

Reconstructing the constituent genomes of the ancestral angiosperm pangenome

David Sankoff¹[0000-0001-8415-5189], Jiazhen Leng¹[0000-0002-9390-8426],
Pratheesh Soman²[0009-0000-7263-748X], Qiaoji Xu¹[0000-0003-3316-2172],
Chunfang Zheng³, Alex Liu⁴, James H. Leebens-Mack⁵[0000-0003-4811-2231],
and Lingling Jin²[0000-0002-4586-2347]

¹ Department of Mathematics and Statistics, University of Ottawa, Ottawa, ON, Canada

² Department of Computer Science, University of Saskatchewan, Saskatoon, SK, Canada

³ Agriculture and Agri-Food Canada, Ottawa, ON, Canada

⁴ School of Computer Science, University of Waterloo, Waterloo, ON, Canada

⁵ Department of Plant Biology, University of Georgia, Athens, Georgia, USA

Abstract. To reconstruct the gene orders of the constituent genomes in an ancestral pangenome, we propose an analysis of RACCROCHE maximum weight matching output of contigs based on adjacency pairs from phylogenetically disparate genomes. The key idea is to use the multiple solutions to the matching optimization as a sample of the set of constituent genomes. We identify those gene-order contigs present in all the solutions as the “core” of the pangenome, and those absent in some of the solutions as the pangenome “shell”. Different cliques of mutually compatible shell contigs identified different constituent genomes. A significant proportion of shell genes in each pangenome was inherited from the set of shell genes in its ancestor pangenome. As a hint to the chromosomal structure, we performed hierarchical clustering on the combined set of contigs based on the number of solutions shared by each pair of contigs, in the search for chromosomal fragments, large clusters present in many ancestors, and some limited but clear results were obtained.

1 Introduction

Pangenomes aim to represent all the variation found in the genomes of a set of related organisms [1] — populations, species, genera — which we call the constituent genomes. There are two main approaches to the formal study of the gene complement of pangenomes. One is identification of the “core” genes (or ortholog group) present in all constituent genomes, versus the “accessory” or “shell” genes, present in a sizable subset of the constituent genes, and the “unique” or “cloud” genes, present in a single genome. (The meanings of terms like accessory and cloud vary in the literature.) The core may contain fewer than

10% of the pangenome genes, as in the case of some bacteria [2,3], from 30 – 70% for many plants and animals [4,5,6], or over 95% for humans [7].

The second gene-centric approach to the pangenome is that of “pangenome graphs”. Here, genes (or ortholog groups) are represented as vertices. Adjacent genes in a chromosome of a constituent genome are connected by an edge, often a directed edge. The massive redundancies and conflicts inherent in the resulting raw structure are then reduced by various algorithms to acyclic or locally acyclic graphs. Many types of graph are used to represent the output of these algorithms, but most of these focus on sequences, where full analysis of the gene content is secondary or absent, e.g., de Bruijn graphs [8], cactus graphs [9]. A number of primarily gene-centric pangenome-graph algorithms and packages are, however, available [10,11,12].

One topic almost never broached in the pangenome literature is the phylogeny of pangenomes. But in the context of the phylogenetics of a number of species or genera, each represented by a pangenome, why settle for simply reducing each pangenome to a linear, or at least locally acyclic, order and then proceed with a traditional phylogenetic analysis of these linearized genomes? After all, it is not a new idea that an ancestral population may be more or less heterogeneous with respect to the genomes of individuals or groups. This is explicit in the modern recognition of incomplete lineage sorting [13], but it was understood earlier, such as in the description of species as clouds or quasispecies of more or less closely related individuals [14]. In this paper, then, following previous suggestions [15], we develop a “small” phylogenetic analysis of pangenomes, where the inferred ancestors are also pangenomes. We apply this analysis to flowering plants, with representative genomes from each of the major angiosperm clades, as in Figure 1.

Our analysis is based on an earlier gene-order inference of ancestral genomes: the RACCROCHE pipeline [16,17]. This analysis generates a non-unique optimal solution. In this paper, we capitalize on the non-uniqueness property to reconstruct the constituent genomes of the pangenome for each of the ancestors, as in Figure 2. Note that this preservation and exploitation of the non-uniqueness of RACCROCHE solutions is a *complete antithesis* of the traditional goal of reducing the ambiguity inherent in multiple solutions to arrive at a single ancestral genome.

