The design, implementation and application of a computational pipeline for the reconstruction of the gene order on the chromosomes of very ancient ancestral species

Qiaoji Xu

Thesis submitted in partial fulfillment of the requirements for the degree of
Doctorate in Philosophy Mathematics and Statistics[1]

Department of Mathematics and Statistics
Faculty of Science
University of Ottawa

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

Abstract

This thesis presents a novel approach to reconstructing ancestral genomes of a number of descendant species related by a phylogeny. Traditional methods face challenges due to cycles of whole genome doubling followed by fractionation in plant lineages. In response, the thesis proposes a new approach that first accumulates a large number of candidate gene adjacencies specific to each ancestor in a phylogeny. A subset of these which to produces long ancestral contigs are chosen through maximum weight matching. The strategy results in more complete reconstructions than existing methods, and a number of quality measures are deployed to assess the results.

The thesis also presents a new computational technique for estimating the ancestral monoploid number of chromosomes, involving a "g-mer" analysis to resolve a bias due to long contigs and gap statistics to estimate the number. The method is applied to a set of phylogenetically related descendant species, and the monoploid number is found to be 9 for all rosid and asterid orders. Additionally, the thesis demonstrates that this result is not an artifact of the method, by deriving a monoploid number of approximately 20 for the metazoan ancestor.

The reconstructed ancestral genomes are functionally annotated and visualized through painting ancestral projections on descendant genomes and highlighting syntenic ancestor-descendant relationships. The proposed method is applied to genomes drawn from a broad range of plant orders. The Raccroche pipeline reconstructs ancestral gene orders and chromosomal contents of the ancestral genomes at all internal vertices of a phylogenetic tree, and constructs chromosomes by counting the frequencies of ancestral contig co-occurrence on the extant genomes, clustering these for each ancestor, and ordering them.

Overall, this thesis presents a significant contribution to the field of ancestral genome reconstruction, offering a new approach that produces more complete reconstructions and provides valuable insights into the evolutionary process giving rise to the gene content and order of extant genomes.

Acknowledgement

I would like to take this opportunity to express my heartfelt gratitude to all those who have supported me throughout my academic journey. First and foremost, I would like to extend my sincere appreciation to my supervisor, Dr. David Sankoff, for providing me with the opportunity to work in his laboratory. It has been an honor and a privilege to have learned and worked under the guidance of such an exceptional scientist, whose vast knowledge and unwavering dedication to his work has been a source of constant inspiration. I am deeply grateful for his patience, professionalism, and unconditional support, which has been instrumental in shaping my academic career. I am grateful for the invaluable support I received from my labmates, including Zhe Yu, Yue Zhang, Eric Lam, Xiomeng Zhang, Daniella Santos Monoz, Xinyi Huang, Mona Meghdari, Andrea Michaela Parvan, and the late Alma Oladi.

I would also like to express my gratitude to Dr. Lingling Jin, Dr. Chungfeng Zheng, and Dr. James Leeben-Mack for their invaluable contributions, encouragement, and thought-provoking discussions during my graduate study. Their continual support and mentorship have played a vital role in my academic growth and development.

I am also indebted to the advisors and staff in the Math department, particularly Gilles Lamothe, Benoit Dionne, Mayada El Maalouf, Felipe Chong and Martine Bertrand-Bourgeois, for their kindness, camaraderie, and unwavering support. I am grateful to the welcoming staff of the Math department for their assistance and guidance throughout my academic journey.

Finally, I would like to express my deepest appreciation to my parents, sister, and my partner, Haoyang Liu, for their unwavering love, support, and encouragement. They have been the source of my everlasting drive for success, and without their support, I could not have made it this far.

To all those who have helped me along the way, I offer my heartfelt thanks. Your support and encouragement have been invaluable, and I am grateful for the role you have played in my academic journey. Thank you, everyone.

Contents

List of Figures

Chapter 1

Introduction

1.1 A novel approach to ancestral genome reconstruction

The subject of this thesis is the design, implementation and application of a novel computational pipeline for the reconstruction of the gene order on the chromosomes of very ancient ancestral species.

As data, the pipeline uses only the chromosomal gene orders in present-day (extant) descendants of these ancestors, as well as the historical relationships among these descendants in the form of a known evolutionary family tree: a "phylogeny".

This work adds to the analyses and programs in the field of comparative genomics, although this latter field is largely based on DNA sequence analysis, whereas in my work sequence analysis plays only an incidental role, the main data being gene order at the genomic level. Thus in this thesis, a genome is a set of chromosomes, up to several dozen in number, and each chromosome is a sequence of labelled elements called genes. The same gene may exist in several somewhat different versions in a genome or in different genomes. All these versions are homologous, meaning they are very similar, reflect a common origin in evolution, and are clearly distinct from other genes, so that in our work we use single label for all homologs and consider them identical for the purposes of our analyses.

We will review existing ancestral genome reconstruction methods in subsequent paragraphs, but first I will list some claims for novelty in the biological, mathematical and statistical aspects of my work. These will also be explained later in this introduction.

- We reconstruct monoploid ancestors, namely genomes that contain at most one copy of any gene. This requires justification since no other method incorporates this restriction.

- This allows us to use a Maximum Weight Matching (MWM) algorithm, which is the only method that directly guarantees chromosomes with no branching structure, a highly desirable property of reconstructions.
- The output of MWM being chromosome fragments, not complete chromosomes, we introduce a matrix of chromosomal co-occurrence of fragments, which can be clustered statistically to produce whole chromosomes.

1.2 Parsimony problems, large and small

The archetypical task of computational comparative genomics is to develop and apply methodology for constructing phylogenetic trees (or"phylogenies") for groups of species whose evolutionary history is of interest. Formally, in the most typical cases these are rooted binary branching trees, with N given species (observed or extant) associated with the N terminal vertices (or leaves), where the non-terminal vertices represent ancient speciation events. These event are what give rise to the similarities and differences among the N extant species.

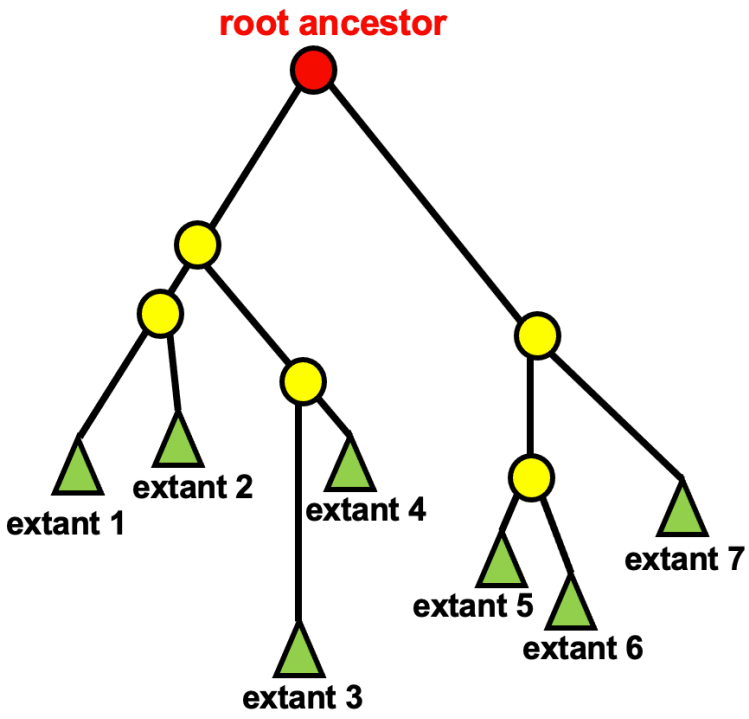


Figure 1.1: Phylogeny in terms of rooted binary branching tree.

Mathematically, these trees are acyclic connected graphs where every vertex has degree one (the terminals) or degree three (the non-terminals), except for one vertex of degree two (the root). The biological interpretation is that time, or evolutionary differentiation flows from the root, at some historical time, along the edges of the tree, until the present day extant species associated with terminals (cf. Figure 1.1). We can also use unrooted trees.

The data on which phylogenies are based may be of many types, from the anatomical morphology of traditional systematics, through the amino acid sequence of a protein, the DNA sequence of a gene up to the ordering of genes on chromosomes. The methods for constructing trees aim to put more similar species, under any of these criteria, close together in the tree. This is motivated by the hypothesis that as the most recent speciation event shared by two species recedes into the past, the degree of similarity inevitably decays, as independent (and generally different) mutations accumulate in the two of them.

In the mid-1960s, this enterprise, under the rubric “numerical taxonomy", was exemplified by methods of hierarchical classification such as single link, average link (sometimes called UPGMA), and compete link, and was surveyed in the first textbook in the field by R. Sokal and P. Sneath (1963) [2].

Shortly thereafter, a very different approach was championed by J.S.Farris, leading to a virulent controversy over what he called the “phenetics" of Sokal and Sneath versus his own parsimony-based “cladistics" [3].

The data for both approaches was generally an $N \times p$ array of species, one row per species, and some characteristic or component, like morphological size, shape or coloration, or in more recent times the p amino acids of a protein or the q bases of a gene [4].

For our purposes, the main difference between the two approaches is that the clustering techniques are generally applied to a single matrix of similarities or distance summarizing all the input component measures of resemblance or difference between pairs of species. In contrast the parsimony methods also analyze each of these components separately. The clustering methods simply seek to estimate the graph theoretical tree, the “topology" of the phylogeny, while the parsimony methods also seek to assign some biological meaning to each speciation vertex in the tree, which we may loosely call “reconstruction". Finding the topology, common to both approaches, is called the “large phylogeny problem", and it is computationally hard. Reconstructing information about hypothetical ancestral species associated with the non-terminal nodes, particular to parsimony, is called the “small phylogeny problem" and is generally more tractable. In modern phylogenetics, powerful Bayesian and other computational methods based on efficent algorithms can solve the maximum likelihood version of both problems for genome-level sequence data from large numbers of species [5] [6].

1.3 Genomes, chromosomes and genes

In this thesis, we depart from traditional phylogeny in the kind of data we consider, the kind of reconstruction we seek, and the methods we use.

The data are no longer $N \times p$ arrays with the entries in each column reflecting some similarity or difference among the N species. Instead each of the N species is represented by a formal object called a genome, which is an abstract set of containing a (variable) number of elements called chromosomes, and where each chromosome consists of a (variable) number of labelled items called genes in a linear order.

This formal structure is all that we need for our main algorithms, but the terms “genomes", “chromosomes" and “genes" are biologically motivated and are abstractions from real biological data. Thus a genome is the complete set of genetic information contained in the DNA of each organism of a species. It provides all of the information the organism requires to function and is the locus of faithful inheritance from one generation to the next. The genome is partitioned among the chromosomes, which are made of protein and a single molecule of DNA. Some sections of the DNA, called genes, which are linear arrays of hundreds or thousands of copies of the chemical nucleotides (or bases) adenine, cytosine, guanine and thymine. The bases in each gene are arranged to spell out the code for some RNA and protein molecules required by some one or other cell type responsible for the organism's structure and function.

The data for our work is extracted from various genome sequence databases, [7] which organize the sequences in terms of the chromosome or chromosomal fragment, as well as the location of previously known or novel genes, or parts of genes, or potential genes, as well as other types of genomic elements.

The same gene may exist in several somewhat different versions in a genome or in different genomes, all “homologs", but each with a different scientific name assigned by biologists who study them. An important part of our work is the conversion of this biologically meaningful data to the type of gene labels used in our work. We profit from original algorithms developed in our lab by Chunfang Zheng [8] to partition all the genes in a set of genomes into disjoint sets of homologs, and to assign each set a unique label.

Examining the series of genes in several genomes, it is striking that they contain many of the same genes, but are ordered differently along the length of the chromosome. Moreover the genes on a single chromosome in one genome may be distributed among several chromosomes in another genome. This is due to genome rearrangement mutations, which are very different from the single nucleotide changes familiar from the evolution or pathological mutation of individual genes.

This brings us back to the main goal of this thesis, the reconstruction of ancestral genomes, a version of the small phylogeny problem. The problem is given the chromosomal structure of N genomes related by a known phylogenetic tree, as well as the gene content and gene order on these chromosomes, how can we best construct

the chromosomal structure, including gene orders, of the ancestral genomes?

1.4 Whole Genome Duplication, plant orders and a focus on monoploidy

In this thesis and in comparative genomics in general, we put aside consideration of the fact that within a species, there is a small amount of variation in gene content and order, something that is studied under the term pangenomics, and that such variation may exist between the maternal and paternal versions of a chromosome within a single individual, so that we only consider a unique version of each chromosome. We may use the genetics term “haploid" to distinguish such a representation from the normal “diploid" genome, which contains pairs of chromosomes, one from each parent.

Some species, however, exist as “polyploids", where there are two, three, four or more identical copies of each chromosome, called tetraploids, hexaploids, octoploids, etc. This is especially prevalent in the plant kingdom, involving for example, many crop plants. Differences in the genetic processes of reproduction may be a feature of a polyploid genome compared to a diploid. We say that the species has undergone “whole genome duplication (WGD)" or whole genome triplication, etc. (Figure 1.2)

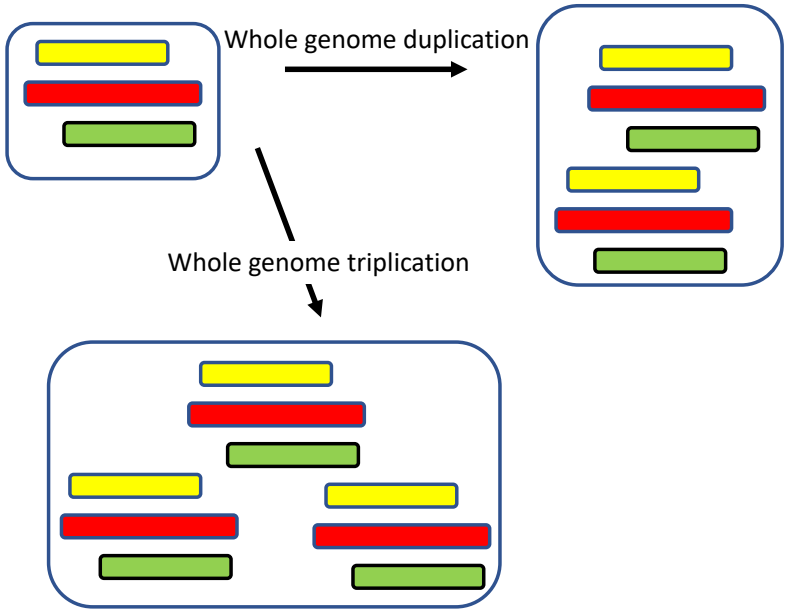


Figure 1.2: Whole genome duplication and triplication of a genome containing three chromosomes

In studying the evolution of a species, an ancient polyploidization event can often be deduced from the presence of many duplicate genes or duplicate chromosomal fragments on various chromosomes. There will, however, no longer be pairs or triples of identical chromosomes, thanks to a long history of rearrangement processes; many if not most genes will be single-copy; while genetic and other features of polyploids will not longer operate; a “re-diploidization" process will have taken place.

The widespread occurrence of ancient WGD in plants, together with extensive chromosomal rearrangement, leads to extensive ambiguity in reconstructing ancestral genomes. When a different copy of the same gene can be adjacent to four or five other genes in a genome, and to other genes in other genomes, it is inevitable that an incorrect ancestral adjacency will often be inferred.

Our solution to this quandary lies in an important observations about plant genomes, as well as a high-risk hypothesis, a gamble really, about our method of reconstruction. The observation is that most plant WGD occurred in the last 40-60 million years approximately, a period when most plant families were founded, and when, more recently, the genera and species we recognize today emerged (cf. Figure 1.3). Before that, 70-100 million years ago, the ancestors at the origin of plant orders, before they each diversified into several families, were generally unaffected by WGD. This was already apparent to Grant (1963), who discovered that the basic chromosome number of plant orders was quite small, around 7-10, and that larger numbers, due to WGD, emerged more recently, with the founding of gene families [9].

Genomes of species unaffected by WGD, although they may not be completely devoid of duplicates, will tend to have fewer of them and, most importantly, will have no duplicate chromosomal fragments containing many contiguous pairs of genes. For our purposes, we define a “monoploid" genome as one which contains at most one copy of each of its genes, and we postulate monoploidy as a good model for the genome of the founding species of each of the plant ordera we study. The term “monoploid" is sometimes confused with “haploid", but from our evolutionary perspective they are quite different. Thus the 21 chromosomes of wheat are derived from a polyploidization combining three 7-chromosome descendants of the same 7-chromosome founder species. Thus the monoploid number of chromosomes is 7, but in meiosis all 21 pairs of chromosomes in the diploid species act individually and so the number of haploid chromosomes is 21.

