Reconstructing the constituent genomes of the ancestral angiosperm pangenome

David Sankoff[1][0000-1111-2222-3333], Jiazhen Leng[1][1111-2222-3333-4444], Pratheesh Soman[2][1111-2222-3333-4444], Qiaoji Xu[1], Chunfang Zheng[1], Alex Liu[1], and Lingling Jin[2][2222–3333-4444-5555]

[1] [1] University of Ottawa
[2] [2] University of Saskatchewan

Abstract. To reconstruct the gene orders of the constituent genomes in an ancestral pangenome, we propose an analysis of RACCROCHE maximum weight match output based on adjacency pairs from phylogenetically related genomes. The key idea is to use the multiple solutions to the matching optimization as a sample, or resample, of the constituent genomes. This preservation and exploitation of the non-uniqueness of the matching solutions is complete reversal of the traditional goal of reducing by various means the the ambiguity inherent in multiple solutions to arrive at a single ancestral genome. 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 genome. The next task was to decompose the core into chromosomes. We performed hierarchical clustering on the combined set of contigs based on the number of solutions shared by each pair of contigs, and used a cutoff to decompose the entire set. We compared average-link and complete-link methods. We report dendrograms for each ancestor-method pair and a cophenetic correlation analysis; the latter is plotted against cluster cut size K to emphasize its invariance to cutting.

1 Introduction

Pangenomes aim to represent all the variation found in a the genomes of a set of related organisms — populations, species, genera — which we call the constituent genomes. There are two main approaches to the formal study of the gene complement of pangenomes. One is identification of the "core" genes (or ortholog group) present in all the constituent genomes, versus the "accessory" or "shell" genes, present in a sizable subset of the constituent genes and the "unique" or "cloud" genes, present in a single genome. (The meanings of terms like accessible, cloud and dispensable vary in the literature.) The core may contain fewer than 10% of the pangenome genes, as in the case of some bacteria [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 in 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].

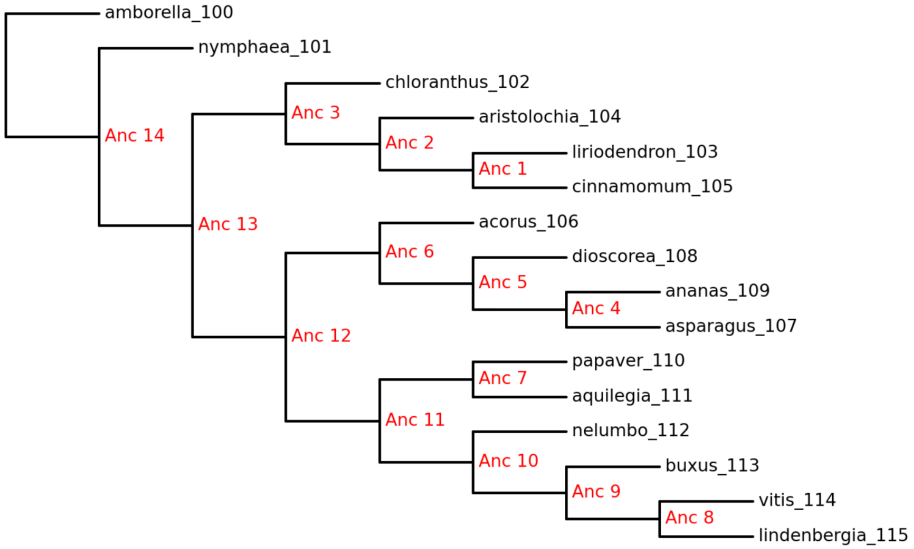

Fig. 1. Angiosperm phylogeny

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 the flowering plants, with representative genomes from each of the major angiosperm clades as in Figure 1.

Our analysis is based on a previous 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.

2 Data and Methods

Source data The original data were 16 high-quality genomes reported in [18] and depicted in Figure 1. references?

The matching The initial analysis was carried out by the RACCROCHE pipeline [16,19,20,17,18]. All the adjacencies and near-adjacencies 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 2 pertaining to each of the various ancestors. For each

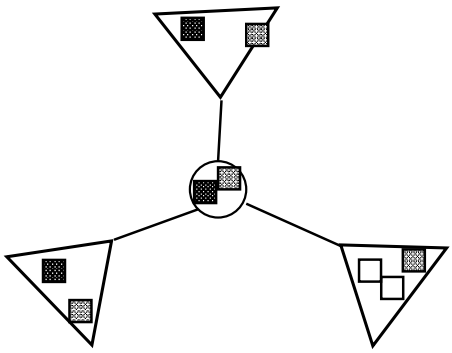

Fig. 2. Phylogenetic validation of adjacencies. Necessary condition for adjacencies to appear at an internal vertex associated with an ancestral pangenome of a binary branching phylogenetic tree. Light shaded adjacency (small square) appears in all three trees (triangles) subtended by the internal vertex (circle). Dark shaded adjacency appears in only two of the trees. Unshaded adjacency appears in only one subtree so does not affect internal vertex. The shaded adjacencies are "phylogenetically validated" with respect to the internal vertex. The unshaded one is not validated. Adapted from [15]

ancestor, 101 distinct replicate solutions of the MWM algorithm were generated, by varying the data input order. The output for each replicate 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 96.5% to 98.5% check) for each were in almost all of the replicates.