2 Data and Methods

Source data. The original data were 16 high-quality genomes reported in [18] and depicted in Figure 1. Indicating the ancestral node of each clade with boldface numbers, we use the nuclear genomes of

- ANA grade ①
 - Amborella trichopoda* [19]
 - Water lily (*Nymphaea colorata*) [20]
 - Chloranthus sessilifolius* ② [21].
- Magnoliids ⑦
 - Cinnamon (*Cinnamomum chago*) [22]

- Tulip tree (*Liriodendron chinense*) [23,24]
Aristolochia fimbriata [25]
 Monocots ⑤
 Asparagus (*Asparagus officinalis*) [26]
 Pineapple (*Ananas comosus*) [27]
 Yam (*Dioscorea alata*) [28]
 Acorus americanus [29]
 Ranunculales ⑧
 Poppy (*Papaver bracteatum*) [30]
 Columbine (*Aquilegia coerulea*) [31] (Eudicots),
 Proteales ⑪
 Lotus (*Nelumbo nucifera*) [32]
 Buxales ⑩
 Boxwood (*Buxus austroyunnanensis*) [33]
 Core eudicots ⑨
 Grapevine (*Vitis vinifera*) (NCBI RefSeq assembly: GCF_030704535.1)
 Lindenbergia philippensis [34]

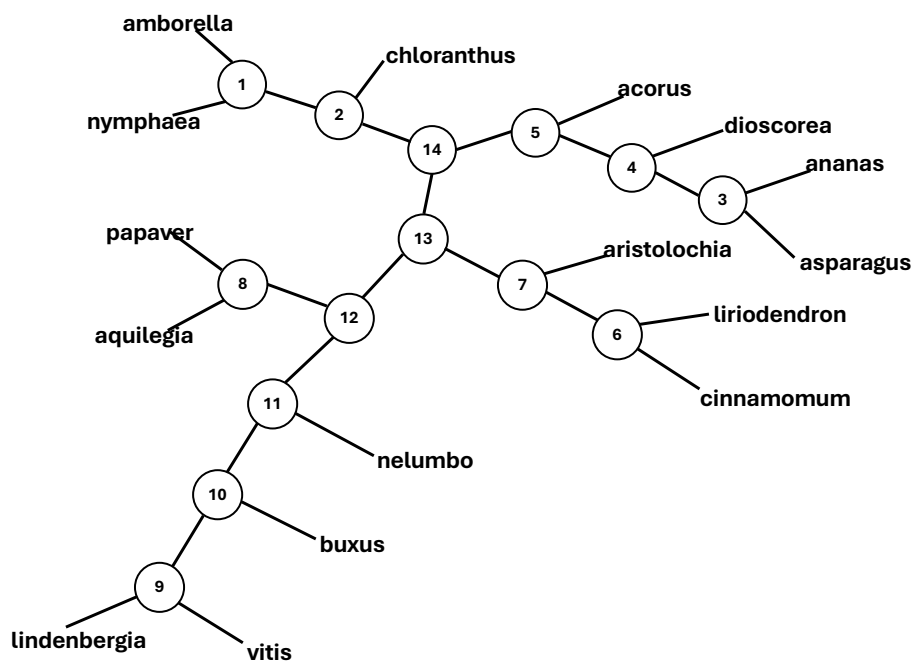


Fig. 1. Angiosperm phylogeny. The goal is to reconstruct the 14 ancestral pangenomes associated with the numbered nodes, based on the 16 extant genomes.

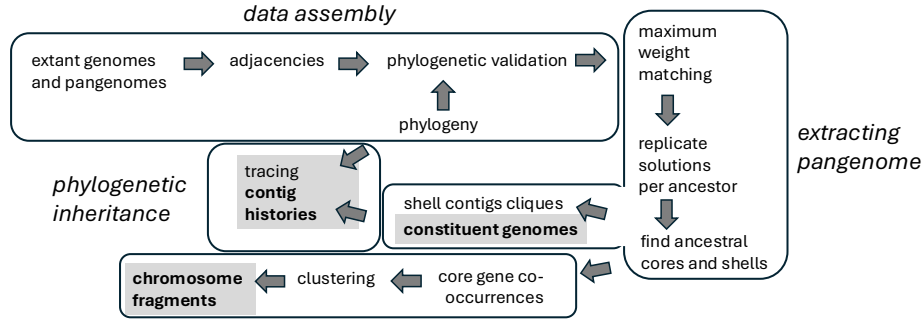


Fig. 2. Protocol for ancestral pangenome construction

The matching. The initial analysis was carried out using the RACCROCHE pipeline [16,35,36,17,18]. All adjacencies and near-adjacencies¹ identified by ortholog groups in all these genomes were assembled as an input graph to a maximum weight matching (MWM) algorithm, with specific “phylogenetic validation” restrictions as in Figure 3 pertaining to each of the various ancestors. For an adjacency X to appear at an internal vertex associated with an ancestral pangenome of a binary branching phylogenetic tree, X must appear in at least two trees subtended by the internal vertex.

For each ancestor, 101 distinct replica solutions of the MWM algorithm were generated by varying the data input order. Since the matching algorithm ensured that each gene in an adjacency was matched to at most one gene in another adjacency, the output for each replicate solution was a set of disjoint “contigs” representing linearly ordered fragments of the chromosomes of the ancestor. Combining the results from all the replicates provided the summary under “contigs” in Table 1. Most of these contigs (from 85% to 92%) for each ancestor, containing from 90% to 95% of the genes, were in all replicates.