1.5 Gene Adjacencies and Maximum Weight Matching

The smallest building block of the gene order on a chromosome is an “adjacency", what genes are next to another gene as in Figure 1.4, preceding or following, in the ordering. In a linear order each gene can be adjacent to at most two other genes. It follows

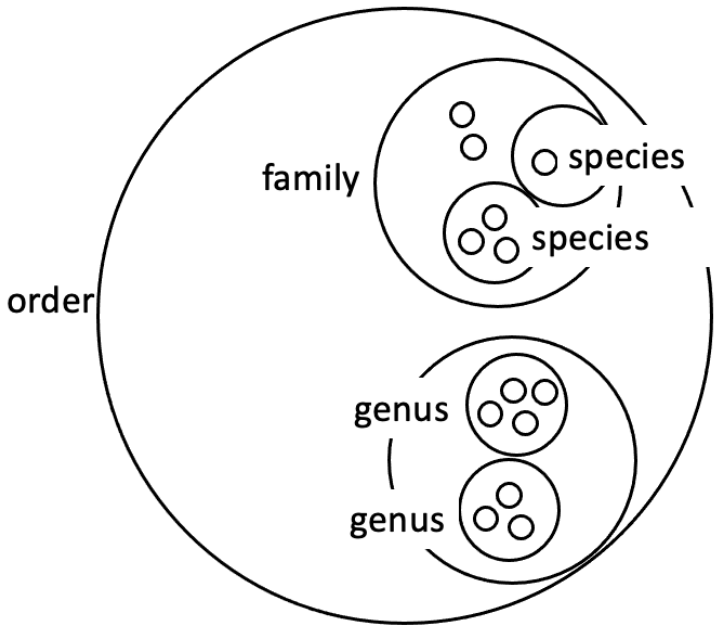


Figure 1.3: Biological classification reflecting evolutionary history.There are 64 currently accepted flowering plant orders, 416 families [1], around 13,000 genera and 300,000 species.

from the definition of monoploidy, that in a monoploid genome, each gene can be adjacent to at most two other genes. We call this condition the linearity prerequisite. In fact, the set of adjacencies in a monoploid is mathematically equivalent to the set of gene orders over all its chromosomes.

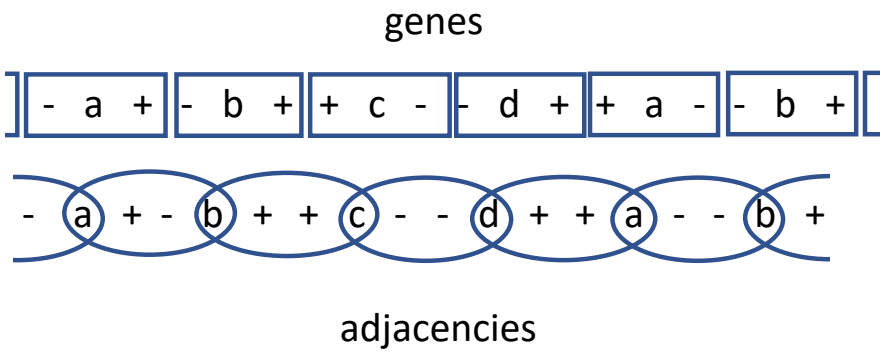


Figure 1.4: Top: A segment of a chromosomal gene order showing six successive genes and their orientations. Bottom: The adjacencies determined by these six genes. Note that the first and last of these adjacencies are not the same, because the orientation of of the two occurrences of gene “a" have opposite orientation.

The basic hypothesis underlying the work in this thesis is that in a phylogeny of extant species representing a plant order, possibly including species that have undergone one or more WGD, a sufficient number of adjacencies from the founding ancestor will have been conserved among the present-day genomes that can serve to reconstruct this ancestor.

Thus our strategy is to collect all the adjacencies in all the extant genomes and to pick out the subset of adjacencies that are both frequent (highly weighted) and that satisfy the linearity prerequisite. This becomes the well-known “Maximum Weighted Matching" (MWM) problem studied in combinatorial optimization.

There was no guarantee before starting my research that this strategy could yield meaningful results; i.e. that there is sufficient signal in the extant adjacencies to recover the ancestral genome. But the outcome presented in the following chapters show that this gamble has paid off.

1.6 Thesis overview

This thesis contains five published papers presented as Chapters 2 through 6.

- Qiaoji Xu, Lingling Jin, Yue Zhang, Xiaomeng Zhang, Chunfang Zheng, James H. Leebens-Mack, and David Sankoff. RACCROCHE: Ancestral flowering plant

chromosomes and gene orders based on generalized adjacencies and chromosomal gene co-occurrences. In: Jha, S.K., Mandoiu, I., Rajasekaran, S., Skums, P., Zelikovsky, A. (eds) Computational Advances in Bio and Medical Sciences. ICCABS 2020. Lecture Notes in Computer Science, vol 12686. Springer, Cham.[10].

- Qiaoji Xu, Lingling Jin, Chunfang Zheng, James H. Leebens-Mack, and David Sankoff. Validation of automated chromosome recovery in the reconstruction of ancestral gene order. Algorithms, 14, 2021 [11].

- Andre S. Chanderbali, Lingling Jin, Qiaoji Xu, Yue Zhang, Jingbo Zhang, Shuguang Jian, Emily Carroll, David Sankoff, Victor A. Albert, Dianella G. Howarth, Douglas E. Soltis and Pamela S. Soltis. Buxus and Tetracentron genomes help resolve eudicot genome history. Nat Commun, 13:643, 2022 [12].

- Qiaoji Xu, Lingling Jin, Chunfang Zheng, Xiaomeng Zhang, James Leebens-Mack and David Sankoff. From comparative gene content and gene order to ancestral contigs, chromosomes and karyotypes. Scientific Reports, 13:6095, 2023 [13].

- Qiaoji Xu and David Sankoff. Gene order phylogeny via ancestral genome reconstruction under Dollo. In: Jahn, K., Vinar, T. (eds) Comparative Genomics. RECOMB-CG 2023. Lecture Notes in Computer Science, vol 13883. Springer. [14].

In Chapter 2, we describe our new pipeline, called RACCROCHE, for reconstructing ancestral genomes and applied it to a set of representative genomes from six monocot orders [10]. The steps in this pipeline, starting from a given phylogenetic tree relating a set of extant genomes, each subdivided into a number of chromosomes, which are each made up of a linear ordering of signed (or oriented: left to right or right to left) genes:

- Partition all the genes in all the genomes into groups of orthologous genes, using sequence comparison methods, and relabel the genes in all genomes according the orthology groups they belong to.

- Extract all pairs of adjacent (or neighbouring) genes from all genomes, taking into account their orientation, as in Figure 1.4, and tabulate the overall frequency of each of these adjacencies.

- For each ancestor node in the tree, retain only those adjacency pairs that are phylogenetically relevant to that node, and run the MWM algorithm to select the linearly compatible subset of pairs of highest overall sum of frequencies. These pairs link together naturally to define a set of “contigs", linearly ordered sets of genes that are effectively fragments of chromosomes.

- All pairs of contigs for an ancestral node are examined to see how frequently that pair co-occurs on the same chromosomes in the extant genomes, and if so, in what left-to-right or right-to-left order. With the help of a heat map visualization, the co-occurrence matrix that results is then clustered into a number of proposed chromosomes for the ancestral genome associated with that node, each chromosome made up of the contigs in the same cluster, appropriately ordered.

Chapter 3 works over some the same material as the previous one, adding some improvements to the visualization and clustering methodology. More important, it contains a close simulation of the evolutionary processes that have affected the monocots. [11].

The method is validated by a simulation model which simulates the real descendant genomes of the first ancestor and then reconstructs the remaining ancestors by RACCROCHE. We can validate the accuracy by comparing the simulated ancestors and the real ancestors, which shows a majority of overlap. The main objective of this paper is to closely simulate the evolutionary process giving rise to the gene content and order of a set of extant genomes, and to assess to what extent an updated version of RACCROCHE can recover the artificial ancestral genome at the root of the phylogenetic tree relating to the simulated genomes.

In Chapter 4 we show that *Buxus sinica* and *Tetracentron sinense* are both characterized by independent WGDs, resolve relationships among early-diverging eudicots and their respective genomes, and use the RACCROCHE pipeline to reconstruct ancestral genome structure at three key phylogenetic nodes of eudicot diversification [12]. Our reconstructions indicate genome structure remained relatively stable during early eudicot diversification, and reject hypotheses of inter-lineage hybridization between ancestral eudicot lineages. This chapter is of interest because it was during the research on this project that the RACCROCHE pipeline was first brought to completion.

Chapter 5 contains our most ambitious use of the pipeline, both for biological results it produced, and for the major extensions it incorporated into the method [13]. The technical innovations included:

- to remove a severe bias due to a proportion of very long contigs at the order or family level reconstruction, a bias not apparent in the wider phylogenetic scope of our earlier studies of the monocots and *Buxus*, we introduced our “g-mer" concept involving the fragmenting all contigs into equal size segments, before the chromosome assembly step,

- since the MWM solution is not unique, and some solutions are more conducive to disjoint clusters/chromosomes, we sampled 100 solutions of each MWM input for further evaluation,

- we introduced a gap statistic trajectory for estimating k, the number of clusters from the chromosomal co-occurrence matrix of contigs, where k is the inflection point of the trajectory.

The biological results of this study were remarkable. We were the first to use genome evidence to estimate the base chromosome number of eleven plant orders from the rosid/asterid groups. This number turned out to be 9, for all of the orders, even though our method is capable of finding smaller and larger numbers when warranted. Non-genomic studies, following Grant [9], had also found uniform numbers, though usually 8, for these orders, but without any confirmation based on genome sequence data.

In Chapter 6, we used the MWM approach in Raccroche to extend our work to the large phylogeny problem, namely the reconstruction of the topology of a phylogeny. This required a partial relaxation of the criterion of phylogenetic validity in constructing the set of adjacencies to admit to the MWM step. This approach turned out to produce accurate or almost accurate topolgies for three plant orders studied as examples [14].

1.7 Related work

Dobzhansky and Sturtevant [15] first studied reconstruction of ancestral gene orders and karyotypes in *Drosophila* chromosomes in 1938.

In computational genomics, there have been a variety of approaches to the reconstruction of ancestral plant genomes. [8, 16, 17, 18, 19]. Some of this has been based on rearrangement distances [20][21], where the idea is to minimize total distance between ancestors and their descendants in the phylogeny, but these methods do not scale well.

Most other reconstruction methods are fundamentally based on gene adjacencies in the extant genomes used as data. Some are more particularly focused on these adjacencies [22, 23].

Others have focused on building contiguous ancestral regions (CARS) [24, 25]. Here, the reconstruction of ancestral gene orders proceeds through the identification of local commonalities in the genomes of a number of extant descendant species through various merger, assembly and concatenation procedures to finally produce a set of chromosome fragments representing the ancestral genome. This approach, introduced successfully in the context of mammalian genomes, where there are no polyploidizations since the common ancestor, and then taken over to plant genomics, applies to a series of methods of which a recent improved exemplar is proCARs [26]. As we discuss in Chapter 3, in the case of flowering plants, the avoidance of premature selection of gene adjacencies in Raccroche allows the recovery of more of the ancestral genome than proCARs.

A recent reconstruction method, AGORA [27], has been applied to several large data sets, and is similar to our pipeline in that it reconstructs all the ancestors in a phylogeny independently, but it requires a gene-tree reconciliation step involving all ancestors and all genes before the reconstruction step and before even considering adjacencies. Most important it does not depend on MWM and it is not restricted to monoploid output. It can reconstruct large fragments of chromosomes, but lacking any clustering step, cannot create an entire chromosome-level karyotype.

Chapter 2

The RACCROCHE Pipeline

Qiaoji Xu, Lingling Jin, Yue Zhang, Xiaomeng Zhang, Chunfang Zheng, James H. Leebens-Mack, and David Sankoff. RACCROCHE: Ancestral flowering plant chromosomes and gene orders based on generalized adjacencies and chromosomal gene co-occurrences. In: Jha, S.K., Mandoiu, I., Rajasekaran, S., Skums, P., Zelikovsky, A. (eds) Computational Advances in Bio and Medical Sciences. ICCABS 2020. Lecture Notes in Computer Science, vol 12686. Springer, Cham.

RACCROCHE: ancestral flowering plant chromosomes and gene orders based on generalized adjacencies and chromosomal gene co-occurrences

Qiaoji Xu[1][0000−0003−3316−2172],
Lingling Jin[2][0000−0002−4586−2347],
Chunfang Zheng[3][0000−0003−4205−4501],
James H. Leeben-Mack[4][0000−0003−4811−2231], and
David Sankoff[5][0000−0001−8415−5189]

Abstract. Given the phylogenetic relationships of several extant species, the reconstruction of their ancestral genomes at the gene and chromosome level is made difficult by the cycles of whole genome doubling followed by fractionation in plant lineages. Fractionation scrambles the gene adjacencies that enable existing reconstruction methods. We propose an alternative approach that postpones the selection of gene adjacencies for reconstructing small ancestral segments and instead accumulates a very large number of syntenically validated candidate adjacencies to produce long ancestral contigs through maximum weight matching. Likewise, we do not construct chromosomes by successively piecing together contigs into larger segments, but instead count all contig co-occurrences on the input genomes and cluster these, so that chromosomal assemblies of contigs all emerge naturally ordered at each ancestral node of the phylogeny. These strategies result in substantially more complete reconstructions than existing methods. We deploy a number of quality measures: contig lengths, continuity of contig structure on successive ancestors, coverage of the reconstruction on the input genomes, and rearrangement implications of the chromosomal structures obtained. The reconstructed ancestors can be functionally annotated and are visualized by painting the ancestral projections on the descendant genomes, and by highlighting syntenic ancestor-descendant relationships. We apply our methods to genomes drawn from a broad range of monocot orders, confirming the tetraploidization event "tau" in the stem lineage between the alismatids and the lilioids.

Keywords: Genome reconstruction · Gene order · Polyploidization · Fractionation · Monocots · Generalized adjacencies · Multiple orthology · Safe phylogeny · Maximum weight matching · Co-occurrence matrix · Complete-link clustering · Linear ordering problem.

1 Introduction

Reconstruction methods depending on conserved gene adjacencies tend to break down in plants, largely because the history of whole genome doubling and tripling events (WGD and WGT, respectively) in the lineages of plants. All known flowering plant genomes (except *Amborella trichopoda* [1]) have at least one, and often several, WGDs or WGTs in their lineages since the ancestral angiosperm, followed by extensive loss of redundant genes, largely randomly distributed along one or other of the duplicated chromosomes. These processes effectively scramble gene order and disrupt most adjacencies. Subsequently, most of the sets of duplicate or triplicate genes created by WGD/WGT events are reduced sooner or later to a single gene, by the redundance-eliminating process known as gene fractionation. Because of this fractionation, duplication of a genome fragment containing genes in the order 1-2-3-4-5-6, for example, may result in two surviving orders 1-3-5 and 2-4-6, with none of the five fragment-internal adjacencies conserved, and only one adjacency at most conserved with the chromosomal regions surrounding each copy of the fragment. The situation is compounded if there are several WGD or WGT events in the history of some of the present-day genomes. All this is superimposed on a background of gene family expansion through tandem duplication or other mechanisms, and loss of genes from species for which they are no longer physiologically or ecologically essential, genome rearrangement and other processes, all of which disrupt adjacencies independently of the fractionation process.

For this paper, we developed a pipeline for ancestral plant genome inference, `RACCROCHE`, **R**econstruction of **A**n**C**estral **CO**ntigs and **CH**romosom**E**s, including some intermediate ancestral genomes giving rise to major plant subgroupings. The new strategy implemented in our approach combines six fundamental components:

1. The replacement of the traditional selection of 1-1 orthologs among input genomes, as a first step, by the identification of many-to-many correspondences among gene families of limited size within these genomes.
2. The use of generalized adjacencies [18,17], namely any pair of genes close to each other on a chromosome, instead of just immediately adjacent genes.

These first two components avoid premature decisions on which orthologies and which adjacencies should be incorporated in the final reconstruction, in contrast to approaches which insist on making these decisions early in the reconstuction process, e.g., [11].

3. The compilation of oriented candidate adjacencies at each of the ancestral nodes of a given binary branching tree phylogeny using a "safe" criterion - that such an adjacency must be evidenced in genomes in two or three of the subtrees connected by this node, not just one or none.
4. The large set of these candidates is then resolved, at each node, by maximum weight matching (MWM) to give an optimally compatible subset, which ipso facto defines linearly (or circularly) compatible "contigs" of the ancestral genomes to be constructed, thus avoiding the branching segments that plague other methods [14].
5. A local sequence matching, satisfying proximity and contiguity conditions, of each contig on all of the chromosomes of the input genomes. This step includes the construction of a total chromosomal co-occurrence matrix of contigs belonging to each ancestral node.
6. A clustering applied to the co-occurrence matrix. This is then decomposed into chromosomal sets of contigs, with the aid of a heat map comparison of the contigs as organized by the clustering. Within each contig, the order of the genes is already predetermined by the MWM step. Ordering the contigs along the chromosomes is carried out by a linear ordering algorithm. The assignment and ordering of contigs to construct entire chromosomes, and not just a collection of small regions, is a advance over previous methods. Corresponding chromosomes in different ancestral genomes can be identified by the similar contigs they contain.

