	0	0	0
24	0	0	0	0	0	1	0	0	0	0	0
25	1	0	0	0	0	0	0	0	0	0	1
26	0	0	1	0	0	0	0	0	0	0	1
27	1	0	1	0	0	0	0	0	0	1	1
28	1	1	0	0	0	0	0	0	0	0	0
29	0	0	1	0	0	0	0	0	0	0	0
30	0	0	0	0	0	0	1	0	0	1	1
31	0	0	0	0	0	0	0	1	1	0	0
32	0	0	0	0	1	1	0	0	0	0	0
33	0	0	0	1	0	0	0	0	0	0	0
34	0	0	1	0	0	0	1	1	1	1	1
35	1	1	1	1	1	1	1	0	0	1	1
36	0	0	0	1	1	1	0	0	0	0	0
37	0	1	1	0	0	0	0	0	0	0	0
38	1	0	0	0	0	0	0	0	0	0	0
39	0	0	0	0	0	0	0	0	0	1	1
40	1	0	0	0	0	0	0	0	0	0	0
41	0	0	1	0	0	0	0	0	0	0	0
42	1	0	0	0	0	0	0	0	0	0	0
43	1	1	1	0	0	0	0	0	0	0	0
44	0	0	0	0	0	0	1	0	0	1	1
45	1	0	0	0	0	0	0	0	0	0	0
46	0	1	0	0	0	0	0	0	0	0	0
47	0	0	1	0	0	0	0	0	0	1	0
48	0	0	0	0	0	1	0	0	0	0	1
49	1	1	1	0	0	0	1	0	0	1	1
50	0	0	0	0	0	0	1	0	0	1	1

Continued on next page

Table B.1: Cotton presence/absence matrix transposed

	Gt	Gm	Gd	XLZ7	ZMS24	NDM8	Tanguis	Xinhai21	Hai7124	Pima90	3-79
51	0	1	1	0	0	0	0	1	1	0	0
52	1	0	0	0	0	0	0	0	0	0	0
53	1	0	0	0	0	0	0	0	0	0	0
54	1	0	0	0	0	0	0	0	0	0	0
55	0	0	0	0	0	0	0	0	0	1	0
56	0	0	0	0	0	0	1	0	0	1	1
57	0	0	0	0	0	0	1	0	0	1	1
58	0	1	1	0	0	1	0	0	0	0	0
59	1	1	1	0	0	0	1	0	0	1	1
60	0	0	0	0	0	0	1	0	0	0	1
61	1	1	1	0	0	0	0	0	0	1	0
62	0	0	0	0	0	0	0	1	1	0	0
63	1	1	1	0	0	0	1	0	0	1	1
64	0	0	0	0	0	0	1	0	0	0	1
65	0	0	0	0	0	0	0	0	0	1	0
66	0	0	0	0	0	0	0	0	0	1	0
67	1	0	0	0	0	0	0	0	0	0	0
68	0	0	0	0	0	0	1	0	0	1	1
69	0	0	0	0	0	0	1	0	0	1	1
70	0	1	0	0	0	0	0	0	0	0	0
71	1	1	0	1	0	1	0	0	0	0	0
72	0	0	1	0	0	0	0	0	0	0	0
73	0	0	0	0	0	0	0	0	0	0	1
74	0	0	0	0	0	0	1	0	0	1	1
75	0	1	0	0	0	0	0	0	0	0	0
76	0	1	1	0	0	0	1	0	0	1	1
77	0	0	0	0	0	0	0	0	0	0	1
78	1	0	0	0	0	0	0	0	0	0	0
79	0	0	0	0	1	0	1	0	0	0	0
80	0	0	0	0	0	0	0	1	0	0	0
81	1	0	0	0	0	0	1	0	0	1	1

Code Availability

All scripts used for tree reconstruction, MAST comparison, permutation testing, and simulation analyses are publicly available at:

`https://github.com/xup6340-creator/inversion-phylogeny`

The repository contains annotated R scripts sufficient to reproduce the analyses described in this thesis.

Bibliography

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

[2] Sokal RR, Rohlf FJ. The comparison of dendrograms by objective methods. *Taxon*. 1962;11:33–40.

[3] Gordon AD. On the assessment and comparison of classifications. In: Grassle JF, editor. *Numerical Taxonomy*. New York: Academic Press; 1980. p. 149–160.