Accessory contigs analysis. In order to examine contigs that are not part of the core (shared) genome, we constructed an overlap graph represented by a binary matrix in which each entry (i, j) is set to 0 if contigs i and j share no genes, and 1 otherwise. Under our evolutionary model, ancestral genomes are monoploid [36] and thus free of paralogy; consequently, any two contigs that share one or more genes are incompatible and cannot co-occur within the same constituent genome. In contrast, contigs with no shared genes are mutually compatible.

Algorithm 1 identifies compatible contig sets by iteratively extracting maximum subsets of pairwise compatible contigs. Specifically, it searches for a maximum clique in the complement of the overlap graph—that is, a largest set of vertices connected exclusively by 0-entries, corresponding to contigs that are mutually

¹ All pairs of genes in a window size 3 in the gene order of a chromosome. Because of frequent single-gene inversions, the orientation of genes was ignored.

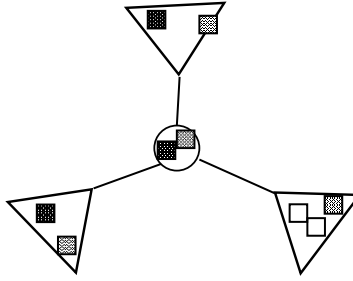


Fig. 3. Phylogenetic validation of adjacencies [15]. Necessary condition for adjacencies to appear at an internal vertex associated with an ancestral pangenome of a binary branching phylogenetic tree. Light shaded adjacency (small square) appears in all three trees (triangles) subtended by the internal vertex (circle). Dark shaded adjacency appears in only two of the trees. Unshaded adjacency appears in only one subtree so does not affect internal vertex. The shaded adjacencies are “phylogenetically validated” with respect to the internal vertex. The unshaded one is not validated.

non-overlapping in gene content. Once such a clique is identified, the associated contigs are removed from the graph, and the procedure is repeated on the remaining contigs until the graph is exhausted. This iterative extraction yields a partition of the contigs into sets that are internally compatible and can therefore represent candidate constituents of accessory genomes.

Algorithm 1 first constructs the compatibility graph on the set of contigs, in which two contigs are adjacent if they do not overlap in gene content. A clique in this graph therefore represents a set of pairwise compatible contigs. At each iteration, the algorithm identifies a maximum clique, removes its vertices from the graph, and repeats the procedure on the remaining contigs until no vertices remain. This greedy sequential extraction yields an ordered partition of the contigs into internally compatible sets. Maximum cliques were computed exactly using a branch-and-bound-type algorithm with pruning, as implemented in the NetworkX library [37], following standard approaches to the maximum clique problem [38,39,40].

With this type of algorithm, no polynomial time guarantee is available. But for the moderate size data sets under study, the time complexity is no hindrance.

Inheritance of shell genes. Because each ancestor and its descendants are reconstructed independently by the MWM procedure, we examined whether the resulting reconstructions nonetheless exhibit signatures of inheritance across successive ancestral nodes. To this end, we focused on shell genes and quantified the extent to which the shell gene set of a given ancestor is inherited from its immediate predecessor. Specifically, we calculated the proportion of an ancestor’s shell genes that are shared with the shell gene set of its parent ancestor.

Chromosomal fragments. The core of the pangenomes contains, by definition, the set of genes that are present in all constituent genomes, excluding any genes

Algorithm 1: Sequential extraction of compatible contigs

Input: A set of contigs $C = \{1, 2, \dots, n\}$ and an overlap matrix M , where

$$M_{ij} = \begin{cases} 0, & \text{if contigs } i \text{ and } j \text{ share no genes (compatible),} \\ 1, & \text{if contigs } i \text{ and } j \text{ share at least one gene (incompatible).} \end{cases}$$

Output: An ordered partition $\mathcal{P} = (K_1, K_2, \dots, K_T)$ of C into pairwise compatible contig sets.Construct an undirected graph $G = (V, E)$ with $V \leftarrow C$;**for** $i < j$ *with* $i, j \in C$ **do** **if** $M_{ij} = 0$ **then** add edge (i, j) to E ; $\mathcal{P} \leftarrow \emptyset$;**while** $V \neq \emptyset$ **do** $K \leftarrow$ a maximum clique of G ; append K to \mathcal{P} ; remove all vertices in K from G ;**return** \mathcal{P} ;

that are absent from any of them. These core genes are distributed across a large number of contigs, which are not ordered relative to one another. However, because evolutionary breakpoints are far less frequent than genes, substantial fragments of chromosomes are expected to be conserved across multiple ancestral genomes.

Accordingly, as in Algorithm 2, we quantify the co-occurrence of pairs of genes within the same contig across all 14 reconstructed ancestors. We then apply a clustering analysis to these co-occurrence patterns to identify groups of genes that are likely to have resided on the same chromosome during the evolution of the pangenomes. Complete-link clustering was performed using MATLAB's `linkage` implementation, which uses optimized nearest-neighbor chain-type methods with quadratic time complexity.

3 Results

3.1 Replicates

One hundred and one replicate runs, produced 101 different ancestral genomes for each of intermediate ancestors.