The results of this pipeline are mapped back to the input genomes, indicating how these extant genomes were derived through chromosomal rearrangements from their immediate ancestral genome.

We provide an evaluation of the reconstruction in terms of the sizes of the ancient chromosomal fragments found, the coherence (or continuity) between adjacent ancestral genomes, the coverage of the ancestors when mapped to extant genomes, and the "choppiness" of this mapping in terms of ancestor-descendant rearrangement.

There has been much recent work on the reconstruction of ancestral plant genomes [19,10,4,12,3]; on the computational side most of this has been based on common gene adjacencies in extant genomes, as summarized in such structures as sets of species trees and contiguous ancestral regions (CARS) [2]. The latter terminology, introduced successfully in the context of mammalian genomes [7], where there are no polyploidizations since the common ancestor, and then taken over to plant genomics [5,4,12], applies to a series of methods of which a recent improved exemplar is `proCARs` [11]. We will show that in the case of flowering plants, the avoidance of premature selection of gene adjacencies in `RACCROCHE` allows the recovery of more of the ancestral genome than `proCARs`.

The rest of the paper is organized as follows. Section 2 presents the features and procedure of the algorithm. (Most of the details appear in appendices.) An application of the `RACCROCHE` pipeline is shown in Section 3 with a focus on the reconstruction of the four monocot ancestors in the known phylogeny relating six extant monocot plant genomes. These include *Acorus calamus* (sweet flag) from the order Acorales, *Spirodela polyrhiza* (duckweed) from the order Alismatales, *Dioscorea rotundata* (yam) from the order Dioscorales, *Asparagus officinalis* (asparagus) from the order Aspargales, *Elaeis guineensis* (African oil palm) from the order Arecales and *Ananas comosus* (pineapple) from the order Poales. This includes an evaluation of the reconstruction in terms of the sizes of the ancient chromosomal fragments found, the coherence between adjacent ancestral genomes, the coverage of the ancestors when mapped to extant genomes, and the "choppiness" of this mapping in terms of ancestory descendant rearrangement. Section 4 concludes the paper and outlines some future directions.

2 Methods

2.1 Input

The input to `RACCROCHE` consists of N annotated extant genomes related by a given unrooted binary branching phylogeny, and a number of parameters, including

W: window size to include generalized as well as immediate adjacencies,
NF: largest total gene family size allowed in ortholog grouping in all extant genomes,
NG: largest gene family size allowed in any one genome,
NC: the number of longest contigs in ancestral genomes to be matched to extant genomes,
K: the desired number of chromosomes for each ancestor,
DIS: the maximum distance between two adjacent genes in an extant genome to be matched with adjacent genes in an ancestral contig.

Fig. 1 depicts the overall flow of the `RACCROCHE` pipeline.

2.2 The pipeline

Step 1: Pre-process gene families Pre-processing for the `RACCROCHE` procedure starts with syntenically validated orthogroups, or gene families, constructed from $\frac{1}{2}(N^2 + N)$ between-genome and self-comparison sets of pairwise `SynMap` synteny blocks by accumulating all genes that are syntenically orthologous to at least one other gene in the family. It retains only those families with at most a preset number NF of members and at most NG members in any particular genome. Without loss of generality, $NF \leq N \times NG$.

The use of syntenically validated adjacencies only, restricted to genes appearing in synteny blocks identified by the comparison of some pair of the descendant genomes, avoids generating huge gene families and astronomical numbers of adjacencies not reflective of the ancestor.

An optional second "redistribution" step for genes in large families is described in Appendix A.

Step 2: List generalized adjacencies For each of the N extant genomes, `RACCROCHE` compiles all generalized adjacencies, i.e., representatives of two gene families, occurring within a window of a preset size, W, in the order of genes on a chromosome. The adjacencies are oriented by the DNA

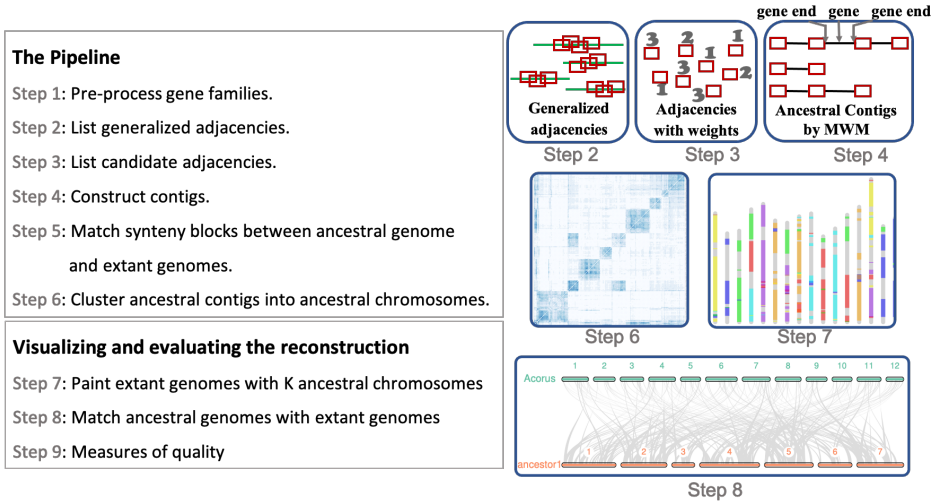


Fig. 1: Overall flow of the RACCROCHE procedure.

strand or strands containing the two genes, so that we can distinguish the two ends of each gene, and identify which ends are involved in the adjacency.

Step 3: List candidate adjacencies For each ancestral tree node, allow only adjacencies in occurring in two or three of the three subtrees connected by a branch incident to that node as candidates to be adjacencies in the corresponding ancestral genome. Occurrence in a subtree means occurrence in at least one of the extant genomes in that subtree.

Step 4: Construct contigs With candidate adjacencies weighted 2 or 3 according to whether they occur in 2 or 3 subtrees, use maximum weight matching to extract the highest weight set of compatible adjacencies, i.e., each gene end is matched to at most one other gene end, which automatically defines a set of disjoint linear contigs for the ancestral genome.

A method for improving the coherence of successive ancestors is discussed in Appendix B. This comes at the cost of other qualities of the contigs, and will not be discussed further here.

Step 5: Match synteny blocks between ancestral genome and extant genomes For each of the NC longest contigs of an ancestral genome, search for locally matched regions - synteny blocks - in all N extant genomes. This process is formally described in Appendix C.

Step 6: Cluster ancestral contigs into ancestral chromosomes Clustering of ancestral chromosomes is based on co-occurrence of ancestral contigs of sufficient size on the same chromosomes of extant genomes. First, a co-occurrence matrix is constructed on the set of contigs counting the cumulative number of times two different contigs are matched on the same chromosome in one or more extant genomes. Next, a complete-link clustering of the contigs is performed in each ancestral genome, based on the co-occurrence matrix. The hierarchical cluster thus produced is decomposed either automatically (e.g., with a cut-off level or with a cluster size criterion) or with some biologically-motivated manual intervention into a preset number K of chromosomes. See Section 3.2 below for an example.

Contigs are ordered by applying the algorithm of linear ordering problem [13] based on the count of relative ordering, the number of times each contig appears upstream/downstream of the other contig for every pair of contigs within a cluster.

The clustering and ordering are detailed in Appendix D. These procedures have been validated through simulation studies [16].

2.3 Visualizing and evaluating the reconstruction

Step 7: Painting the extant genomes according to the ancestral chromosomes Each of the K chromosomes of an ancestor genome is assigned a different colour. Each extant genome can then be painted by the colours of an ancestor based on the coordinates of synteny blocks calculated in Step 5. Unpainted regions less than 1Mb long between two blocks of the same colour are also painted with that colour. Although we can establish a general correspondence between the chromosomes of the successive ancestor genomes, the synteny blocks and the painting of the extant genomes will nevertheless depend on which ancestor is used. Generally the immediate ancestor of a genome gives the most meaningful painting.

Step 8: Adapting MCScanX to match ancestral genomes with extant genomes We use MCScanX [15] to connect matching parts of each descendant and its immediate ancestor, as well as to calculate the optimal order of chromosomes.

MCScanX requires both gene location and gene sequence to search pairwise synteny. The "genes" in the constructed ancestors, however, are really gene familes, each represented by a integer label. For the purposes of MCScanX, we simply choose a member of the gene family, either randomly, or from a descendant of that ancestor.

For viewing purposes, the number of "crossing" lines in the trace diagram should be minimized. MCScanX searches for the ordering of the chromosomes that minimizes this, using a genetic algorithm.

Step 9: Measures of Quality. In the construction of the contigs, we count how many gene families and how many candidate adjacencies are incorporated in total by the MWM and in the longest NC chromosomes. We also document details of the *contig length distribution*, e.g., the longest contig and N50.

The *coherence* between all pairs of contig sets, each set associated with one ancestor is a way of more global way of assessing the reconstruction. To be credible, the contigs at one ancestral node should resemble to some extent the contigs at a neighbouring ancestor.

A measure of commonality between two contigs i and j from two ancestors I and J respectively, is given by

$$\text{sim}_{ij} = \frac{x_{ij}}{\sqrt{x_{i.}x_{.j}}}, \tag{1}$$

where $x_{i.}, x_{.j}$ and x_{ij} are the numbers of gene families in contig i, in contig j and in both contigs, respectively.

Then, calculating the coherence between two tree nodes for the NC longest contigs.

$$\text{coherence}_{\text{IJ}} = \frac{\sum_i \max_{j=1}^{NC} \text{sim}_{ij}}{NC}. \tag{2}$$

Percent coverage is defined as the percentage that genome G is covered by the synteny block set of ancestor A. It also reflects how closely ancestor A is related to G.

Choppiness of painting in G is quantitatively measured by the number of different colours, T, the number of single-colour regions, R, and the number of small stripes, X, on each extant chromosome [9]. T is defined as the sum number of different colours on each chromosome of G minus 1, reflecting how much inter-chromosomal exchange, such as translocation, there has been; R is defined as the sum number of single-colour regions on each chromosome of G and is a measure of how much intra-chromosomal movement (e.g., reversals or transpositions) there has been; X is defined as the number of stripes less than a certain threshold size (i.e. 300 Kbp), which we deduct to avoid inflating R. The choppiness measure of painting in G is written as $R - X$.

2.4 Ancestral gene function

To aid in future studies of the genomic organization of gene function, a GO-term enrichment analysis of the members of each gene family is implemented to produce a functional annotation for the inferred ancestral genes. The details are reported in Appendix E, but are not applied in this paper.

3 Reconstruction of monocot ancestors

We applied our method to the reconstruction of four monocot ancestors, given six extant monocot plant genomes from *Acorus calamus* (sweet flag), *Spirodela polyrhiza* (duckweed), *Dioscorea rotundata* (yam), *Asparagus officinalis* (asparagus), *Elaeis guineensis* (African oil palm) and *Ananas comosus* (pineapple). The phylogenetic tree is shown in Fig. 2. The divergence time from Ancestor 1 to any of the extant genomes is about 130 Mya [6]. The reconstruction problem is difficult due not only to this lengthy elapsed time, since the early Cretaceous, comparable to that of the early divergence of placental mammals, but also to the occurrence of at least one WGD in every order, and generally two or more.

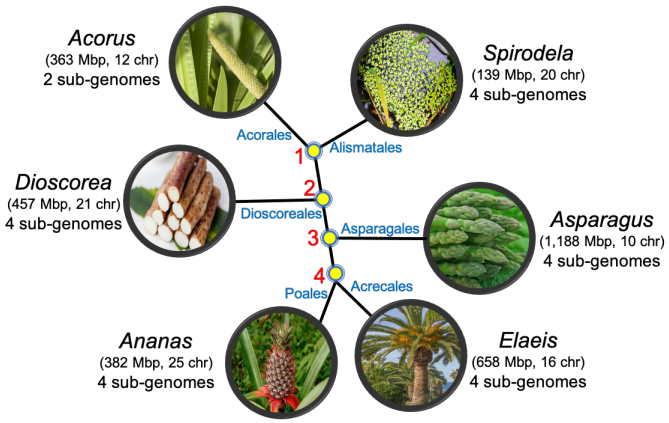


Fig. 2: Phylogeny showing relationships among six monocots and their ancestors.

One question we aimed to answer was whether both ancient WGD detected in the extant *Dioscorea* genome occurred after its branching off the stem lineage to Asparagales, Arecales and Poales, or whether one of these WGD occurred earlier, between Ancestors 1 and 2, and is identical to the "tau" event known to affect all these later branching orders.

3.1 Properties of the contig reconstruction

After numerous trials, input parameters that seemed (somewhat subjectively) to balance contig length properties, coherence and coverage were chosen to be window size $W = 7$, maximum total family size $NF = 50$ and within-genome maximum family size $NG = 10$. Table 1 summarizes the gene content of each of the input genomes, first, syntenically validated genes (i.e., in synteny blocks); second, after removing very large gene families; third, after filtering for within-genome family size; fourth, genes present in a candidate adjacency; fifth, genes incorporated in the 250 longest contigs for any ancestor.

Table 1: Numbers of genes at each step of building contigs.

	in synteny blocks	in families < 5000	in filtered families	in candidate adjacencies	in contigs, after MWM
Acorus	21,308	11,807	11,300	10,189	9,649
Spirodela	20,751	8,385	8,005	7,706	7,276
Dioscorea	19,240	8,256	7,873	7,485	7,141
Asparagus	28,141	10,109	9,645	9,128	8,750
Ananas	27,024	11,744	11,180	10,623	10,116
Elaeis	21,425	12,833	12,227	11,831	11,369

Recall that to be a candidate, an adjacency must appear at least once in at least two different genomes, thus satisfying the safety criterion for at least one ancestor. Applying the MWM algorithm to the set of candidates greatly reduces the number in selecting the best linearized subset, as documented in Table 2.

Table 2: Input adjacencies to MWM, and output.

	Ancestor 1	Ancestor 2	Ancestor 3	Ancestor 4
candidate adjacencies	35,165	41,963	47,118	48,452
MWM adjacencies	6,335	6,847	7,244	7,310

The contigs that are formed by the MWM matches are of moderate length, as suggested by Table 3. The longest one contains 84-89 genes and the last one retained ($NC = 250$) contains around 10 genes. We then locate all the matches of these contigs on the chromosomes of the extant genomes.

Table 3: Contig statistics for the four ancestors. The number of genes in a contig measures its length.

	longest	total number	N50		N60		N70	
	contig	of contigs	length	number	length	number	length	number
Ancestor 1	84	3,950	10	249	5	403	1	662
Ancestor 2	89	3,441	12	219	8	292	3	510
Ancestor 3	85	3,043	15	169	10	252	5	393
Ancestor 4	88	2,975	17	151	12	215	6	342

A good proportion of the MWM adjacencies will be shared by successive (or all) ancestors, and many contigs will be similar from ancestor to ancestor. Table 4 displays the coherence among the contig sets for the four ancestor genomes.

Table 4: Coherence among ancestors.

	Ancestor 1	Ancestor 2	Ancestor 3	Ancestor 4
Ancestor 1	1.000			
Ancestor 2	0.430	1.000		
Ancestor 3	0.361	0.443	1.000	
Ancestor 4	0.318	0.357	0.419	1.000

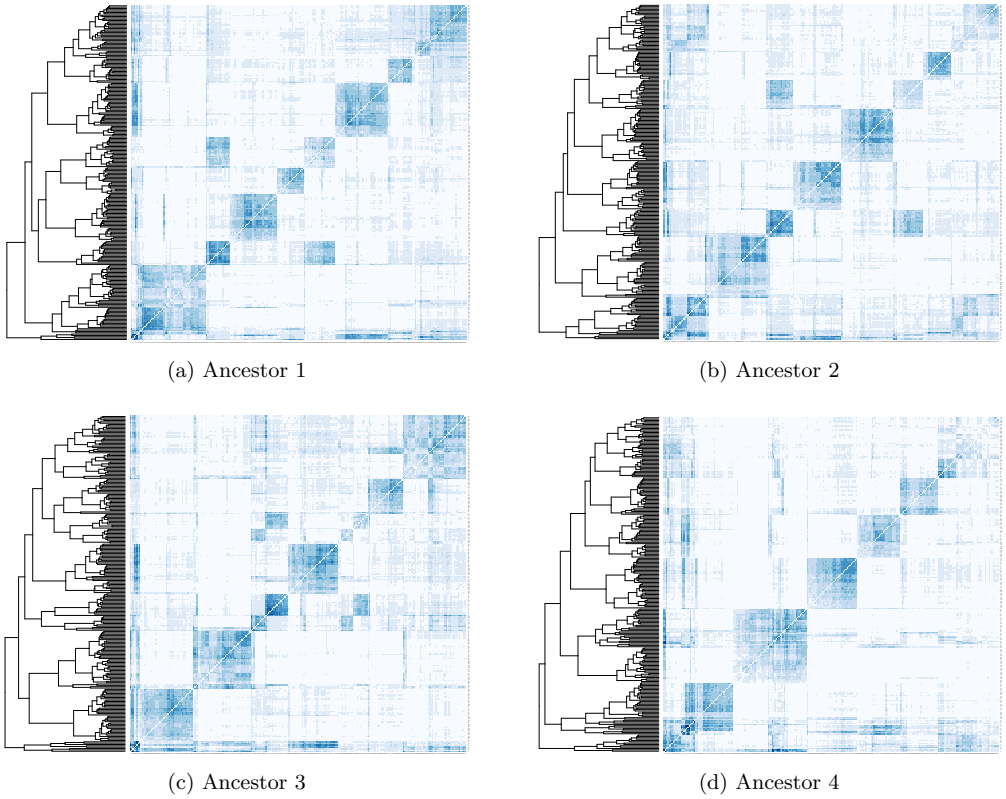
3.2 Clustering

The choice of complete link method of hierarchical clustering is appropriate in the context of searching for balanced clusters at all levels, and avoiding an asymmetric "chaining" effect. Chromosomes in a genome tend to be roughly the same order of magnitude, which therefore suggests complete link.