Accessory contigs analysis. We constructed an overlap graph matrix with 0's for each pair (i, j) if the two contigs shared no genes, and 1's otherwise. Given that ancestors in our evolutionary model are monoploid [20], excluding paralogy, contigs sharing genes are incompatible, meaning they cannot not appear in a single constituent genome, Looking for sets of compatible contigs, we calculated the maximal clique of 0's, removed them, calculated the the maximal clique of 0's on the remaining contigs,and so on until the graph was empty.

Inheritance of shell genes. Given that that each ancestor and its descendants are calculated independently by the MWM procedure, we ask whether there is a signal in the data indicating a degree of inheritance from one ancestor to another. We focused on the shell genes and calculated what proportion of an ancestor's shell genes own ancestors's shell genes.

3 Results

Replicate matchings One hundred and one replicate runs, produced 101 different ancestral genomes for each of intermediate ancestors. But most of the 6533-8194 contigs for each ancestor, including all the longest contigs, were part of every replicate solution, as is clear from the small numbers uner "shell" in Table 1.

We hypothesize that the genes in the great majority of these ~ 7500 contigs form the core of the ancestral pangenome. Those counted under "shell" would form part of the accessory portion or unique portion of the pangenome since they are present in at least one replicate.

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

	contigs			genes		
ancestor	core	shell	total	core	shell	total
14	8027	78	8105	11935	219	12154
13	6475	58	6533	11788	366	12154
3	6584	89	6673	11679	475	12154
2	6887	112	6999	11736	418	12154
1	7054	104	7158	11726	428	12154
12	6500	76	6576	11589	565	12154
6	7567	80	7647	11887	267	12154
5	8054	140	8194	11762	392	12154
4	8632	151	8783	11828	326	12154
11	6560	76	6636	11642	512	12154
7	7708	92	7800	11841	313	12154
10	6579	94	6673	11774	380	12154
9	6826	95	6921	11747	407	12154
8	7471	84	7555	11905	249	12154

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	4	(50,428),(49,427),(3,50),(2,29)
2	4	(54,418),(54,418),(2,16),(2,16)
3	4	(42,475),(42,475),(3,55),(2,18)
4	4	(74,326),(74,325),(2,10),(1,5)
5	4	(67,392),(67,391),(4,17),(2,7)
6	4	(39,267),(39,267),(1,4),(1,2)
7	2	(46,313),(46,313)
8	2	(42,249),(42,249)
9	9	(43,407),(43,407),(3,33),(1,29),(1,26) (1,25),(1,24),(1,21),(1,8)
10	2	(47,380),(47,380)
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 2 gives the maximal clique size and the results of successively removing previous cliques.

Do the clique sizes reveal anything about the structure of the ancestral genomes in terms of the constituent genomes? Each clique is incompatible with the others so there must be a distinct constituent genome in the pangenome for each clique. This is a minimum of course - each clique may be broken up in a number of pieces, each of which might determine a distinct constituent genome. Nevertheless, these results are the first evidence of distinct constituent genomes in our analysis.

Table 3 shows that whereas shell genes make up from 2% to 5% of the gene complement of an ancestor, these are not randomly selected from the genes of its ancestor; around 20% of them are inherited from the shell genes of its ancestor. This is much more than the 2-5% expected from random. This means that there is some evolutionary signal from an ancestor to its descendants. The ancestors are constructed independently of each other, i.e., based on different sets of adjacencies output by MWM, so that the evolutionary signal resides already in these adjacencies and is transmitted through evotionary lineages.