3.2 The core

Most of the 597-750 contigs for each ancestor, including all the longest contigs, containing an even greater proportion of the 3834-6073 genes, were part of every

Algorithm 2: Search for conserved chromosomal fragments

Input: Set A of ancestor pangenomes;
 for each $a \in A$: partition of gene set G into contigs

Output: Genes clustered by co-occurrence frequency in ancestral contigs

Initialize co-occurrence matrix $M \in \mathbb{R}^{|G| \times |G|}$ with zeros;

foreach ancestor $a \in A$ **do**

foreach contig c in a **do**

foreach pair of genes (g_i, g_j) in c where $i < j$ **do**

$M[g_i][g_j] \leftarrow M[g_i][g_j] + 1$;

$M[g_j][g_i] \leftarrow M[g_j][g_i] + 1$;

Convert to distance matrix: $D[i][j] \leftarrow |A| - M[i][j]$ for all i, j ;

Perform complete-linkage hierarchical clustering on D ;

foreach $k \in \{0, 1, \dots, |A| - 2\}$ **do**

 Extract clusters at tree level k ;

 // Genes in each cluster co-occur in $\geq (|A| - k)$ ancestors

return All clusters for $k \in \{0, 1, \dots, |A| - 2\}$

replicate solution, as is clear from the small numbers listed under “shell” in Table 1.

We hypothesize that genes in the great majority of these contigs form the core of the ancestral pangenome. Those counted under “shell” would form part of the accessory portion or a unique portion of the pangenome since they are present in at least one replicate, but not all.

3.3 How many constituent genomes make up a pangenome?

Do the clique sizes reveal anything about the structure of the ancestral genomes in terms of these constituent genomes? Table 2 reports the maximum clique sizes and the results obtained by successively removing previously identified cliques. Because cliques are mutually incompatible, each maximum clique must correspond to a distinct constituent genome within the pangenome. This is a minimum, however, since each clique could potentially be subdivided into many smaller sets, each of which might determine a separate constituent genome. Nevertheless, these results provide the first evidence for the presence of distinct constituent genomes in our analysis.

3.4 Inheritance of shell genes

Table 1 shows that shell genes make up approximately 5% to 10% of the gene complement in each ancestral pangenome. Importantly, Table 3 shows that these shell genes are not randomly drawn from the ancestor’s full gene set: up to 29% (average 16%) of them are inherited from the shell genes of the

Table 1. Number of contigs and genes in ancestors, partitioned by core and shell membership.

| ancestor | contigs | | | genes | | |
|----------|-----------|-----------|-------|------------|-----------|-------|
| | core | shell | total | core | shell | total |
| 2 | 557 (85%) | 97 (15%) | 654 | 4961 (90%) | 557 (10%) | 5518 |
| 1 | 593 (92%) | 54 (8%) | 647 | 4293 (95%) | 216 (5%) | 4509 |
| 14 | 549 (85%) | 98 (15%) | 647 | 5410 (91%) | 557 (9%) | 5967 |
| 5 | 627 (91%) | 65 (9%) | 692 | 4742 (95%) | 264 (5%) | 5006 |
| 4 | 633 (87%) | 93 (13%) | 726 | 4083 (92%) | 362 (8%) | 4445 |
| 3 | 638 (86%) | 101 (14%) | 739 | 3514 (92%) | 320 (8%) | 3834 |
| 13 | 518 (87%) | 79 (13%) | 597 | 5578 (92%) | 495 (8%) | 6073 |
| 7 | 594 (87%) | 89 (13%) | 683 | 5277 (93%) | 377 (7%) | 5654 |
| 6 | 624 (88%) | 86 (12%) | 710 | 5120 (93%) | 394 (7%) | 5514 |
| 12 | 570 (88%) | 81 (12%) | 651 | 5522 (92%) | 468 (8%) | 5990 |
| 8 | 593 (90%) | 68 (10%) | 661 | 4547 (94%) | 292 (6%) | 4839 |
| 11 | 577 (88%) | 75 (12%) | 652 | 5612 (94%) | 357 (6%) | 5969 |
| 10 | 618 (88%) | 85 (12%) | 703 | 5391 (93%) | 386 (7%) | 5777 |
| 9 | 688 (92%) | 62 (8%) | 750 | 4888 (95%) | 247 (5%) | 5135 |

immediately preceding ancestor, clearly exceeding the 5%-10% that would be expected from random (Table 1). This indicates the presence of a detectable evolutionary signal from ancestors to descendants. This enrichment demonstrates that a distinct evolutionary signal is encoded within the pangenomes, reflecting non-random transmission of gene content from ancestor to descendant. Recall that each ancestor pangenome is reconstructed independently using different sets of phylogenetically validated adjacencies as input to the MWM algorithm, indicating that the observed evolutionary signal is inherently embedded in these adjacency patterns and is systematically propagated along evolutionary lineages.

3.5 Chromosome structure of the core

All the constituent genomes of a reconstructed pangenome ancestor share the same core, containing the same genes organized within the same contigs (in our construction). However, there is currently no information or constraint that allows us to determine the relative order of these contigs within any constituent, and the same is true of the shell contigs.