[4] Finden CR, Gordon AD. Obtaining common pruned trees. *J Classif*. 1985;2:255–276.

[5] Steel M, Warnow T. Kaikoura tree theorems: computing the maximum agreement subtree. *Inf Process Lett*. 1993;48(2):77–82.

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

[7] 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.

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

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

[10] 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.

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

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

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

[14] Eizenga JM, Novak AM, Sibbesen JA, et al. Pangenome graphs. *Annu Rev Genomics Hum Genet*. 2020;21:139–162.

[15] 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.

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

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

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

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

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

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

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

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

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

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

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

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

[28] 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.

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

[30] 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.

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

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

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

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

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

[36] Alon N, Snir S, Yuster R. On the compatibility of quartet trees. *SIAM J Discrete Math*. 2014; 28: 10.1137/130941043.

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

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

[39] Baele G, Ji X, Hassler GW, et al. BEAST X for Bayesian phylogenetic, phylogeographic and phylodynamic inference. *Nat Methods*. 2025; 22: 1653–1656.

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

[41] Kuhner MK, Felsenstein J. A simulation comparison of phylogeny algorithms under equal and unequal evolutionary rates. *Molecular Biology and Evolution*. 1994; 11: 459–468.

[42] Bourque M. *Arbres de Steiner et réseaux dont certains sommets sont à localisation variable*. PhD thesis. Université de Montréal; 1978.

[43] Robinson DF, Foulds LR. Comparison of phylogenetic trees. *Math Biosci*. 1981;53:131–147.

[44] Sand A, Brodal GS, Fagerberg R, et al. A practical O(n $\log^2$ n) time algorithm for computing the triplet distance on binary trees. *BMC Bioinformatics*. 2013;14(Suppl 2):S1.

[45]

[46] Calamoneri T, di Mambro A, Sinaimeri B. Comparing related phylogenetic trees. *Lect Notes Comput Sci*. 2012;7305:31–42.

Bayer PE, Golicz AA, Scheben A, Batley J, Edwards D. Plant pan-genomes are the new reference. *Nat Plants*. 2020;6(8):914–20. doi:10.1038/s41477-020-0733-0

[47] Golicz AA, Bayer PE, Bhalla PL, Batley J, Edwards D. Pangenomics comes of age: from bacteria to plant and animal applications. *Trends Genet*. 2020;36(2):132–45. doi:10.1016/j.tig.2019.11.006

[48] Schreiber M, Stein N, Mascher M. Plant pangenomes for crop improvement, biodiversity and evolution. *Nat Rev Genet*. 2024;25:563–77. doi:10.1038/s41576-024-00691-4

[49] Schreiber M, Mascher M. What are we learning from plant pangenomes? *Annu Rev Plant Biol*. 2025. doi:10.1146/annurev-arplant-090823-015358

[50] Golicz AA, Bayer PE, Barker GC, Edger PP, Kim H, Martinez PA, et al. The pangenome of an agronomically important crop plant *Brassica oleracea*. *Nat Commun*. 2016;7:13390. doi:10.1038/ncomms13390

[51] Zhao Q, Feng Q, Lu H, Li Y, Wang A, Tian Q, et al. Pan-genome analysis highlights the extent of genomic variation in cultivated and wild rice. *Nat Genet*. 2018;50:278–84. doi:10.1038/s41588-018-0041-z

[52] 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(13):3542–58. doi:10.1016/j.cell.2021.04.046

[53] Shang L, Li X, He H, Yuan Q, Song Y, Wei Z, et al. A super pan-genomic landscape of rice. *Cell Res*. 2022;32:878–96. doi:10.1038/s41422-022-00685-z

[54] Hufford MB, Seetharam AS, Woodhouse MR, Chougule KM, Ou S, Liu J, et al. De novo assembly, annotation, and comparative analysis of 26 diverse maize genomes. *Science*. 2021;373(6555):655–62. doi:10.1126/science.abg5289

[55] Montenegro JD, Golicz AA, Bayer PE, Hurgobin B, Lee H, Chan CK, et al. The pangenome of hexaploid bread wheat. *Plant J*. 2017;90(5):1007–13. doi:10.1111/tpj.13515