The hierarchical cluster of the 250 longest contigs according to their chromosomal co-occurrence (Section 2.2) is seen beside each panel in Figure 3. The intensity of the shading of each cell in the heat map reflects how frequently the corresponding contigs co-occur in the extant genomes. In each case seven large, darkly shaded, blocks emerge neatly from the map, thus constituting the chromosomes of the ancestral genome. Table 5 contains statistics on the chromosomes and contigs.

Table 5: Contigs and genes in ancestral chromosomes.

chromosome	Ancestor 1		Ancestor 2		Ancestor 3		Ancestor 4	
	contigs	genes	contigs	genes	contigs	genes	contigs	genes
1	43	857	42	1,398	40	1,909	44	1,911
2	40	729	43	585	43	683	46	703
3	23	363	21	443	22	467	18	620
4	44	951	39	671	42	853	38	917
5	41	773	43	894	32	656	40	810
6	23	536	23	666	30	958	31	985
7	36	743	39	844	41	497	33	411
total	250	4,952	250	5,501	250	6,013	250	6,357



(a) Ancestor 1

(b) Ancestor 2

(c) Ancestor 3

(d) Ancestor 4

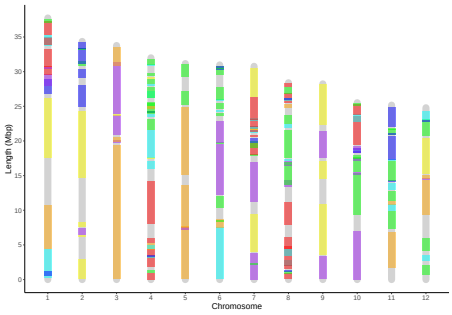
Fig. 3: Heat maps of the four ancestors showing the clusters of contigs making up ancestral chromosomes from the longest 250 contigs by the complete-link clustering algorithm.

3.3 Painting the chromosomes of the present-day genomes

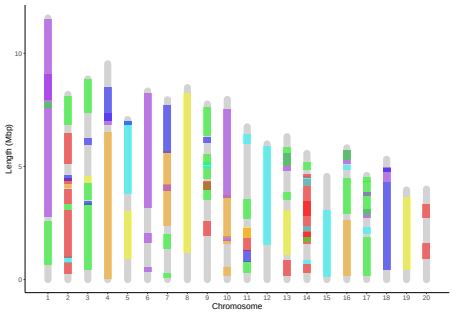
Each chromosome in an ancestor genome is assigned a colour. Despite the genome rearrangements intervening between an earlier ancestor and a later one, corresponding chromosomes in different ancestral genomes can be identified by similarity in the gene content of their constituent contigs. This correspondence, though it disrupted in many places by interchromosomal exchanges, is reflected in the chromosomal colour assignment in the four ancestors. The colours are then projected on to the chromosomes of the extant genomes that served as inputs to the pipeline, based on the contig matches detected in Section 3.1. Painting is carried out as described in Section 2.3 and is depicted in Figure 4.

3.4 Evaluation

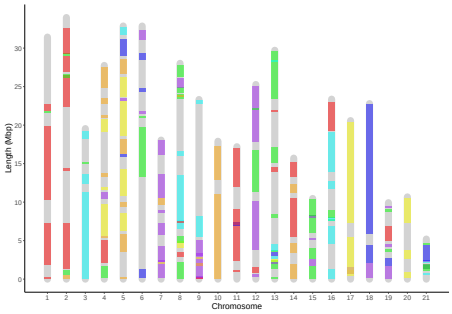
Tables 6 and 7 provide quality assessments of the reconstruction as manifest in the painted extant genomes. In Table 6 we see a high degree of coverage of the extant genomes, while Table 7 shows a degree of choppiness that is moderate, given the time scale involved. Ancestors 1 and 2 achieve better coverage of all the extant genomes, even though most of the genomes were more directly involved in the reconstruction of Ancestors 3 and 4. This may be an artifact of the sparsity of matches from Ancestors 1 and 2, so that the inter-block colouring discussed in Section 2.3 can cover longer, uninterrupted, regions of the chromosomes. A similar sparsity explanation can also be entertained for the low degree of choppiness of the paintings on the *Spirodela* genome, despite its higher degree of polyploidy than *Acorus*.



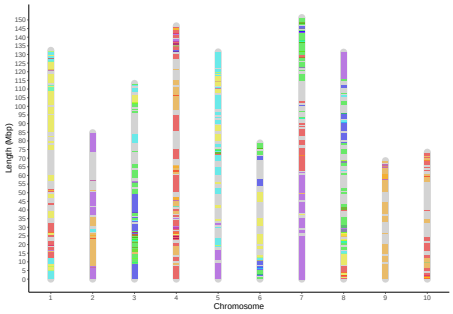
(a) Ancestor 1 painted on *Acorus*.



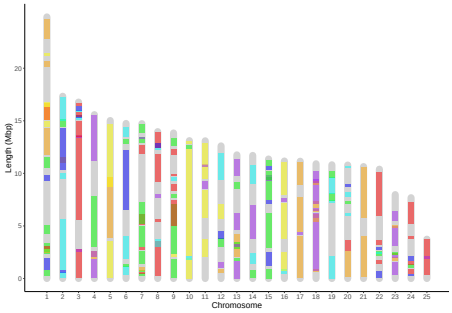
(b) Ancestor 1 painted on *Spirodela*.



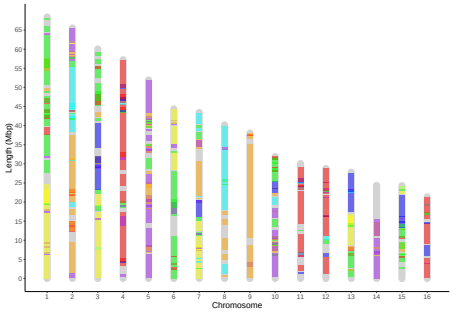
(c) Ancestor 2 painted on *Dioscorea*.



(d) Ancestor 3 painted on *Asparagus*.



(e) Ancestor 4 painted on *Ananas*.



(f) Ancestor 4 painted on *Elaeis*.

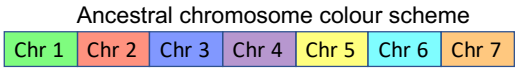


Fig. 4: Chromosome painting of extant genomes according to the colour assignment in their immediate ancestors. Ancestral blocks shorter than 150 Kbp are not shown.

Table 6: Percent coverage of extant genomes by ancestral chromosomes.

	Ancestor 1	Ancestor 2	Ancestor 3	Ancestor 4
Acorus	81%	80%	82%	83%
Spirodela	74%	78%	80%	81%
Dioscorea	54%	61%	62%	63%
Asparagus	63%	62%	66%	71%
Ananas	62%	69%	71%	70%
Elaeis	75%	79%	83%	84%

Table 7: Choppiness of painting on extant genomes. T reflects how much inter-chromosomal exchange has occurred, $R - T$ is a measure of intra-chromosomal movement (e.g., reversals or transpositions) and X is the number of small stripes shorter than 300 Kbp, which misleadingly inflates R.

T	*Acorus*	*Spirodela*	*Dioscorea*	*Asparagus*	*Ananas*	*Elaeis*
Ancestor 1	45	33	48	40	48	59
Ancestor 2	38	22	45	38	38	57
Ancestor 3	48	36	48	43	42	60
Ancestor 4	50	39	57	47	55	65
$R - T$	*Acorus*	*Spirodela*	*Dioscorea*	*Asparagus*	*Ananas*	*Elaeis*
Ancestor 1	122	56	128	233	88	193
Ancestor 2	95	34	107	220	94	161
Ancestor 3	129	51	131	284	104	194
Ancestor 4	172	75	140	331	124	247
$R - X$	*Acorus*	*Spirodela*	*Dioscorea*	*Asparagus*	*Ananas*	*Elaeis*
Ancestor 1	134	63	136	239	106	216
Ancestor 2	112	45	121	221	100	196
Ancestor 3	142	64	143	283	110	215
Ancestor 4	170	74	166	337	137	270

3.5 MCScanX visualization

A different view of the evolution of the monocot genomes via ancestral intermediates is obtained through connecting homologous synteny blocks in a MCScanX visualization, as laid out in Figure 5. Consistent with the history of extensive rearrangement evident in Figure 4 and Table 7, the patterns of MCScanX connections is rather complex. Nevertheless, we can find important relationships using the "highlight" feature of the software.

Thus, the comparison between Ancestor 1 and *Acorus* shows several chromosomal regions in the ancestor each linked to two regions in the extant genome, whereas the opposite pattern is non-existent. Similarly the comparison between Ancestor 1 and *Spirodela* also shows instances of a 1:4 pattern, consistent with the two WGDs inherited by this species.

The most interesting pattern, however, is that between Ancestors 1 and 2, which strongly suggests a duplication event occurring before the branching of the *Dioscorales* from the main monocot stem lineage. In contrast the Ancestor 2-Ancestor 3 and Ancestor 3-Ancestor 4 comparisons both show 1-1 patterns. Moreover, though dot-plot examination of *Dioscorea* evidences four subgenomes, thus two WGD in its history, the MCScanX diagram of Ancestor 2-*Dioscorea* only shows evidence of one event, confirming that one event must have predated Ancestor 2. This latter event is the one shared by all the more recently branching orders, known as "tau".

4 Discussions and Conclusions

This work explored an alternative approach to genome reconstruction by stepwise piecing together of small units. Instead, we compile a large number of potential components and use a combinatorial optimization approach to combining them, an approach explicitly disavowed by, e.g., [11]. We were motivated by the special case of plant comparative genomics, which has to deal with the aftermath

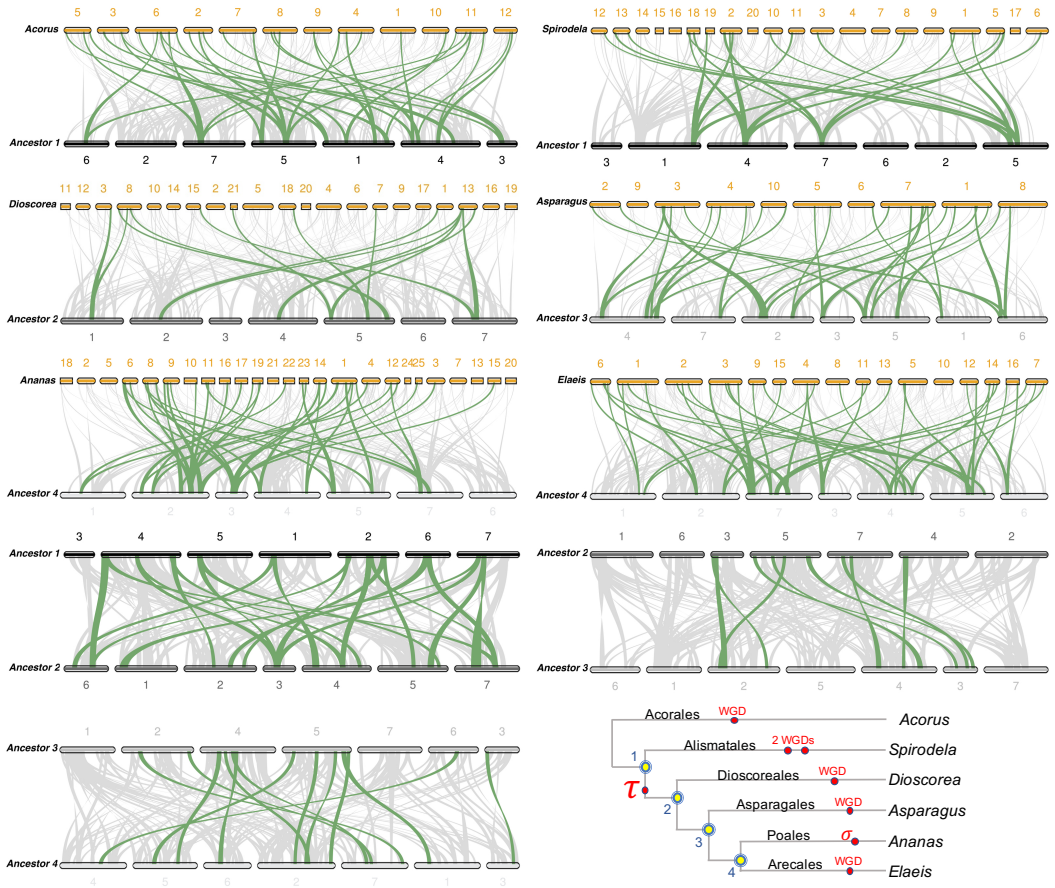


Fig. 5: Matching genomes, extant and ancestral, with their immediate ancestors.

or recurrent polyploidization and fractionation. Compared to approaches like `proCARs` [11] which is very successful in reconstructing ancestral animal genomes, `RACCROCHE` may work better with plant genomes, since it is designed to be robust against the gene order scrambling effect of fractionation.

Since the entities reconstructed by `proCARs` are not meant to be individual ancestral genes, but blocks of syntenically related genes identified at the level of extant genomes, it is hard to compare our inferred ancestral genomes, composed of hypothetical genes with identifiable functions, with the output of `proCARs`. In our hands `proCARs` identified 214 synteny blocks in our data, organized into "CARs" (contiguous ancestral regions) making up the ancestral genomes. These contained a total of 3,248 "universal seeds", which may be comparable to our ancestral genes, although our ancestors contained about twice as many. Insofar as these comparisons are valid, they confirm a role for `RACCROCHE` in plant comparative genomics.

One particular feature that stands out in this work, is the innovative clustering of counts of contig co-occurrences on extant chromosomes, followed by heatmap construction to identify ancestral chromosomes. Another is the use of MCScanX to locate a WGD on an internal branch of a phylogeny.

Acknowledgements

We thank the Department of Energy Joint Genome Institute staff and collaborators including David Kudrna, Jerry Jenkins, Jane Grimwood, Shengqiang Shu, and Jeremy Schmutz for pre-publication access to the *Acorus* genome sequence and annotation. Thanks to Aîda Ouangraoua for much help in implementing ProCARs [11] and Haibao Tang for prompt replies to queries about MCScanX [15].

Funding

Research supported by Discovery grants to LJ and DS from the Natural Sciences and Engineering Research Council of Canada. DS holds the Canada Research Chair in Mathematical Genomics.

Availability

The annotated genomic data is accessible on the `CoGe` platform https://genomevolution.org/coge/ and Phytozome. The pipeline is available at https://github.com/jin-repo/RACCROCHE.