Nevertheless, we can assume that each constituent genome of a pangenome possesses a chromosomal architecture, in which genes are partitioned and ordered along some number of chromosomes. During evolution, as a pangenome gives rise to its descendants, the chromosomes of its constituent genomes undergo various genome rearrangements. These include inversions of long or short segments, transpositions of segments to new positions along the same chromosome, translocations involving exchanges of segments between chromosomes, and segmental duplications. Despite these rearrangements, the number of breakpoints in chromosomes remains approximately two orders of magnitude smaller than the

Table 2. Sequentially identified maximum-weight cliques. For each ancestor, we show the number of cliques; then the ordered list of (clique size (contigs), number of genes).

| ancestor | number of cliques | number of (contigs, genes) |
|----------|-------------------|--|
| 1 | 3 | (32,212),(21,192),(1,9) |
| 2 | 3 | (52,552),(44,533),(1,10) |
| 3 | 4 | (58,305),(40,278),(2,12),(1,5) |
| 4 | 4 | (51,353),(37,319),(3,19),(2,7) |
| 5 | 3 | (34,260),(30,248),(1,4) |
| 6 | 4 | (45,389),(35,368),(4,50),(2,29) |
| 7 | 4 | (49,373),(38,353),(1,15),(1,15) |
| 8 | 2 | (38,285),(30,269) |
| 9 | 2 | (36,241),(26,223) |
| 10 | 9 | (40,380),(34,363),(3,31),(2,30),(2,29), (1,20),(1,20),(1,8),(1,4) |
| 11 | 4 | (36,512),(36,512),(2,66),(2,66) |
| 12 | 3 | (37,565),(36,563),(3,33) |
| 13 | 3 | (29,366),(28,362),(1,9) |
| 14 | 4 | (38,219),(38,219),(1,7),(1,3) |

Table 3. Number and proportion of genes inherited by ancestor j from ancestor i , partitioned by core and shell membership.

| lineage | | core i | | shell i | |
|--------------|--------------|-------------|--------------|-------------|--------------|
| ancestor i | ancestor j | in core j | in shell j | in core j | in shell j |
| 2 | 1 | 3635 (96%) | 169 (4%) | 398 (93%) | 32 (7%) |
| 14 | 5 | 4149 (96%) | 189 (4%) | 415 (87%) | 61 (13%) |
| 5 | 4 | 3422 (93%) | 270 (7%) | 186 (83%) | 37 (17%) |
| 4 | 3 | 3009 (93%) | 224 (7%) | 234 (77%) | 71 (23%) |
| 14 | 13 | 4749 (92%) | 423 (8%) | 497 (93%) | 40 (7%) |
| 13 | 7 | 4664 (94%) | 290 (6%) | 381 (84%) | 72 (16%) |
| 7 | 6 | 4539 (95%) | 264 (5%) | 250 (71%) | 101 (29%) |
| 13 | 12 | 4700 (94%) | 309 (6%) | 331 (73%) | 121 (27%) |
| 12 | 8 | 3993 (94%) | 255 (6%) | 364 (93%) | 27 (7%) |
| 12 | 11 | 4902 (95%) | 268 (5%) | 383 (86%) | 62 (14%) |
| 11 | 10 | 4858 (94%) | 310 (6%) | 264 (80%) | 67 (20%) |
| 10 | 9 | 4228 (96%) | 191 (4%) | 287 (88%) | 39 (12%) |

total number of genes. Consequently, we can expect some chromosomal fragments to survive intact from one pangenome ancestor to its descendants, providing continuity in the genomic structure over evolutionary time.

To detect signals of gene retention across ancestors, we examined all pairs among the 8150 genes appearing in all the ancestors. We counted how many times they co-occurred in the same contig of an ancestor, from 0 co-occurrences to 14.

The co-occurrence matrix thus constructed was subjected to a complete-link clustering analysis, from which we could extract clusters of genes that consistently co-occurred with each other across contigs. Specifically, we could extract clusters

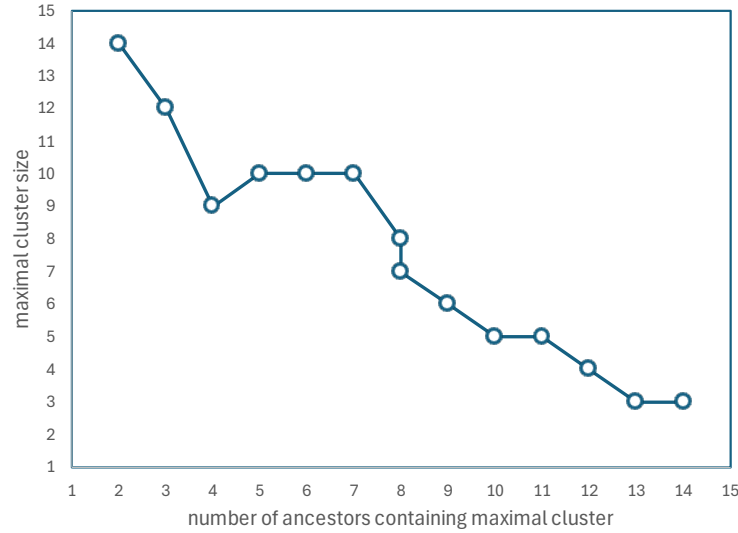


Fig. 4. Maximum number of occurrences of genes clusters of each size.

that co-occurred in just one ancestor pangenome, in two ancestor pangenomes, and so on, thereby quantifying the degree of retention of the cluster across the ancestral pangenomes.