[56] Walkowiak S, Gao L, Monat C, Haberer G, Kassa MT, Brinton J, et al. Multiple wheat genomes reveal global variation in modern breeding. *Nature*. 2020;588(7837):277–83. doi:10.1038/s41586-020-2961-x

[57] Li YH, Zhou G, Ma J, Jiang W, Jin LG, Zhang Z, et al. De novo assembly of soybean wild relatives for pan-genome analysis of diversity and agronomic traits. *Nat Biotechnol*. 2014;32:1045–52. doi:10.1038/nbt.2979

[58] Liu Y, Du H, Li P, Shen Y, Peng H, Liu S, et al. Pan-genome of wild and cultivated soybeans. *Cell*. 2020;182(1):162–76. doi:10.1016/j.cell.2020.05.023

[59] Gao L, Gonda I, Sun H, Ma Q, Bao K, Tieman DM, et al. The tomato pan-genome uncovers new genes and a rare allele regulating fruit flavor. *Nat Genet*. 2019;51:1044–51. doi:10.1038/s41588-019-0410-2

[60] Zhou Y, Zhang Z, Bao Z, Li H, Lyu Y, Zan Y, et al. Graph pangenome captures missing heritability and empowers tomato breeding. *Nature*. 2022;606(7914):527–34. doi:10.1038/s41586-022-04808-9

[61] Li N, He Q, Wang J, Wang B, Zhao J, Tang S, et al. Super-pangenome analyses highlight genomic diversity and structural variation across wild and cultivated tomato species. *Nat Genet*. 2023;55:852–60. doi:10.1038/s41588-023-01340-y

[62] Song JM, Guan Z, Hu J, Guo C, Yang Z, Wang S, et al. Eight high-quality genomes reveal pan-genome architecture and ecotype differentiation of *Brassica napus*. *Nat Plants*. 2020;6:34–45. doi:10.1038/s41477-019-0577-7

[63] Li X, Wang Y, Cai C, Ji J, Han F, Cheng F, et al. Large-scale gene expression alterations introduced by structural variation drive morphotype diversification in *Brassica oleracea*. *Nat Genet*. 2024;56:517–29. doi:10.1038/s41588-024-01657-2

[64] Guo N, Wang SY, Wang TY, Duan MM, Zong M, et al. A graph-based pan-genome of *Brassica oleracea* provides new insights into its domestication and morphotype diversification. *Plant Commun*. 2024;5:100791. doi:10.1016/j.xplc.2024.100791

[65] Jiao WB, Schneeberger K. Chromosome-level assemblies of multiple *Arabidopsis* genomes reveal hotspots of rearrangements with altered evolutionary dynamics. *Nat Commun*. 2020;11:989. doi:10.1038/s41467-020-14779-y

[66] Kang M, Wu H, Liu H, Liu W, Zhu M, Han Y, et al. The pan-genome and local adaptation of *Arabidopsis thaliana*. *Nat Commun*. 2023;14:6259. doi:10.1038/s41467-023-42029-4

[67] Lian Q, Huettel B, Walkemeier B, Mayjonade B, Lopez-Roques C, Gil L, et al. A pan-genome of 69 *Arabidopsis thaliana* accessions reveals a conserved genome structure throughout the global range. *Nat Plants*. 2024;10:1425–39. doi:10.1038/s41477-024-01755-3

[68] Li H, Wang S, Chai S, Yang Z, Zhang Q, Xu Y, et al. Graph-based pan-genome reveals structural and sequence variations related to agronomic traits and domestication in cucumber. *Nat Commun*. 2022;13:682. doi:10.1038/s41467-022-28362-0

[69] He Q, Tang S, Zhi H, Chen J, Zhang J, Liang H, et al. A graph-based genome and pan-genome variation of the model plant *Setaria*. *Nat Genet*. 2023;55:1232–42. doi:10.1038/s41588-023-01423-w

[70] Bozan I, Achakkagari SR, Anglin NL, Ellis D, Tai HH, Strömvik MV. Pan-genome analyses reveal impact of transposable elements and ploidy on the evolution of potato species. *Proc Natl Acad Sci USA*. 2023;120:e2211117120. doi:10.1073/pnas.2211117120

[71] Qiao Q, Edger PP, Xue L, Qiong L, Lu J, Zhang Y, et al. Evolutionary history and pan-genome dynamics of strawberry (*Fragaria* spp.). *Proc Natl Acad Sci USA*. 2021;118(45):e2105431118. doi:10.1073/pnas.2105431118