References

1. Amborella Genome Project: The Amborella genome and the evolution of flowering plants. Science **342**(6165), 1241089 (2013)
2. Anselmetti, Y., Luhmann, N., Bérard, S., Tannier, E., Chauve, C.: Comparative methods for reconstructing ancient genome organization. In: Comparative Genomics, pp. 343–362. Springer (2018)
3. Avdeyev, P., Alexeev, N., Rong, Y., Alekseyev, M.A.: A unified ILP framework for core ancestral genome reconstruction problems. Bioinformatics **36**(10), 2993–3003 (2020)
4. Badouin, H., Gouzy, J., Grassa, C.J., Murat, F., Staton, S.E., Cottret, L., Lelandais-Brière, C., Owens, G.L., Carrère, S., Mayjonade, B., et al.: The sunflower genome provides insights into oil metabolism, flowering and asterid evolution. Nature **546**(7656), 148–152 (2017)
5. Chauve, C., Tannier, E.: A methodological framework for the reconstruction of contiguous regions of ancestral genomes and its application to mammalian genomes. PLoS Computational Biology **4**(11), e1000234 (2008)
6. Givnish, T.J., Zuluaga, A., Spalink, D., Gomez, M.S., Lam, V.K.Y., Saarela, J.M., Sass, C., Iles, W.J.D., de Sousa, D.J.L., Leebens-Mack, J., et al.: Monocot plastid phylogenomics, timeline, net rates of species diversification, the power of multi-gene analyses, and a functional model for the origin of monocots. American Journal of Botany **105**(11), 1888–1910 (2018)
7. Ma, J., Zhang, L., Suh, B.B., Raney, B.J., Burhans, R.C., Kent, W.J., Blanchette, M., Haussler, D., Miller, W.: Reconstructing contiguous regions of an ancestral genome. Genome research **16**(12), 1557–1565 (2006)
8. Martí, R., Reinelt, G., Duarte, A.: A benchmark library and a comparison of heuristic methods for the linear ordering problem. Computational optimization and applications **51**(3), 1297–1317 (2012)
9. Mazowita, M., Haque, L., Sankoff, D.: Stability of rearrangement measures in the comparison of genome sequences. Journal of Computational Biology **13**(2), 554–566 (2006)
10. Murat, F., Armero, A., Pont, C., Klopp, C., Salse, J.: Reconstructing the genome of the most recent common ancestor of flowering plants. Nature Genetics **49**, 490–496 (2017)
11. Perrin, A., Varré, J.S., Blanquart, S., Ouangraoua, A.: ProCARs: Progressive reconstruction of ancestral gene orders. BMC genomics **16**(S5), S6 (2015)
12. Rubert, D.P., Martinez, F.V., Stoye, J., Doerr, D.: Analysis of local genome rearrangement improves resolution of ancestral genomic maps in plants. BMC genomics **21**, 1–11 (2020)
13. Schiavinotto, T., Stützle, T.: The linear ordering problem: Instances, search space analysis and algorithms. Journal of Mathematical Modelling and Algorithms **3**(4), 367–402 (2004)
14. Tannier, E., Bazin, A., Davín, A., Guéguen, L., Bérard, S., Chauve, C.: Ancestral genome organization as a diagnosis tool for phylogenomics (2020)
15. Wang, Y., Tang, H., DeBarry, J.D., Xu Tan, J.L., Wang, X., Lee, T.H., Jin, H., Marler, B., Guo, H., Kissinger, J.C., Paterson, A.H.: MCScanX: a toolkit for detection and evolutionary analysis of gene synteny and collinearity. Nucleic Acids Research **40**(7), e49 (2012)
16. Xu, Q., Jin, L., Zheng, C., Leebens-Mack, J.H., Sankoff, D.: Validation of automated chromosome recovery in the reconstruction of ancestral gene order. Algorithms **14** (2021), in press
17. Xu, X., Sankoff, D.: Tests for gene clusters satisfying the generalized adjacency criterion. In: Brazilian Symposium on Bioinformatics. pp. 152–160. Springer (2008)
18. Yang, Z., Sankoff, D.: Natural parameter values for generalized gene adjacency. Journal of Computational Biology **17**(9), 1113–1128 (2010)

19. Zheng, C., Chen, E., Albert, V.A., Lyons, E., Sankoff, D.: Ancient eudicot hexaploidy meets ancestral eurosid gene order. BMC genomics **14**(S7), S3 (2013)

Appendices

A Redistributing genes from families exceeding upper size limits

As an optional second "redistribution" step, all families with more than NF members or more than NG members in any particular genome, are flagged. Then the construction of the families is repeated, with the restriction that no gene can be recruited to a family by virtue only of a similarity of less than some threshold homology level θ to a gene already in the family. The intent is to break up large families held together by a few weak links, and thus to retrieve some better supported smaller families.

B Modes of contig construction

`RACCROCHE` executes for a single set of W, NF, NG parameters, or for a range of values of W and NG. In the latter case, there is an option, designed to increase coherence among sets of contigs for successive ancestors, that the MWM for any combination of W and NG must be restricted to include all adjacencies already recovered for lesser values of W or NG, insofar as possible. Thus, starting with some small W and NG, we can construct MWM solutions for larger window size and/or larger gene family size, and hence sets of contigs, by incrementing one or the other of the parameters.

It is possible, however, to have conflicts between $W, NG-1$, and $W-1, NG$ analyses. For example if adjacencies (a, b) and (b, c) are in the MWM for $(W, NG-1)$ and (a, b) and (b, d) are in the MWM for $(W-1, NG)$, then a matching for W, G cannot be forced to include all matchings from the two previous MWM. To accommodate this possibility, when we restrict the MWM for (W, NG) to include all adjacencies from $(W, NG-1)$ and $(W-1, NG)$, we make an exception for any adjacencies from either that are in potential conflict with adjacencies from the other. Thus (a, b) in the example above might be obligatorily included, but (b, c) and (b, d) would not. Thus the MWM for (W, NG) might include (b, c) or (b, d), but not both.

C Matching contigs to chromosomes of extant genomes

For the ancestor genome, A, computed from a set of extant genomes neighbouring A, $G_{1\cdots n}$, perform the following steps.

1. Extract gene features of ancestor A in descendant genomes.
   For every gene, g, in ancestor A computed from Step 2, retrieve six features of this gene in every extant genome $G_{1\cdots n}$ involved in constructing ancestor A. The features of a gene include chromosome ID, start and end chromosomal positions, distance between g to its next adjacent gene in G_i, gene family ID labelled in Step 1, and contig ID in A, denoted as $g^{A\to G_i}(chr, start, end, distance, gf, ctg)$.
2. Map ancestor A to each of the descendant genomes.
   The ancestor will be mapped as ancestral syntenic blocks on the descendant genome in two steps. The first step initializes a syntenic block by merging two adjacent genes given a distance threshold DIS: merge two genes, g_1 and g_2, forming one ancestral syntenic block on G_i if g_1 and g_2 satisfy the following conditions:
   (a) g_1 and g_2 locate the same chromosome of G_i;
   (b) g_1 and g_2 are adjacent to each other; in other words, there could be a non-coding region but no other gene(s) between g_1 and g_2;
   (c) The distance between the two adjacent genes must be less than or equal to the distance threshold DIS (i.e. DIS=1 Mbp).

   The second step extends the above identified ancestral syntenic block by merging flanking gene(s) into the block if the gene(s) satisfies the above three conditions. It stops extending the block if no flanking gene could be merged into the block. After the two steps, an ancestral synteny block mapping A to G_i is denoted as $syntenyBlk(chr, start, end, ctg, len)$. The set of synteny blocks between A and G_i is
   $syntenyBlkSet^{A\to G_i} = \{syntenyBlk_k(chr, start, end, ctg, len) | 1 \le k \le m$, where m is the total number of synteny blocks mapping from A to $G_i\}$

D Construction of ancestral chromosomes

1. Filter the set of blocks longer than a block length threshold.
   Given a block length threshold, $blockLEN$, $\overline{syntenyBlkSet}^{A\to G_i}$ is a subset of $syntenyBlkSet^{A\to G_i}$, where each block in the set is longer than $blockLEN$ (i.e. $blockLEN$ = 150Kbp).

2. Count co-occurrence of ancestral contigs on same chromosomes.
   Based on $syntenyBlk.chr$ and $syntenyBlk.ctg$ of each pair of synteny block in $\overline{syntenyBlkSet}^{A\to G_i}$, gather the co-occurrence of ancestral contigs on the same extant chromosome. Write the co-occurrence result into the lower triangle of a $NC \times NC$ matrix, m, where the rows and columns are contigs with ID from 0 to $(NC-1)$, $m_{i,j}$ is the number of co-occurrence between contigs i and j, where $0 < j < i < NC-1$. The maximum co-occurrence frequency in m is denoted as $\max_{freq}$.

3. Cluster ancestral contigs into ancestral chromosomes according to pairwise distance matrix based on co-occurrence.
   A NC by NC distance matrix, $dmat$, is calculated as

$$dmat_{i,j} = -\log(\frac{\max_{freq} - m_{i,j}}{\max_{freq}}).$$

   This distance matrix is fed into the complete-link clustering algorithm. This can then be composed into K clusters, according to users' preferences. The resultant clusters of contigs correspond to ancestral chromosomes and their compositions.

Last, attach ancestral chromosome number as an attribute to each of the synteny block:

$$syntenyBlkSet^{A\to G_{1\cdots N}} = \{syntenyBlk_k(chr, start, end, ctg, len, ancestral_chr)\},$$

where $ancestral\_chr$ corresponds to the cluster ID which blk.ctg belong to.

To order the contigs along each chromosome, we proceed as follows.

After the $syntenyBlkSet^{A\to G_{1\cdots N}}$ is generated in Step 3, relative ordering between every pair of contigs is counted. The number of times each contig appears upstream/downstream of other contig is structured into an $NC \times NC$ ordering matrix, C, where the rows and columns are contig IDs from 0 to $NC-1$. $c_{i,j}$ represents the number of times contig i occurred in upstream of contig j in the extant chromosomes.

Given the ordering matrix C, the *linear ordering problem (LOP)* is the problem of finding a permutation π of the column and row indices $\{1, \cdots, NC\}$, such that the value

$$f(\pi) = \sum_{i=1}^{NC} \sum_{j=i+1}^{NC} C^{(\pi(i),\pi(j))} \tag{3}$$

is maximized [13]. In other words, the goal is to find a permutation of the columns and rows of C such that the sum of the elements in the upper triangle is maximized.

By applying a meta-heuristic solver of LOP, Tabu Search [8], the solution order corresponds to the ordering/permutation of contigs sorted by their positions along ancestral chromosomes.

E Functional annotation of ancestral genes

We create a set of all genes in all families represented by ancestral genes in the reconstructed ancestor. This is the background set. For each gene family, all the genes in the family constitute a query set for GO-term enrichment analysis against the background set. Significant terms that emerge constitute the functional annotation for the ancestral gene.

Chapter 3

Validation through Simulation

Qiaoji Xu, Lingling Jin, Chunfang Zheng, James H. Leebens-Mack, and David Sankoff. Validation of automated chromosome recovery in the reconstruction of ancestral gene order. Algorithms, 14, 2021.

algorithms

MDPI

Article

Validation of Automated Chromosome Recovery in the Reconstruction of Ancestral Gene Order

Qiaoji Xu [1], Lingling Jin [2], James H. Leebens-Mack [3] and David Sankoff [1,*]

[1] Department of Mathematics and Statistics, University of Ottawa, Ottawa, ON K1N 6N5, Canada
[2] Department of Computer Science, University of Saskatchewan, Saskatoon, SK S7N 5C9, Canada
[3] Department of Plant Biology, University of Georgia, Athens, GA 30602, USA
* Correspondence: sankoff@uottawa.ca

Abstract: The RACCROCHE pipeline reconstructs ancestral gene orders and chromosomal contents of the ancestral genomes at all internal vertices of a phylogenetic tree. The strategy is to accumulate a very large number of generalized adjacencies, phylogenetically justified for each ancestor, to produce long ancestral contigs through maximum weight matching. It constructs chromosomes by counting the frequencies of ancestral contig co-occurrences on the extant genomes, clustering these for each ancestor and ordering them. The main objective of this paper is to closely simulate the evolutionary process giving rise to the gene content and order of a set of extant genomes (six distantly related monocots), and to assess to what extent an updated version of RACCROCHE can recover the artificial ancestral genome at the root of the phylogenetic tree relating to the simulated genomes.

Keywords: genome reconstruction; gene order; polyploidization; fractionation; monocots; simulation; co-occurrence correlations; clustering

Citation: Xu, Q.; Jin, L.; Leebens-Mack, J.; Sankoff, D. Validation of Automated Chromosome Recovery in the Reconstruction of Ancestral Gene Order. *Algorithms* **2021**, *1*, 0. https://doi.org/

Academic Editor: Hélène Touzet and Aïda Ouangraoua

Received: 01 April 2021
Accepted: 18 May 2021
Published:

1. Introduction

The reconstruction of ancestral gene orders proceeds through the identification of local commonalities—synteny blocks, "CARs" (contiguous ancestral regions), contigs, microsyntenies—in the genomes of a number of extant descendant species through various merger, assembly and concatenation procedures to finally produce a set of chromosome fragments representing the ancestral genome. The final step, assembling these fragments into whole chromosomes, is usually unresolved by these methods [1–4].

We have previously proposed an approach, RACCROCHE, to reconstruction that postpones the selection of gene adjacencies for reconstructing small ancestral segments [5]. Instead, it accumulates a very large number of syntenically validated candidate adjacencies to produce long ancestral contigs through maximum weight matching. Moreover, it does not construct chromosomes by successively piecing together contigs into larger segments, but instead counts all contig co-occurrences on the extant genomes and clusters these so that chromosomal assemblies of ancestral contigs can be recognized at each ancestral node of the phylogeny.

Though the reconstruction procedure and clustering are automated, the crucial step between clusters and chromosomes has a "polishing" step requiring human intervention, including the assignment of many problematic contigs to clusters. In the present paper, we improve and automate the process of chromosome recovery through a more meaningful measure than raw contig co-occurrence, leading to the assignment of almost all contigs to clusters.

To validate the accuracy of the reconstruction, this paper describes a simulation of the evolutionary processes giving rise to the gene content and order of a set of extant genomes. This serves as a verification of the RACCROCHE reconstruction method and assesses to what extent RACCROCHE can recover the artificial ancestral genomes given that the ground truth is known for the simulated data.

2. RACCROCHE

We first sketch our algorithm for ancestral plant genome inference, RACCROCHE, **R**econstruction of **A**n**C**estral **CO**ntigs and **CH**romosom**E**s [5], including reconstruction of the intermediate ancestral genomes giving rise to modern species, designed with a particular focus on flowering plant evolution.

Algorithm 1: RACCROCHE – reconstruction of ancestral contigs and chromosomes

input : *Tr*, an unrooted binary branching phylogeny
H, the number of annotated extant genomes related by *Tr*
W, size of window including all generalized adjacencies
output: reconstructed ancestral chromosomes

1 generate gene families from pairwise SynMap comparisons of extant genomes;
2 **for** *genome i* ← 1 **to** *H* **do**
3 list all generalized adjacencies occurring within a window of a preset size *W*;
4 **end**
5 **foreach** *ancestor in Tr* **do**
6 assign adjacency weights as the number of subtrees connected by a branch to *ancestor*, containing at least one occurrence of the adjacency;
7 select candidate adjacencies with weights 2 or 3;
8 construct an adjacency graph from the candidate adjacencies;
9 construct contigs using Maximum Weight Matching from the adjacency graph;
10 match contigs to extant genomes;
11 count the frequency of co-occurrence of contigs on extant chromosomes;
12 cluster ancestral contigs into ancestral chromosomes according to contig co-occurrence;
13 **end**

The strategy implemented in this approach combines the following components:

1. In Line 1 of Algorithm 1, the replacement of the traditional selection of one-to-one orthologs among input genomes, as a first step, by the identification of many-to-many correspondences among gene families of limited size within these genomes from pairwise SynMap [6,7] comparisons.
2. In Line 3, the use of generalized adjacencies [8,9], namely, any pair of genes close to each other within a predefined window size on a chromosome, instead of just immediately adjacent genes.
3. In Lines 6–7, the compilation of oriented candidate adjacencies at each of the ancestral nodes of a given binary branching tree phylogeny using the "safe" criterion that such an adjacency must be evidenced in genomes in two or three of the subtrees connected by this node, not just one or none.
4. In Lines 8–9, the large set of these candidates is then resolved, at each node, by maximum weight matching (MWM) to give an optimally compatible subset, which ipso facto defines linearly (or circularly) compatible "contigs" of the ancestral genomes to be constructed, thus avoiding the branching segments that plague other methods [10]. Use of MWM for ancestral gene order reconstruction was introduced some time ago, but with modest results [11].
5. In Line 10, local sequence matching, satisfying proximity and contiguity conditions, of each ancestral contig on all of the chromosomes of the extant genomes, followed in Line 11 by the construction of a total chromosomal co-occurrence matrix of contigs belonging to each ancestral node.
6. In Line 12, a clustering applied to the co-occurrence matrix. This is then decomposed into chromosomal sets of closely clustered contigs. Within each contig, the order of the genes is already predetermined by the MWM step. Ordering the contigs along the chromosomes is carried out by a linear ordering algorithm.

The last two steps, leading to the reconstruction of a set of distinct chromosomes, without any projection on, or even reference to, the gross chromosomal architecture of the extant genomes, seem entirely novel. Moreover, the utilization of generalized adjacencies and gene family size restrictions is an innovation inspired by the particular case of plants, in the context of widespread and recurrent whole genome duplication (WGD) and fractionation. Although our restriction to safe adjacencies is the same as "informative for the ancestor of interest; i.e., the ancestor is on the pathway between both species in the phylogenetic tree" [4], our use of MWM to select globally optimum sets of adjacencies rather than a greedy approach, relying on adjacencies with the highest levels of attestation, is better suited to the botanical context.

3. Clustering

The hypothesis underlying chromosome recovery is that regions on DNA located near one another on the same chromosome are likely to be inherited (or "linked") together; thus, contigs originating from the same ancestral chromosome will have a good chance of appearing on the same chromosomes in the extant genomes descending from that ancestor. Even if the syntenic relationship of two contigs can be disrupted in one lineage by rearrangements such as translocation or chromosome fission, deletion or fractionation, the co-occurrence of the two may be conserved in other lineages. The frequency of co-occurrence of two contigs on the same chromosomes of the extant genomes is thus a good indication of whether these contigs appeared on the same chromosome in the ancestral genome.