Figure 4 shows, for each number A of ancestor pangenomes, from 2 to 14, how large is the largest cluster of genes co-occurring in A ancestors. Thus for any group of seven ancestors, the maximum cluster of genes contained ten genes. For wider co-occurrence, such as 11 ancestors, the maximum cluster size is only 5.

4 Discussion

Our analysis demonstrates the first methodology for reconstructing gene orders of ancestral pangenomes by leveraging the multiple solutions generated by RAC-CROCHE. By distinguishing core contigs (those consistently present across all solutions) from shell contigs, which are variably present, we provide a systematic method to identify conserved versus variable genomic components. Grouping mutually compatible shell contigs into distinct constituent genomes reveals the modular architecture of ancestral pangenomes, whereas hierarchical clustering of core contigs highlights chromosomal fragments and large, evolutionarily conserved gene clusters.

As the first approach to the reconstruction of ancestral gene-order pangenomes, this work is based on a number of assumptions that could be explored or relaxed in further work. Perhaps the most contentious may be the identification of the alternate solutions to the MWM optimization output as potential constituent genomes of the pangenome. This is not, however, the arbitrary interpretation of combinatorial algorithmic choices. Signals present in the complete input genomes,

reflecting differences among ancestral constituent genomes, would be carried through the adjacencies to the MWM analysis, and can be captured by the optimization process. This is supported by the observation that the proportion of inherited shell contigs and shell genes is markedly larger than that of the core, indicating a meaningful signal.

We have previously justified our working assumption of ancestral monoploidy in the context of the RACCROCHE gene-order reconstruction of unique ancestral genomes. This assumption is particularly important when partitioning shell genes among the constituent genomes of a pangenome.

The reliance on a single history of evolutionary divergence, in the context of the small phylogeny problem, could be controversial, given that the origin of the major angiosperm clades is not settled.

We have determined that there are generally at least two constituent genomes each containing approximately equal numbers of shell contigs and shell genes, with one to seven additional constituent genomes with fewer shell elements. These numbers should be considered lower bounds for each ancestor, as the data are also consistent with larger numbers of constituent genomes. Further research could refine these estimates, potentially by extending our heritability-based approach together with chromosome-level searches.

The discovery of a few sets of genes suggestive of conserved chromosomal fragments is encouraging, although it does not yet provide a detailed picture of the rate of angiosperm pangenome evolution. Several strategies could increase the sensitivity of this search: for example, using alternative clustering approaches such as average-link, k-means (our use of complete-link clustering may be overly stringent); examining whether smaller clusters are concentrated in specific clades, which could indicate genome rearrangement events at the founding of these clades; or identifying clusters of genes that are mutually exclusive in terms of co-occurrence patterns.

Together, this approach not only refines our understanding of ancestral pangenome organization but also offers a scalable methodology for studying pangenome evolution, structural variation, and lineage-specific innovations across larger phylogenies.

References

1. Chelsea A. Matthews, Nathan S. Watson-Haigh, Rachel A. Burton, and Anna E. Sheppard. A gentle introduction to pangenomics. *Briefings in Bioinformatics*, 25:bbae588, 2024.
2. E. Tantoso, B. Eisenhaber, M. Kirsch, et al. To kill or to be killed: pangenome analysis of *Escherichia coli* strains reveals a tailocin specific for pandemic ST131. *BMC Biology*, 20:146, 2022.
3. P. Qin, H. Lu, H. Du, H. Wang, W. Chen, Z. Chen, Q. He, S. Ou, H. Zhang, X. Li, and X. Li. Pan-genome analysis of 33 genetically diverse rice accessions reveals hidden genomic variations. *Cell*, 184(13):3542–3558, 2021.
4. P. Qin, H. Lu, H. Du, H. Wang, W. Chen, Z. Chen, Q. He, S. Ou, H. Zhang, X. Li, and X. Li. Pan-genome analysis of 33 genetically diverse rice accessions reveals hidden genomic variations. *Cell*, 184(13):3542–3558, 2021.