3.1 Next: chromosome structure of core

Table 3. Number 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
14	13	11573	362	215	4
13	3	11394	394	285	81
3	2	11375	304	361	114
2	1	11424	312	302	116
13	12	11297	491	292	74
12	6	11387	202	500	65
6	5	11535	352	227	40
5	4	11520	242	308	84
12	11	11139	450	503	62
11	7	11374	268	467	45
11	10	11341	301	433	79
10	9	11436	338	311	69
9	8	11538	209	367	40

References

1. Chelsea A Matthews, Nathan S Watson-Haigh, Rachel A Burton, Anna E Sheppard, A gentle introduction to pangenomics, Briefings in Bioinformatics, 25, 2024, bbae588,
2. Tantoso, E., Eisenhaber, B., Kirsch, M. et al. To kill or to be killed: pangenome analysis of Escherichia coli strains reveals a tailocin specific for pandemic ST131. BMC Biol 20, 146 (2022). https://doi.org/10.1186/s12915-022-01347-7
3. Qin P, Lu H, Du H, Wang H, Chen W, Chen Z, He Q, Ou S, Zhang H, Li X, Li X. Pan-genome analysis of 33 genetically diverse rice accessions reveals hidden genomic variations. Cell. 2021 Jun 24;184(13):3542-58.
4. Qin P, Lu H, Du H, Wang H, Chen W, Chen Z, He Q, Ou S, Zhang H, Li X, Li X. Pan-genome analysis of 33 genetically diverse rice accessions reveals hidden genomic variations. Cell. 2021;184(13):3542-58.
5. Gerdol, M., Moreira, R., Cruz, F. et al. Massive gene presence absence variation shapes an open pan-genome in the Mediterranean mussel. Genome Biol 21, 275 (2020).
6. Gong, Y., Li, Y., Liu, X. et al. A review of the pangenome: how it affects our understanding of genomic variation, selection and breeding in domestic animals?. J Animal Sci Biotechnol 14, 73 (2023).
7. Liao, WW., Asri, M., Ebler, J. et al. A draft human pangenome reference. Nature 617, 312–324 (2023).
8. Depuydt, L., Renders, L., Abeel, T. et al. Pan-genome de Bruijn graph using the bidirectional FM-index. BMC Bioinformatics 24, 400 (2023).
9. Hickey G, Monlong J, Ebler J, Novak AM, Eizenga JM, Gao Y, Human Pangenome Reference Consortium, Marschall T, Li H, Paten B (2024). Pangenome graph construction from genome alignments with Minigraph-Cactus". Nat Biotechnol. 42 (4): 663–673.
10. Page AJ, Cummins CA, Hunt M, Wong VK, Reuter S, Holden MT, Fookes M, Falush D, Keane JA, Parkhill J. Roary: rapid large-scale prokaryote pan genome analysis. Bioinformatics. 2015;31(22):3691-3.

11. Tonkin-Hill, G., MacAlasdair, N., Ruis, C. et al. Producing polished prokaryotic pangenomes with the Panaroo pipeline. Genome Biol 21, 180 (2020).
12. Li H, Marin M, Farhat MR. Exploring gene content with pangene graphs. Bioinformatics. 2024;40(7):btae456.
13. Maddison WP, Knowles L, Lacey T (2006) Inferring phylogeny despite incomplete lineage sorting. *Systematic Biology*, **55**: 21–30.
14. Eigen M, Schuster P (1977) A principle of natural self-organization. *Naturwissenschaften*, **64**: 541–565.
15. Zhou, X., Sankoff, D. (2026). Ancestral Pangenomes and Their Phylogenetic Reconstruction. In: Song, G. (eds) Comparative Genomics. RECOMB-CG 2025. Lecture Notes in Computer Science, vol 15666.
16. Q. Xu, L. Jin, C. Zheng, J.H. Leeben-Mack, D. Sankoff (2021) RACCROCHE: Ancestral flowering plant chromosomes and gene orders based on generalized adjacencies and chromosomal gene co-occurrences. Lecture Notes in Computer Science 12686, 97-115.
17. Q.Xu, L.Jin, C.Zheng, X.Zhang, J.Leebens-Mack, D.Sankoff (2023)From comparative gene content and gene order to ancestral contigs, chromosomes and karyotypes. Scientific Reports 13 (1), 6095.
18. Soman P, Leebens-Mack JH, Sankoff D, Jin L. 2025. Deciphering the angiosperm phylogeny using ancestral genome reconstruction. USRA Report, University of Saskatchewan.
19. Q. Xu, L. Jin, J.H. Leeben-Mack, D. Sankoff (2021) Validation of automated chromosome recovery in the reconstruction of ancestral gene order. Algorithms 14, 6: 160.
20. Q. Xu, X. Zhang, Y. Zhang, C. Zheng, J.H. Leeben-Mack, L. Jin, D. Sankoff (2021) The monoploid chromosome complement of reconstructed ancestral genomes in a phylogeny. Journal of Bioinformatics and Computational Biology 19, 6.
21. Hubert, L., Arabie, P. (1985) Comparing partitions. Journal of Classification, 2:193–218.