We can therefore expect pairs of contigs remote from each other on an ancestral chromosome to have lower frequencies of co-occurrence than contigs closer together. There may be various other reductions in co-occurrence in some groups of species such as those due to different chromosomal arm locations, as in some insects, or other operon-like organizations. However, two contigs on different ancestral chromosomes should definitely have the least frequent co-occurrence in the extant genomes. Thus, to partition the contigs into distinct chromosomes while still allowing for some chromosome internal structure of co-occurrence data, it seems eminently reasonable to have recourse to hierarchical clustering procedures, such as average-linkage or complete-linkage [12].

This approach works well with some datasets, for example, the monocot reconstruction in [5], or the eudicot reconstruction in [13]. There are, however, some weaknesses in the procedure due to several factors.

1. Loss of evolutionary signal due to a lengthy time period between the ancestor and its descendants. This leads to a sparsity of co-occurrence values of non-negligible size, meaning that some contigs do not fit into any cluster at a meaningful level.
2. Scale bias. Large contigs will have more co-occurrences than smaller contigs that will be included late, often erroneously (especially with complete-linkage), in the clustering procedure.
3. Variable scores. Due to vagaries in deletion and other evolutionary processes, not all high scores reflect true ancestral co-occurrence. Coversely, some co-occurrences cannot be captured due to low scores.
4. Inflexible visualization settings. The heat maps color pixels by dividing the range of scores into equal intervals by default. However, this is not useful in comparing heat maps produced by different settings in the construction of contigs or in the use of different similarity or distance measures of contig co-occurrence. One heat map may be simply darker or lighter than the other overall, thus obscuring the real object of comparison, which is how clear-cut and distinct the clusters are and how they are qualitatively different from map areas not corresponding to clusters.

4. Updates to the Clustering

4.1. Update to the Co-Occurrence Measure

In this paper, we propose replacing raw co-occurrence frequencies with another measure of the likely common ancestral chromosome membership of two contigs x and y.

This follows the observation, on one hand, of many contig pairs showing low co-occurrence with each other but otherwise having an identical or similar pattern of co-occurrence frequencies with other contigs, while, on the other hand, some contig pairs show elevated co-occurrences, despite little similarity between their patterns of co-occurrence with other contigs. To eliminate these anomalies, we use the correlation r_{xy} between the co-occurrence frequencies of x and y with all the other contigs as a clustering criterion. Let $n = n_{\text{contigs}} - 2$, where n_{contigs} is the total number of contigs.

The effects of changing to the correlation measure are made clear in examining the familiar formula for Pearson's coefficient of correlation,

$$r_{xy} = \frac{n\sum x_i y_i - \sum x_i \sum y_i}{\sqrt{n\sum x_i^2 - (\sum x_i)^2}\sqrt{n\sum y_i^2 - (\sum y_i)^2}}. \quad (1)$$

The sums are taken over all n contigs i, where $i \neq x$ and $i \neq y$, x_i is the co-occurrence between x and i and y_i is the co-occurrence between y and i.

By applying Pearson's coefficient of correlation, the covariance of co-occurrence frequencies is normalized, and therefore the large variability of the data is mitigated. The scale bias is largely removed because multiplying all the x_i, for example, by the same constant in Equation (1) has no effect on r_{xy}.

4.2. Update to Heat Map Visualization

Instead of grouping data into "bins" of equal width in terms of the clustering criterion, we assign a preset proportion of pixels to each gray shade. This allows us to compare the clustering of the different ancestors, to assess the effects of changing the similarity measure as in Section 4.1, or to compare the analyses of real versus simulated data, without obscuring the effect of the variable ranges of similarity measures from one heat map to another. Table 1 shows the fixed proportion of pixels with its corresponding color shade in each bin. This particular set of proportions used throughout this study was set largely subjectively, seeking a clear perception of the difference between the clustered regions of the heat maps and the rest of the area while preserving the internal integrity of the clusters.

Table 1. The fixed proportion of pixels with its corresponding grayscale intensity in each data group shown in the heat maps.

Greyscale intensity	1	2	3	4	5	6	7	8	9
Proportion of pixels	50%	15%	10%	6%	4%	4%	4%	6.5%	0.5%

5. The Monocots

The RACCROCHE method has been applied to six genomes drawn from a broad range of monocot orders:

1. *Acorus calamus* (sweet flag) from the order Acorales;
2. *Spirodela polyrhiza* (duckweed) from the order Alismatales;
3. *Dioscorea rotundata* (yam) from the order Dioscorales;
4. *Asparagus officinalis* (asparagus) from the order Aspargales;
5. *Elaeis guineensis* (African oil palm) from the order Arecales;
6. *Ananas comosus* (pineapple) from the order Poales.

These species belong to lineages that diverged over 110 Mya. The phylogenetic relationship of the six extant species according to APG IV [14] is summarized in Figure 1, where ancestors (internal nodes) are labeled with numbers from 1 to 4 in chronological order, and the root is located between *Acorus* and Ancestor 1.

Almost all known flowering plant genomes have had at least one whole genome doubling or tripling event (WGD and WGT, respectively), with some often having several, in their lineages since the ancestral angiosperm. It is known that there were two WGDs

between Ancestor 1 and *Spirodela* and one WGD between each ancestor and its immediate extant genome(s) in the tree, except *Spirodela*. The ancestral genomes reconstructed in [5] also confirmed the tetraploidization event "tau" [15] in the stem lineage between the alismatids (Acorales and Alismatales) and the lilioids (Dioscorales and Aspargales).

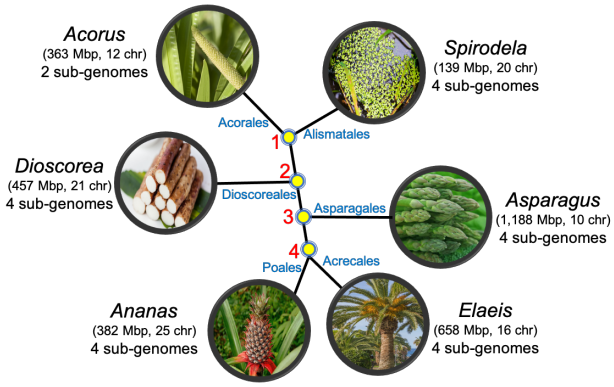
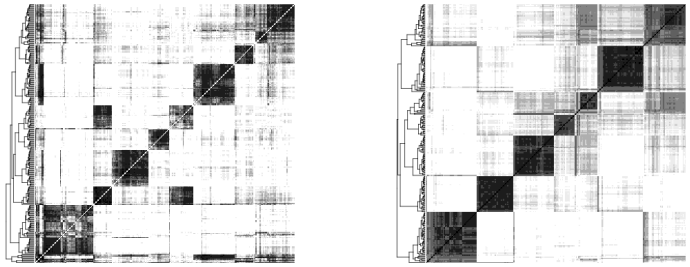


Figure 1. The phylogenetic relationship of the six extant monocot species according to APG IV is summarized in this phylogenetic tree, where ancestors (internal nodes) are labeled with numbers from 1 to 4. The root of this tree is between *Acorus* and Ancestor 1.

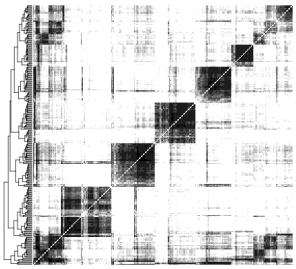
The long period of evolutionary divergence and the complexity of the recurrent cycle of WGD and fractionation affecting these genomes represent a challenge for any ancestral genome reconstruction method. RACCROCHE calculates several measures of the quality of its reconstructions, including statistics on contig length and coherence between successive ancestors, and indicators of the numbers of different types of chromosomal rearrangement the reconstruction implies.

The most telling evidence of the credibility of a reconstruction is the final chromosome structure it produces. The heat map visualization of RACCROCHE applied to the monocot data in Figure 2 shows a completely unambiguous clustering of the contigs into seven chromosomes of comparable size, for all four ancestors. Although this is gratifying, there are no genomes available from plants as old as these ancestors to verify the reconstruction.

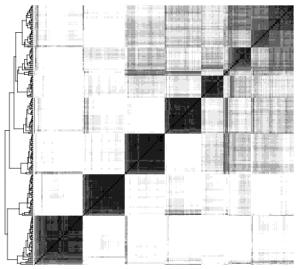


(**a**) Ancestor 1 co-occurrence (**b**) Ancestor 1 correlations

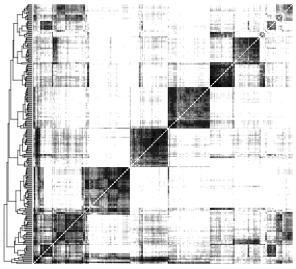
Figure 2. *Cont.*



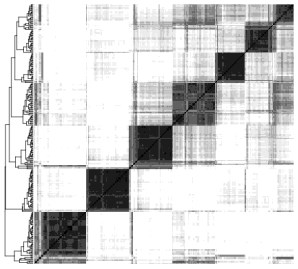
(**c**) Ancestor 2 co-occurrence



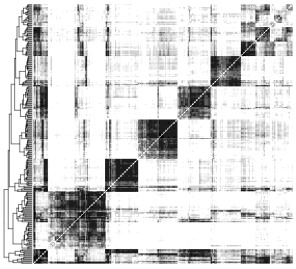
(**d**) Ancestor 2 correlations



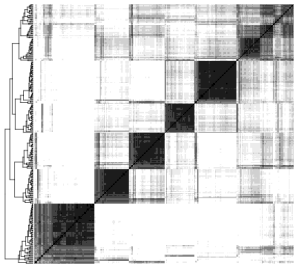
(**e**) Ancestor 3 co-occurrence



(**f**) Ancestor 3 correlations



(**g**) Ancestor 4 co-occurrence



(**h**) Ancestor 4 correlations

Figure 2. Heat maps of the four ancestors from the monocots data showing the clusters of contigs making up ancestral chromosomes from the longest 250 contigs, applying the complete-linkage clustering algorithm on chromosomal co-occurrence frequencies (**left**), and on correlations between co-occurrence vectors (**right**). The ordering of the contigs on the horizontal axis in each heat map is the same as that shown on the vertical axis.

Lacking ground truth about ancestral genomes dating from 100 Mya or more, one way we can assess the quality of reconstruction is through simulation, as discussed in Section 6. Thus, the main contribution from the current paper is a close simulation of monocot evolution, followed by an application of RACCROCHE to the simulated genomes, and a comparison of the results of the simulation input with the ancestor output by RACCROCHE.

6. Simulations

The goal of our simulation is to test RACCROCHE by using it to reconstruct a randomized input ancestral genome, based on data from simulated genomes as similar as possible to the

real extant genomes, in terms of the number of chromosomes, the number of gene families retained or lost across the sample of genomes, the number of genes in these families and the number of translocations and insertions affecting the extant and ancestral genomes, but not the gene order. To the extent the reconstruction of the simulated ancestor is accurate, we can have a high degree of confidence in the reconstruction of the real ancestral genome.

Our simulation study has several components. Central is the actual procedure for generating the evolution of the entire set of gene order genomes corresponding to the real extant genomes, but there are a number of preparatory aspects. We first have to characterize ancestral genomes statistically, that is, to determine the gene families that are present in each ancestral genome, something that has already been conducted during the application of RACCROCHE to the real genomes using a phylogenetic validity criterion (Section 2, item 3). Then, we need to estimate how many genes are in each gene family in Ancestor 1 and the other ancestors. Finally, after the simulations are carried out, we have to evaluate how successful RACCROCHE has been, particularly in recovering the artificial Ancestor 1 at the origin of the simulation.

6.1. Parameters for Simulations

The simulation proceeds along the branches of the evolutionary tree with the given topology and the known occurrences of whole genome doubling. In addition, we are given the number of chromosomes in each extant genome and statistics on the gene families, with J families in total, that were used in the reconstruction of ancestral genomes. For each family j, $M(h, j)$ is the number of genes in the family for each of the extant genomes, where h indexes the six genomes as in Section 5. Note that the identities of the gene families, compiled using SynMap [6,7] as described in [5], are retained across all the genomes. The number of genes per gene family is usually 0, 1 or 2 per genome, though some families have more. There are no families with more than 10 genes in any particular genome or more than 50 for all six genomes; such families are excluded at the outset, from both the RACCROCHE analysis and the simulation, as they would tend to degrade the reconstruction. No information on chromosome content, gene order or gene adjacencies is input to the simulation.

The numbers of inversions ($\sim$150) and translocations ($\sim$50) were estimated from RACCROCHE to have affected each extant genome [5]. These numbers are included as parameters in the simulation.

Not all gene families are present in all of the extant genomes. We can deduce which gene families are present in each of the ancestral chromosomes through the phylogenetic validation criterion described in [4,5], to wit, that a gene family is present in an internal node of the tree if and only if it is present in some terminal node in at least two of the three subtrees, ancestral or descendant, subtended by that node (cf. Section 2, item 3).

The parameters that must be fixed for use in the simulation include four non-negative cost parameters a, b, c and d, and a matrix $N(I, J)$ of estimated gene frequencies, where I is the number of ancestors and J is the total number of gene families with representatives in at least two genomes. Algorithm 2 estimates the optimal gene family sizes to be used in the simulation.

The topological structure of the phylogenetic tree is input through the arrays **anc** and **anct**, which define the ancestor–descendant relations. In the monocot example, $\mathbf{anc} = \begin{pmatrix} 1 & 1 & 2 & 3 & 4 & 4 \end{pmatrix}$ summarizes the ancestor–extant descendant relations in the tree, while $\mathbf{anct} = \begin{pmatrix} 1 & 1 & 2 & 3 \end{pmatrix}$ summarizes the ancestry relations among the ancestors.

Ploidy changes along the branches of the tree are counted using the input parameters r_h and r_k. These parameters are integers that indicate the number of WGDs between an ancestor and its descendant h or k, for extant genomes and ancestors, respectively. For example, $r_h = 1, 2$ or 4 indicates that there are zero, one or two WGDs from an ancestor to its descendant h.

The simulation is very dependent on its starting point, namely, the chromosomes of Ancestor 1, and their gene content. True, RACCROCHE reconstructs a version of Ancestor 1,

Algorithm 2: Estimate optimal gene family sizes in simulated ancestral genomes.

input : anc and $anct$, arrays of integers defining the ancestor-descendant relations in Tr, a fully labelled binary branching phylogeny
r_h and r_k, integers representing the number of WGD on the branch leading to extant genome h or ancestor genome k, respectively
$M(H, J)$, matrix of integers representing the number of genes in each gene family in extant genomes calculated from real data, where H is the number of extant genomes, J is the total number of gene families
output: $N(I \times J)$, matrix of integers representing the number of genes in each gene family in ancestral genomes, where I is the number of ancestors in Tr, J is the total number of gene families in each ancestor

initialize $N(1 \times J) \leftarrow 1$ and $N(k \times J) \leftarrow r_k N((k-1) \times J)$ for all J gene families;
initialize cost penalties a, b, c, d;
minimize $\mathbf{cost} \leftarrow \sum_{j=1}^{J}(\sum_{h=1}^{H} \mathbf{cost_1}(h, j) + \sum_{k=1}^{K} \mathbf{cost_2}(k, j))$ to obtain optimal a, b, c, d, where $\mathbf{cost_1}$ and $\mathbf{cost_2}$ are defined in Equations (2) and 3 respectively;
for *gene family* $j \leftarrow 1$ **to** J **do**
    minimize $\mathbf{cost_j} \leftarrow \sum_{h=1}^{H} \mathbf{cost_1}(h, j) + \sum_{k=1}^{K} \mathbf{cost_2}(k, j)$ to obtain $N(i, j)$ using nonlinear programming with respect to $N(k, j)$ for ancestor k and optimal a, b, c, d;
end
return $N(I \times J)$, array of integers representing gene family sizes in the ancestors;

but its basis in the contigs constructed by MWM means that each gene family occurs at most once. However, real gene family sizes vary within a genome, and the distribution of these sizes on Ancestor 1 will have an important influence on the simulation of the set of extant genomes.

To simulate Ancestor 1, we thus try to associate gene family sizes for each of the genes determined by the RACCROCHE reconstruction. For this, we make use of the distributions of family sizes in the extant genomes.

We define two cost matrices, $\mathbf{cost_1}(H \times J)$ for extant genomes and $\mathbf{cost_2}(I \times J)$ for ancestral genomes, that help us determine the gene family sizes of the ancestral chromosomes, essential for ensuring our simulations mirror as closely as possible the evolutionary processes we inferred for the real data.