5. M. Gerdol, R. Moreira, F. Cruz, et al. Massive gene presence–absence variation shapes an open pan-genome in the mediterranean mussel. *Genome Biology*, 21:275, 2020.
6. Y. Gong, Y. Li, X. Liu, et al. A review of the pangenome: how it affects our understanding of genomic variation, selection and breeding in domestic animals. *Journal of Animal Science and Biotechnology*, 14:73, 2023.
7. Wen-Wei Liao, M. Asri, J. Ebler, et al. A draft human pangenome reference. *Nature*, 617:312–324, 2023.
8. L. Depuydt, L. Renders, T. Abeel, et al. Pan-genome de Bruijn graph using the bidirectional FM-index. *BMC Bioinformatics*, 24:400, 2023.
9. G. Hickey, J. Monlong, J. Ebler, A. M. Novak, J. M. Eizenga, Y. Gao, Human Pangenome Reference Consortium, T. Marschall, H. Li, and B. Paten. Pangenome graph construction from genome alignments with Minigraph-Cactus. *Nature Biotechnology*, 42(4):663–673, 2024.
10. A. J. Page, C. A. Cummins, M. Hunt, et al. Roary: rapid large-scale prokaryote pan genome analysis. *Bioinformatics*, 31(22):3691–3693, 2015.
11. G. Tonkin-Hill, N. MacAlasdair, C. Ruis, et al. Producing polished prokaryotic pangenomes with the Panaroo pipeline. *Genome Biology*, 21:180, 2020.
12. H. Li, M. Marin, and M. R. Farhat. Exploring gene content with pangene graphs. *Bioinformatics*, 40(7):btac456, 2024.
13. W. P. Maddison, L. Knowles, and T. Lacey. Inferring phylogeny despite incomplete lineage sorting. *Systematic Biology*, 55:21–30, 2006.
14. M. Eigen and P. Schuster. A principle of natural self-organization. *Naturwissenschaften*, 64:541–565, 1977.
15. X. Zhou and D. Sankoff. Ancestral pangenomes and their phylogenetic reconstruction. In G. Song, editor, *Comparative Genomics*, volume 15666 of *Lecture Notes in Computer Science*. Springer, 2026.
16. Q. Xu, L. Jin, C. Zheng, J. H. Leeben-Mack, and D. Sankoff. Raccroche: ancestral flowering plant chromosomes and gene orders based on generalized adjacencies and chromosomal gene co-occurrences. In *Lecture Notes in Computer Science*, volume 12686, pages 97–115. Springer, 2021.
17. Q. Xu, L. Jin, C. Zheng, X. Zhang, J. Leeben-Mack, and D. Sankoff. From comparative gene content and gene order to ancestral contigs, chromosomes and karyotypes. *Scientific Reports*, 13(1):6095, 2023.
18. P. Soman, D. Sankoff, and L. Jin. Deciphering the angiosperm phylogeny using ancestral genome reconstruction. Poster presentation, USRA Research Program, University of Saskatchewan, 2025.
19. Zhonglong Guo, Jing-Fang Guo, Zhi-Yan Wei, Ren-Gang Zhang, Scott McMahan, Shuai Nie, Xue-Mei Yan, Shan-Shan Zhou, Quan-Zheng Yun, Jia-Yi Wu, Jing Ge, Yong Yang, Jia-Yu Xue, and Jian-Feng Mao. Near-gapless telomere-to-telomere reference nuclear genome and variable mitochondrial genome of *Amborella trichopoda*. *Journal of Genetics and Genomics*, May 2025.
20. JGI. Jgi data portal. <https://data.jgi.doe.gov/phytozone?organism=Ncolorata&expanded=566>, 2020. [Accessed 22-07-2025].
21. Jianxiang Ma, Pengchuan Sun, Dandan Wang, Zhenyue Wang, Jiao Yang, Ying Li, Wenjie Mu, Renping Xu, Ying Wu, Congcong Dong, Nawal Shrestha, Jianquan Liu, and Yongzhi Yang. The *Chloranthus sessilifolius* genome provides insight into early diversification of angiosperms. *Nature Communications*, 12(1), November 2021.
22. Lidao Tao, Shiwei Guo, Zizhu Xiong, Rengang Zhang, and Weibang Sun. Chromosome-level genome assembly of the threatened resource plant *Cinnamomum chago*. *Scientific Data*, 11(1), May 2024.