[72] Chen S, Wang P, Kong W, Chai K, Zhang S, Yu J, et al. Gene mining and genomics-assisted breeding empowered by the pangenome of tea plant *Camellia sinensis*. *Nat Plants*. 2023;9:1986–99. doi:10.1038/s41477-023-01565-z

[73] Tariq A, Meng M, Jiang X, Bolger A, Beier S, Buchmann JP, et al. In-depth exploration of the genomic diversity in tea varieties based on a newly constructed pangenome of *Camellia sinensis*. *Plant J*. 2024;119:2096–115. doi:10.1111/tpj.16790

[74] Ou L, Li D, Lv J, Chen W, Zhang Z, Li X, et al. Pan-genome of cultivated pepper (*Capsicum*) and its use in gene presence-absence variation analyses. *New Phytol*. 2018;220(2):360–3. doi:10.1111/nph.15413

[75] Jayakodi M, Schreiber M, Stein N, Mascher M, et al. Structural variation in the pangenome of wild and domesticated barley. *Nature*. 2024;636:654–62. doi:10.1038/s41586-024-08187-1

[76] Rijzaani H, Bayer PE, Rouard M, Tuytten R, Roux N, Sardos J, et al. The pangenome of banana highlights differences between genera and genomes. *Plant Genome*. 2022;15:e20100. doi:10.1002/tpg2.20100

[77] Li J, Yuan D, Wang P, Wang Q, Sun M, Liu Z, et al. Cotton pan-genome retrieves the lost sequences and genes during domestication and selection. 251*Genome Biol*. 2021;22:119. 252doi:10.1186/s13059-021-02351-w

[78] Zhang X, Liu T, Wang J, et al. Pan-genome of *Raphanus* highlights genetic variation and introgression among domesticated, wild and weedy radishes. *Molecular Plant*. 2021;14(12):2032–2055.

[79] Jin S, Han Z, Hu Y, et al. Structural variation (SV)-based pan-genome and GWAS reveal the impacts of SVs on the speciation and diversification of allotetraploid cottons. *Molecular Plant*. 2023;16(4):678–693.

[80] Aho AV, Sagiv Y, Szymanski TG, Ullman JD. Inferring a tree from lowest common ancestors with an application to the optimization of relational expressions. *SIAM J Comput*. 1981;10(3):405–421.

[81] Kimura M. Preponderance of synonymous changes as evidence for the neutral theory of molecular evolution. *Nature*. 1977;267:275–276.

[82] Kuo CH, Ochman H. Inferring clocks when lacking rocks: the variable rates of molecular evolution in bacteria. *Biol Direct*. 2009;4:35.

[83] Sankoff D. Mechanisms of genome evolution: models and inference. *Bull Int Stat Inst*. 1989;47(3):461–475.

[84] Koren S, Walenz BP, Berlin K, Miller JR, Bergman NH, Phillippy AM. Canu: scalable and accurate long-read assembly via adaptive k-mer weighting and repeat separation. *Genome Res*. 2017;27(5):722–736.

[85] Durand NC, Shamim MS, Machol I, et al. Juicer provides a one-click system for analyzing loop-resolution Hi-C experiments. *Cell Syst*. 2016;3(1):95–98.

[86] Li R, Zhu H, Ruan J, et al. De novo assembly of human genomes with massively parallel short-read sequencing. *Genome Res*. 2010;20(2):265–272.

[87] Zheng GX, Lau BT, Schnall-Levin M, et al. Haplotyping germline and cancer genomes with high-throughput linked-read sequencing. *Nat Biotechnol*. 2016;34(3):303–311.

[88] Servant N, Varoquaux N, Lajoie BR, et al. HiC-Pro: an optimized and flexible pipeline for Hi-C data processing. *Genome Biol*. 2015;16:259.

Swenson KM, Chen E, Pattengale ND, Sankoff D. The kernel of maximum agreement subtrees. *IEEE/ACM Transactions on Computational Biology and Bioinformatics*. 2012;9(4):1023–1034.

[89] Khezeli A. An improved lower Bound on the largest common subtree of random leaf-labeled binary trees. *SIAM Journal on Discrete MathematicD*. 2024;38(3): 2530–2541.