In considering a gene family i, we define a quantity Δ that measures the difference between the number of genes in it generated by zero, one or more WGDs, $r_h \times N(anc(h), j)$ and the number we previously posited for this family $M(h, j)$ in the case of extant genomes, or $r_k \times N(anct(k), j)$ and $N(k, j)$ in the case of ancestral genomes. Then, depending on whether Δ represents a gene loss or gain, we assign a specific cost, a, b, c or d. We then minimize the total cost to optimize the cost parameters, and then to optimize the $N(k, j)$.

$$\mathbf{cost_1}(h, j) = \begin{cases} a(1 + \log_2(-\Delta)) & \text{if } \Delta < 0, \text{(cost of generating too many genes)}, \\ b(1 + \log_2 \Delta) & \text{if } \Delta > 0, \text{(cost of generating too few genes)}, \\ 0 & \text{if } \Delta = 0, \end{cases} \quad (2)$$

where $\Delta \leftarrow M(h, j) - r_h \times N(anc(h), j)$.

$$\mathbf{cost_2}(k, j) = \begin{cases} c(1 + \log_2(-\Delta)) & \text{if } \Delta < 0, \text{(cost of generating too many genes)}, \\ d(1 + \log_2 \Delta) & \text{if } \Delta > 0, \text{(cost of generating too few genes)}, \\ 0 & \text{if } \Delta = 0. \end{cases} \quad (3)$$

where $\Delta \leftarrow N(k, j) - r_k \times N(anct(k), j)$.

In Algorithm 2, by minimizing the overall cost matrix,

$$\mathbf{cost} = \Sigma_{j=1}^{J}\left(\Sigma_{h=1}^{H}\mathbf{cost_1}(h,j) + \Sigma_{k=1}^{K}\mathbf{cost_2}(k,j)\right),$$

with respect to a, b, c, d using nonlinear programming, the optimal non-negative cost parameters are estimated. In the monocot example, they are $a = 0$, $b = 1$, $c = 4$, $d = 4$.

For each gene family j, minimize the same **cost** as a function of $N(I \times J)$ with fixed values of a, b, c, d, using nonlinear programming. This estimates the number of genes in ancestors for each gene family.

6.2. The Simulation Process

The simulation process is formalized in Algorithm 3 utilizing the parameters estimated from Algorithm 2.

Algorithm 3: The simulation of gene repertoire in extant genomes

input : Tr, an fully labelled binary branching phylogeny
$G_{1\cdots H}$, annotated extant gene-order genomes related by Tr
$N(I \times J)$, matrix of integers representing gene family size in ancestral genomes estimated from Algorithm 2, where I is the number of ancestors in Tr, J is the total number of gene families in each ancestor
output: $G'_{1\cdots H}$, simulated gene-order genomes related by Tr

```
1  for ancestor i ← 1 to I do
2      if i == 1 then
3          initialize ancestor 1 with genes in N(1 × J) randomly distributed in 7
             chromosomes;
4      else
5          construct ancestor i by doubling or equal N((i − 1) × J) according to the
             whole genome duplication event related to ancestor i;
6          doing translocations and inversions in ancestor i;
7      end
8      adjust the number of gene families in each ancestor by inserting/removing
         families at random positions;
9      foreach genome g connected with ancestor i in Tr do
10         construct g by doubling, tripling or quadrupling ancestor i according to the
             whole genome duplication event that relates g to ancestor i in Tr;
11         doing translocations and inversions in each g;
12         if g is a terminal node in Tr then
13             adjust g by inserting/removing genes or fission/fusion chromosomes
                 so that gene and chromosome contents in g is consistent with its
                 corresponding extant genome G_h in Tr;
14             append g to G′;
15         end
16     end
17 end
18 return G′;
```

As considered in the monocot example in this paper, the simulation starts with Ancestor 1, made up of abstract genes belonging to specified gene families j in numbers $N(1, j)$ determined previously. These genes are ordered randomly on seven chromosomes of approximately equal size as in Line 3 of Algorithm 3.

In Line 5, each of the remaining ancestors is generated by doubling or equaling its previous ancestor.

In Line 6, each of these ancestral genomes is then subjected to inversions and translocations randomly in numbers previously calculated [5].

In Line 8, in each ancestor, missing gene families are added randomly and extra families are removed according to $\Delta = M(h, j) - r_h \times N(anc(h), j)$ in Equation (3). If Δ is negative, remove genes from the gene family; otherwise, add genes into the family.

In Line 10, to obtain the three simulated descendants of Ancestor 1, namely, *Acorus*, *Spirodela* and Ancestor 2, the genome of Ancestor 1 is doubled, quadrupled and doubled, respectively.

In Line 11, each extant descendant genome is then subjected to inversions and translocations in numbers previously calculated [5].

In Line 13, in each gene family of each descendant, the number of genes generated by the whole genome doublings of its immediate ancestor is compared to the number of genes known to be in that family in that genome. Missing genes are simply added to the simulated genome, inserted at random on one of the chromosomes. Extra genes are just deleted from the gene family from random positions in the genome.

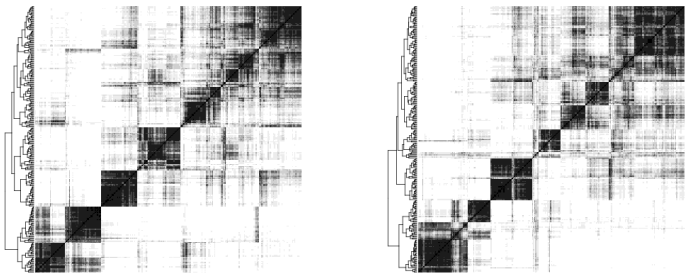
The number of chromosomes in each descendant is then adjusted by fusions of the shortest chromosomes to form a new longer one.

The simulation continues in the same manner for the descendants of Ancestor 2, and then Ancestors 3 and 4.

We thus create a set of six simulated genomes with the same gene families distributed in the same way as in the given monocots. The only difference is that the gene orders are completely random with respect to the real data, although we have retained the quantitative structure of the gene families.

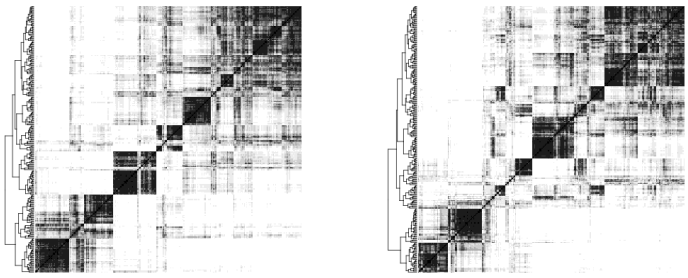
7. Results

The output of one simulation is summarized by the heat maps in Figure 3. A clear clustering pattern is evident, reminiscent of that obtained for the real data. There is somewhat more noise, suggesting that our simulation is more disruptive to gene order than is natural evolution, either because of the biases in our rearrangement parameters, the order in which we carried out the evolution from ancestor to descendant or some other factor.



(**a**) Simulated Ancestor 1 (**b**) Simulated Ancestor 2

Figure 3. *Cont.*



(**c**) Simulated Ancestor 3 (**d**) Simulated Ancestor 4

Figure 3. Heat maps of the 4 ancestors from simulated data showing the clusters of contigs making up ancestral chromosomes from the longest 250 contigs, using complete-linkage on chromosomal co-occurrence correlations.

Nevertheless, as shown in Figure 4, the reconstruction recovers the original ancestors very well, with every chromosome clearly deriving from one of the initial chromosomes.

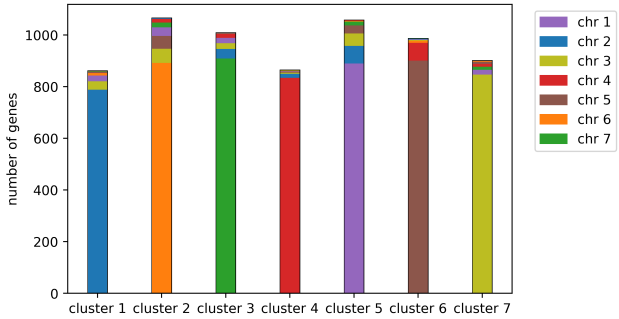
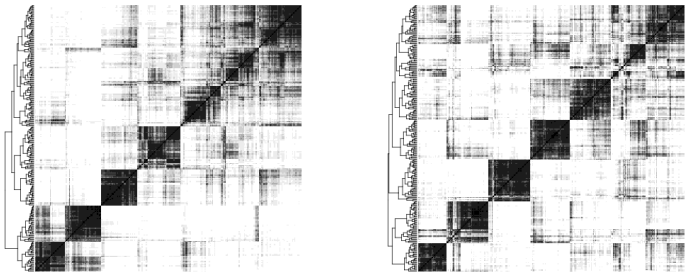


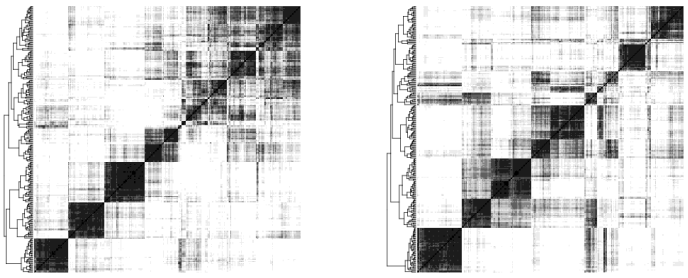
Figure 4. Projection of simulated ancestral chromosomes (colored bars) on version reconstructed by `RACCROCHE` (outlined bars).

To ensure that the results are not artifacts of the particular random numbers used to initiate the simulations, three additional replications are carried out. The results in Figure 5 demonstrate that `RACCROCHE` with the improved clustering method yields consistent results in reconstructing ancestral chromosomes based on simulated data.



(**a**) Simulation 1 Ancestor 1 (**b**) Simulation 2 Ancestor 1

Figure 5. *Cont.*



(**c**) Simulation 3 Ancestor 1 (**d**) Simulation 4 Ancestor 1

Figure 5. Heat maps for Ancestor 1 from four different simulations showing the clusters of contigs making up ancestral chromosomes from the longest 250 contigs, using complete-linkage on chromosomal co-occurrence correlations.

8. Conclusions

As a simple contribution of this work, the fixed proportion of gray shadings on the heat map could have wider applications beyond visualizing ancestral chromosome clustering. We have already used it here in comparing co-occurrence measures, for comparing analyses of real versus simulated data, for comparisons among the four ancestors and for showing parallels between replications of the simulation. There are other possibilities, even within our particular application to reconstruction. For example, were we to use 200 contigs or 400 contigs instead of 250, the improved heat maps could avoid the effect of the contig number on the overall impression, since the average brightness or darkness would not change.

In clustering the contigs using complete-linkage, we hoped to ensure that all the contigs in a cluster are related at least at some minimal level. A disadvantage might be a slight tendency for clusters to be broken up by some fortuitous link early in the execution of the algorithm due to the nature of greedy algorithms. We have not found superior results with other clustering methods such as average-linkage or k-means.

In estimating the number of genes in gene families in ancestral genomes, $N(i, j)$, we did not explore the possibility of iterating the successive estimation of $N(i, j)$ and a, b, c, d. Adopting this approach might decrease the overall noise in the heat map.

One interesting aspect of our reconstruction of the ancestral genomes is that Ancestors 2, 3 and 4 all displayed seven clusters despite the whole genome duplication between Ancestor 1 and Ancestor 2, both in the real data and the simulations. An explanation in terms of selective pressures towards a seven-chromosome state may have some credibility, though a methodological artifact seems more likely; the two subgenomes created by whole genome duplication would be very similar, meaning that co-occurrence patterns of contigs containing runs of homologous genes would be parallel, and separate chromosomes would not emerge from the clustering. This could be an objective of future study.

To what extent would our methods be applicable to sets of genomes even more distantly related than the monocot orders? On the one hand, the reconstructed contigs would be shorter, the co-occurrence frequencies lesser and the clustering less clear. On the other hand, the intermediate ancestors should be more distinct, meaning that we might better distinguish the evolutionary development of the extant genomes.

Author Contributions: Conceptualization, Q.X., L.J., J.L.-M., D.S.; methodology, Q.X., L. J., D.S.; software, Q.X., L. J.; writing, Q.X., L. J., D.S.; funding acquisition, D.S., L.J. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by Discovery grants to L.J. and D.S. from the Natural Sciences and Engineering Research Council of Canada. D.S. holds the Canada Research Chair in Mathematical Genomics.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: The annotated genomic data are accessible on the `CoGe` platform (https://genomevolution.org/coge/) and Phytozome (https://phytozome.jgi.doe.gov/pz/portal.html). The pipeline is available at https://github.com/jin-repo/RACCROCHE.

Acknowledgments: We thank the Department of Energy Joint Genome Institute staff and collaborators including David Kudrna, Jerry Jenkins, Jane Grimwood, Shengqiang Shu and Jeremy Schmutz for pre-publication access to the *Acorus* genome sequence and annotation.

Conflicts of Interest: The authors declare no conflict of interest.

Abbreviations

The following abbreviations are used in this manuscript:

MWM	Maximum weight matching
Mya	Million years ago
WGD	Whole genome duplication
WGT	Whole genome triplication

References

1. Perrin, A.; Varré, J.S.; Blanquart, S.; Ouangraoua, A. ProCARs: Progressive reconstruction of ancestral gene orders. *BMC Genom.* **2015**, *16*, S6.
2. Rubert, D.P.; Martinez, F.V.; Stoye, J.; Doerr, D. Analysis of local genome rearrangement improves resolution of ancestral genomic maps in plants. *BMC Genom.* **2020**, *21*, 1–11.
3. Badouin, H.; Gouzy, J.; Grassa, C.J.; Murat, F.; Staton, S.E.; Cottret, L.; Lelandais-Brière, C.; Owens, G.L.; Carrère, S.; Mayjonade, B.; et al. The sunflower genome provides insights into oil metabolism, flowering and Asterid evolution. *Nature* **2017**, *546*, 148–152.
4. Berthelot, C.; Muffato, M.; Abecassis, J.; Crollius, H. The 3D organization of chromatin explains evolutionary fragile genomic regions. *Cell Rep.* **2015**, *10*, 1913–1924.
5. Xu, Q.; Jin, L.; Zheng, C.; Leebens-Mack, J.H.; Sankoff, D. `RACCROCHE`: Ancestral flowering plant chromosomes and gene orders based on generalized adjacencies and chromosomal gene co-occurrences. Proceedings of the 10th International Conference on Computational Advances in Bio and Medical Sciences, Virtual, Dec. 10-12, 2020. In Springer *Lecture Notes in Bioinformatics*; Jha, S.K., Mandoiu, I., Rajasekaran, S., Sahni, S., Skums, P., Zelikovsky, A., Eds.; Volume 12686.
6. Lyons, E.; Freeling, M. How to usefully compare homologous plant genes and chromosomes as DNA sequences. *Plant J.* **2008**, *53*, 661–673.
7. Lyons, E.; Pedersen, B.; Kane, J.; Freeling, M. The value of nonmodel genomes and an example using SynMap within CoGe to dissect the hexaploidy that predates rosids. *Trop. Plant Biol.* **2008**, *1*, 181–190.
8. Yang, Z.; Sankoff, D. Natural parameter values for generalized gene adjacency. *J. Comput. Biol.* **2010**, *17*, 1113–1128.
9. Xu, X.; Sankoff, D. Tests for gene clusters satisfying the generalized adjacency criterion. In Proceedings of the Brazilian Symposium on Bioinformatics, Santo André, Brazil, 28–30 August 2008; Springer: Berlin/Heidelberg, Germany, 2008; pp. 152–160.
10. Tannier, E.; Bazin, A.; Davín, A.; Guéguen, L.; Bérard, S.; Chauve, C. Ancestral genome organization as a diagnosis tool for phylogenomics. In *Phylogenetics in the Genomic Era*; Scornavacca, C., Delsuc, F., Galtier, N., Eds.; No commercial publisher | Authors open access book, 2020; pp. 2.5:1–2.5:19.
11. Zheng, C.; Chen, E.; Albert, V.A.; Lyons, E.; Sankoff, D. Ancient eudicot hexaploidy meets ancestral eurosid gene order. *BMC Genom.* **2013**, *14*, 1–13.
12. Shalev-Shwartz, S.; Ben-David, S. *Understanding Machine Learning: From Theory to Algorithms*; Cambridge University Press: Cambridge, UK, 2014.
13. Chanderbali, A.; et al. *Buxus* and *Tetracentron* genomes help resolve eudicot phylogeny, gamma hexaploidy, and paleogenomics. **2021**. In preparation.
14. Chase, M.W.; Christenhusz, M.; Fay, M.; Byng, J.; Judd, W.S.; Soltis, D.; Mabberley, D.; Sennikov, A.; Soltis, P.S.; Stevens, P.F. An update of the Angiosperm Phylogeny Group classification for the orders and families of flowering plants: APG IV. *Bot. J. Linn. Soc.* **2016**, *181*, 1–20.
15. Jiao, Y.; Li, J.; Tang, H.; Paterson, A.H. Integrated syntenic and phylogenomic analyses reveal an ancient genome duplication in monocots. *Plant Cell* **2015**, *26*, 2792–2802.