23. Jinhui Chen, Zhaodong Hao, Xuanmin Guang, Chenxi Zhao, Pengkai Wang, Liangjiao Xue, Qihui Zhu, Linfeng Yang, Yu Sheng, Yanwei Zhou, Haibin Xu, Hongqing Xie, Xiaofei Long, Jin Zhang, Zhangrong Wang, Mingming Shi, Ye Lu, Siqin Liu, Lanhua Guan, Qianhua Zhu, Liming Yang, Song Ge, Tielong Cheng, Thomas Laux, Qiang Gao, Ye Peng, Na Liu, Sihai Yang, and Jisen Shi. *Liriodendron* genome sheds light on angiosperm phylogeny and species-pair differentiation. *Nature Plants*, 5(1):18–25, December 2018.
24. Hainan Wu, Ziyuan Hao, Zhonghua Tu, Yaxian Zong, Lichun Yang, Chunfa Tong, and Huogen Li. Re-annotation of the *Liriodendron chinense* genome identifies novel genes and improves genome annotation quality. *Tree Genetics & Genomes*, 19(4), June 2023.
25. Liuyu Qin, Yiheng Hu, Jinpeng Wang, Xiaoliang Wang, Ran Zhao, Hongyan Shan, Kumpeng Li, Peng Xu, Hanying Wu, Xueqing Yan, Lumei Liu, Xin Yi, Stefan Wanke, John E. Bowers, James H. Leebens-Mack, Claude W. dePamphilis, Pamela S. Soltis, Douglas E. Soltis, Hongzhi Kong, and Yuannian Jiao. Insights into angiosperm evolution, floral development and chemical biosynthesis from the *Aristolochia fimbriata* genome. *Nature Plants*, 7(9):1239–1253, September 2021.
26. PRJNA317340. Asparagus (ID 317340) - BioProject - NCBI — ncbi.nlm.nih.gov. <https://www.ncbi.nlm.nih.gov/bioproject/PRJNA317340/>, 2017. [Accessed 22-07-2025].
27. Xuzixin Zhou, Yanbin Xue, Meiqin Mao, Yehua He, Mark Owusu Adjei, Wei Yang, Hao Hu, Jiawen Liu, Lijun Feng, Huiling Zhang, Jiaheng Luo, Xi Li, Lingxia Sun, Zhuo Huang, and Jun Ma. Metabolome and transcriptome profiling reveals anthocyanin contents and anthocyanin-related genes of chimeric leaves in *Ananas comosus* var. bracteatus. *BMC Genomics*, 22(1), May 2021.
28. Yan-Mei Zhang, Zhi-Yan Wei, Cheng-Ao Yang, Xing-Yu Feng, Yue Wang, Sai-Xi Li, Xiao-Qin Sun, Zhu-Qing Shao, and Jia-Yu Xue. A telomere-to-telomere genome assembly for greater yam (*Dioscorea alata*). *Plant Communications*, 6(7):101326, July 2025.
29. JGI. Jgi data portal. <https://data.jgi.doe.gov/refine-download/phytozome?q=Acorus+americanus>, 2020. [Accessed 22-07-2025].
30. NGDC. Browse - BioProject - CNCB-NGDC — ngdc.cncb.ac.cn. <https://ngdc.cncb.ac.cn/bioproject/browse/PRJCA018701>, 2024. [Accessed 22-07-2025].
31. Danièle L Filiault, Evangeline S Ballerini, Terezie Mandáková, Gökçe Aköz, Nathan J Derieg, Jeremy Schmutz, Jerry Jenkins, Jane Grimwood, Shengqiang Shu, Richard D Hayes, Uffe Hellsten, Kerrie Barry, Juying Yan, Sirma Mihaltcheva, Miroslava Karafiátová, Viktoria Nizhynska, Elena M Kramer, Martin A Lysak, Scott A Hodges, and Magnus Nordborg. The *Aquilegia* genome provides insight into adaptive radiation and reveals an extraordinarily polymorphic chromosome with a unique history. *eLife*, 7, October 2018.
32. TNGR. TNGR-Download — nelumbodb.cn. <http://www.nelumbodb.cn/Download-Sequence>, 2022. [Accessed 22-07-2025].
33. Zhenyue Wang, Ying Li, Pengchuan Sun, Mingjia Zhu, Dandan Wang, Zhiqiang Lu, Hongyin Hu, Renping Xu, Jin Zhang, Jianxiang Ma, Jianquan Liu, and Yongzhi Yang. A high-quality *Buxus austro-yunnanensis* (Buxales) genome provides new insights into karyotype evolution in early eudicots. *BMC Biology*, 20(1), October 2022.
34. JGI. JGI Data Portal — data.jgi.doe.gov. <https://data.jgi.doe.gov/search?q=Lindenbergia+philippensis&expanded=Phytozome-689>, 2021. [Accessed 22-07-2025].

35. Q. Xu, L. Jin, J. H. Leeben-Mack, and D. Sankoff. Validation of automated chromosome recovery in the reconstruction of ancestral gene order. *Algorithms*, 14(6):160, 2021.
36. Q. Xu, X. Zhang, Y. Zhang, C. Zheng, J. H. Leeben-Mack, L. Jin, and D. Sankoff. The monoploid chromosome complement of reconstructed ancestral genomes in a phylogeny. *Journal of Bioinformatics and Computational Biology*, 19(6), 2021.
37. NetworkX Developers. Networkx: Network analysis in python. <https://networkx.org>, 2023. Accessed: 2026-03-27.
38. Richard Carraghan and Panos M. Pardalos. An exact algorithm for the maximum clique problem. *Operations Research Letters*, 9(6):375–382, 1990.
39. Patric R. J. Östergård. A fast algorithm for the maximum clique problem. *Discrete Applied Mathematics*, 120(1–3):197–207, 2002.
40. Etsuji Tomita, Akira Tanaka, and Haruhisa Takahashi. The worst-case time complexity for generating all maximal cliques. *Theoretical Computer Science*, 363(1):28–42, 2006.