Chapter 4

Buxus and *Tetracentron*

Andre S. Chanderbali, Lingling Jin, Qiaoji Xu, Yue Zhang, Jingbo Zhang, Shuguang Jian, Emily Carroll, David Sankoff, Victor A. Albert, Dianella G. Howarth, Douglas E. Soltis and Pamela S. Soltis. Buxus and Tetracentron genomes help resolve eudicot genome history. Nat Commun, 13:643, 2022.

nature COMMUNICATIONS

ARTICLE

https://doi.org/10.1038/s41467-022-28312-w OPEN

Buxus and *Tetracentron* genomes help resolve eudicot genome history

Andre S. Chanderbali[1✉], Lingling Jin[2], Qiaoji Xu[3], Yue Zhang[3], Jingbo Zhang[4], Shuguang Jian[5], Emily Carroll[6], David Sankoff[3], Victor A. Albert[6], Dianella G. Howarth[4], Douglas E. Soltis[1,7,8,9] & Pamela S. Soltis[1,8,9]

Ancient whole-genome duplications (WGDs) characterize many large angiosperm lineages, including angiosperms themselves. Prominently, the core eudicot lineage accommodates 70% of all angiosperms and shares ancestral hexaploidy, termed *gamma*. *Gamma* arose via two WGDs that occurred early in eudicot history; however, the relative timing of these is unclear, largely due to the lack of high-quality genomes among early-diverging eudicots. Here, we provide complete genomes for *Buxus sinica* (Buxales) and *Tetracentron sinense* (Trochodendrales), representing the lineages most closely related to core eudicots. We show that *Buxus* and *Tetracentron* are both characterized by independent WGDs, resolve relationships among early-diverging eudicots and their respective genomes, and use the RACCROCHE pipeline to reconstruct ancestral genome structure at three key phylogenetic nodes of eudicot diversification. Our reconstructions indicate genome structure remained relatively stable during early eudicot diversification, and reject hypotheses of *gamma* arising via inter-lineage hybridization between ancestral eudicot lineages, involving, instead, only stem lineage core eudicot ancestors.

[1] Florida Museum of Natural History, University of Florida, Gainesville, FL, USA. [2] Department of Computer Science, University of Saskatchewan, Saskatoon, SK, Canada. [3] Department of Mathematics and Statistics, University of Ottawa, Ottawa, ON, Canada. [4] Department of Biological Sciences, St. John's University, Queens, NY, USA. [5] South China Botanical Garden, Chinese Academy of Sciences, Guangzhou, China. [6] Department of Biological Sciences, University at Buffalo, Buffalo, NY, USA. [7] Department of Biology, University of Florida, Gainesville, FL, USA. [8] Biodiversity Institute, University of Florida, Gainesville, FL, USA. [9] Genetics Institute, University of Florida, Gainesville, FL, USA. ✉email: achander@ufl.edu

Flowering plants (angiosperms), with nearly 400,000 species and a fossil record that dates to the Early Cretaceous, have a complex evolutionary history marked by early and rapid lineage divergences[1–3]. Whole-genome duplication (WGD) events have also been frequent in angiosperms, and indeed all extant species are ancient polyploids descended from a common ancestor that experienced at least one WGD[4,5]. Subsequent polyploidy events have been identified throughout angiosperm phylogeny, often coinciding closely with the origin and/or radiation of major clades[6–10]. Notably, the core eudicots (*Gunneridae*[11]), nested in the eudicot clade, descend from an ancient hexaploid formation, termed *gamma*[12–15], and account for ~70% of extant angiosperm species. Moreover, a novel suite of floral features, 'whorled pentamery' with flower parts arranged in concentric whorls of five[16–18], evolved shortly after the origin of the core eudicots[11,19] and could be genetically linked to this ancient hexaploidy event, e.g., through multiplications or rearrangements of floral transcriptional regulators[15]. Such a causal relationship between *gamma* and whorled pentamery, although still speculative, is consistent with the widely acknowledged role of gene and genome duplications providing the genetic raw material for evolutionary innovation[9,20].

The phylogenetic timing and mechanism of *gamma* hexaploidy are currently unresolved. Hypotheses on the topic mostly envision a two-step process, in which the product of an initial WGD fused with a third genome in a second polyploidization, possibly via a wide cross after an extended period of random fractionation (loss of either copy of duplicated genomic regions following WGD) in the tetraploid intermediate[21]. The breadth of this putative wide cross is also unclear and possibly includes extant early-diverging eudicot lineages[13,15,22]. Alternatively, one of the *gamma* subgenomes may have been more resistant to fractionation, and all three subgenomes may have been joined rapidly in evolutionary time[21], perhaps in an autohexaploidy event[23]. It has also been argued that *gamma* hexaploidy derives from an initial tetraploidy shared by all eudicots[24,25]. Further still, the lack of clear evidence of *gamma* outside of the core eudicots may be due to stochastic gene loss over more than 100 million years of independent evolution[23]. Efforts to evaluate evolutionary scenarios of *gamma* origins have been hampered by limited data and unsettled sister-group relationships to the core eudicots. Plastome sequence data support either Buxales[19,26,27] or Trochodendrales[28,29] as immediate sisters to the core eudicots, while single-copy nuclear (SCN) genes from transcriptome data sets have recovered a Buxales+Trochodendrales clade placed sister to the core eudicots[15,30]. Thus, despite considerable research interest, the timing and mechanism of *gamma* formation have remained unresolved.

We here provide genome assemblies for *Buxus sinica* (Buxales) and *Tetracentron sinense* (Trochodendrales), which represent, either individually or collectively, the sister lineage of core eudicots[15]. These two genome assemblies complement those available for other early-diverging eudicot lineages[22,31–33] and permit evaluations of eudicot phylogeny and *gamma* origins based on phylogenomics, molecular evolution, and synteny. In addition, we employ the RACCROCHE[34] pipeline of algorithms to infer the ancestral genomes at three sequential nodes of the eudicot radiation.

Results and discussion

Genome assembly, annotation, and structure. Chromosome-scale nuclear genome assemblies for *Buxus* and *Tetracentron* were produced from PacBio long-read contigs assembled with the FALCON/FALCON-unzip pipeline[35] and scaffolded by Hi-C technology[36] (Fig. 1; Supplementary Data 1). The *Buxus* assembly totals 764 Mb (90% of the estimated genome size of 850 Mb), with 7180 contigs (N50 = 164 kb) in 63 scaffolds (N50 = 56 Mb), of which 14 contain 763 Mb (99.8%) of the assembly. The *Tetracentron* assembly totals 908 Mb (93% of the estimated genome size of 975 Mb), with 6178 contigs (N50 = 238 kb) in 662 scaffolds (N50 = 54 Mb), of which 19 contain 856 Mb (94.5%) of the assembly. The largest 14 and 19 scaffolds of the *Buxus* and *Tetracentron* assemblies, respectively, correspond with the known chromosome numbers of these taxa[37,38]. Benchmarking Universal Single-Copy Orthologs (BUSCO) analyses[39,40] estimate 96.3% and 93.5% completeness for the *Buxus* and *Tetracentron* genomes, respectively (Supplementary Data 2). Transposable elements and other repeat sequences account for 76.4% and 78.5% of the *Buxus* and *Tetracentron* assemblies, respectively (Supplementary Data 3). In *Buxus*, LTR retrotransposons (26.8%), followed by LINEs (4.9%) and DNA transposable elements (2.8%), are most abundant, with Ty3/*Gypsy* and Ty1/*Copia* retrotransposons accounting for 87.2% and 13.0% of the LTRs, respectively. LTRs (27.4%), LINES (4.6%), and DNA transposable elements (2.9%) account for most of the *Tetracentron* repeats, with Ty3-*Gypsy* (62.6%) and Ty1/*Copia* (36.6%) retrotransposons best represented among the LTRs. Annotation of the repeat-masked assemblies yielded 27,027 and 30,704 protein-coding gene models, including 86.9% and 80.5% of the BUSCO genes, in *Buxus* and *Tetracentron*, respectively (Supplementary Data 2). Our *Tetracentron* assembly is similar to one produced for another individual of this species[33] in terms of BUSCO statistics and annotation metrics, but differs in size (908 vs 1170 Mb) and the number of chromosome-size scaffolds (19 vs 24). We are unable to account for these differences, but our assembly closely matches the genome size measured by flow cytometry, and the only reported chromosome count of $n = 24$[41] for *Tetracentron* has been discredited[37].

Analyses of synonymous changes per synonymous site (*Ks*) and intragenomic synteny indicate that *Buxus* and *Tetracentron* are both paleopolyploids, with one and two rounds of WGDs in their respective evolutionary histories. *Buxus* syntenic paralogs (paleologs) constitute extensive blocks of colinear genome sequence across pairs of chromosomes and are characterized by *Ks* values close to 1.0 (Fig. 1c). *Ks* values for *Tetracentron* paleologs are concentrated near $Ks = 0.5$, but colinear genome sequences are distributed among four chromosomes (Fig. 1d), together suggesting two WGDs in close succession. The two *Buxus* subgenomes are highly conserved, with synteny blocks that often extend across much of the whole chromosomes, while the four subgenomes of *Tetracentron* appear to be highly rearranged at the chromosomal level (Fig. 1). The extent to which this structure reflects genome reshuffling, which is a prominent mechanism of post-polyploid diploidization (PPD) after WGDs[42], or artifacts of genome assembly, is unclear. In favor of PPD processes, the *Tetracentron* genome is appreciably downsized compared to its sister species, and the only other living member of Trochodendraceae, *Trochodendron aralioides* (0.9 versus 1.6 GB), which shares two WGDs with *Tetracentron*[33] but exhibits more extensive blocks of inter-chromosomal synteny (Supplementary Fig. 1).

Phylogenetic positions of *Buxus* and *Tetracentron*. To reconstruct the branching sequence of the early eudicot radiation, we analyzed phylogenetic data sets for representative angiosperms composed of hundreds of BUSCO genes[43], the Angiosperms353 loci[44], and orthogroups identified de novo by the Orthofinder pipeline[45]. Coalescence-based analyses of all three data sets place Ranunculales as sister to all other living eudicot lineages, with Proteales (including Sabiaceae) diverging next, and a Buxales +Trochodendrales clade as sister to the core eudicot clade (Fig. 2a;

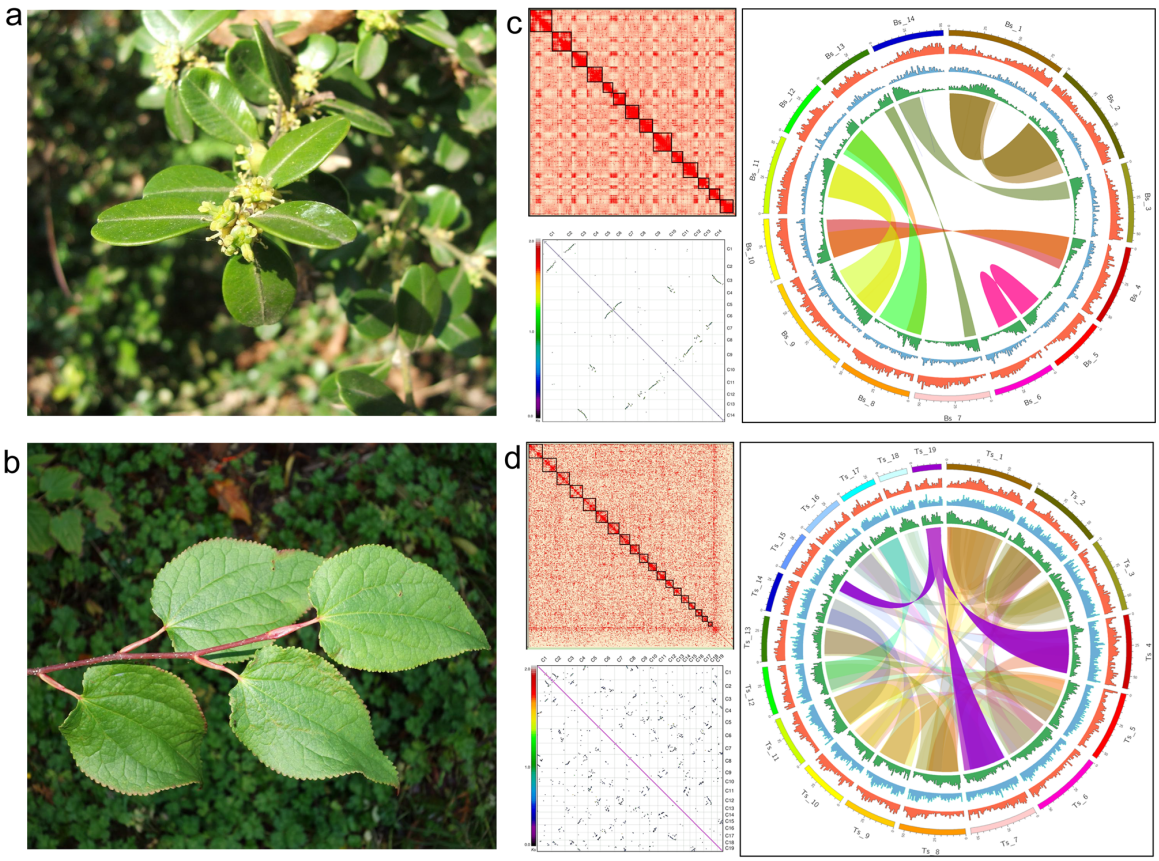


Fig. 1 Habit and genome assembly features of *Buxus* and *Tetracentron*. Flowering branch of *Buxus sinica* (**a**) courtesy of PiPi; and leafy shoot of *Tetracentron sinense* (**b**) courtesy of Daderot. Hi-C contact heatmaps, intragenomic synteny with syntenic blocks colored according to the *Ks* scale, and Circos plots for *Buxus* (**c**) and *Tetracentron* (**d**). Concentric tracks in the Circos plots, from innermost outwards, show gene, *Copia*, and *Gypsy* retrotransposon densities per 1 Mb, and chromosomes, while ribbons connect inter-chromosomal syntenic regions. Source data underlying Fig. 1c, d are provided as a Source data file.

left panel, Supplementary Figs. 2a and 3). Concatenated data sets of the SCN loci, whether analyzed in Maximum Likelihood (Fig. 2a, right panel, Supplementary Fig. 2b) or Bayesian Inference (Supplementary Fig. 4) frameworks, recover Buxales alone as the core eudicot sister group, with Trochodendrales as sister to this Buxales +core eudicot clade. Although this branching sequence receives maximal statistical support in both Maximum Likelihood (bootstrap) and Bayesian Inference (posterior probability) analyses, incomplete lineage sorting (ILS) is a potential confounding factor in phylogenetic analyses of concatenated data sets in the face of rapid radiations[46], as is the case for the eudicots. Indeed, the quartet-support values associated with the Buxales+Trochodendrales clade in the coalescence tree indicate considerable gene tree discordance with respect to the positions of these taxa. Further exploration of conflicts affecting the eudicot clade, visualized as a cloudogram of gene trees (Fig. 2b), however, reveals that ~30% of the gene trees support the Buxales+Trochodendrales clade, while only ~18% support either Buxales or Trochodendrales as the core eudicot sister group (Supplementary Data 4). We also estimated the branching sequence of early-diverging eudicots using the 'Trees in the Peaks' method, which reconstructs speciation and polyploidization events from *Ks* and similarity score distributions of syntenic homologs[47,48] (Fig. 2c). This method, which requires that ancestral *Ks* and similarity scores and/or their ranges must precede (greater *Ks* or lower similarity) or overlap those in the descendants, was applied to evaluate each of all possible binary rooted phylogenies. The only branching sequence that satisfies these conditions is one in which Buxales and Trochodendrales are collectively sister to the core eudicots. Specifically, the peak *Ks* value of syntenic orthologs that diverged via the *Buxus/Tetracentron* speciation is younger than those derived from the phylogenetic divergence of *Vitis* (a core eudicot) from *Buxus* or from *Tetracentron*.

Phylogenomics of eudicot subgenomes. Synteny-guided phylogenomic analyses of eudicot subgenomes were conducted to assess the several hypothesized scenarios for the origin of *gamma* hexaploidy (Fig. 3). Pairwise analyses of inter-genomic collinearity (macrosynteny) and fractionation patterns identify extensive regions of early-diverging eudicot genomes shared with the *gamma*-derived hexaploid genome of *Vitis*, and each other (Supplementary Figs. 5–9). The ratios of syntenic depths (the number of times a genomic region is syntenic to regions in another genome) in these comparisons reflect the number of subgenomes, or level of ploidy, for the respective species. Thus, we see 2:3 syntenic depth between *Buxus* and *Vitis*, and 4:3 syntenic depth between *Tetracentron* and *Vitis*, while *Tetracentron* to *Buxus* is 4:2 in syntenic depth. Likewise, as previously reported, *Aquilegia* and *Nelumbo* each exhibit 2:3 syntenic depth with *Vitis*, and 2:2 with each other. Collectively, these macrosyntenic alignments approximate the modern distribution of the seven

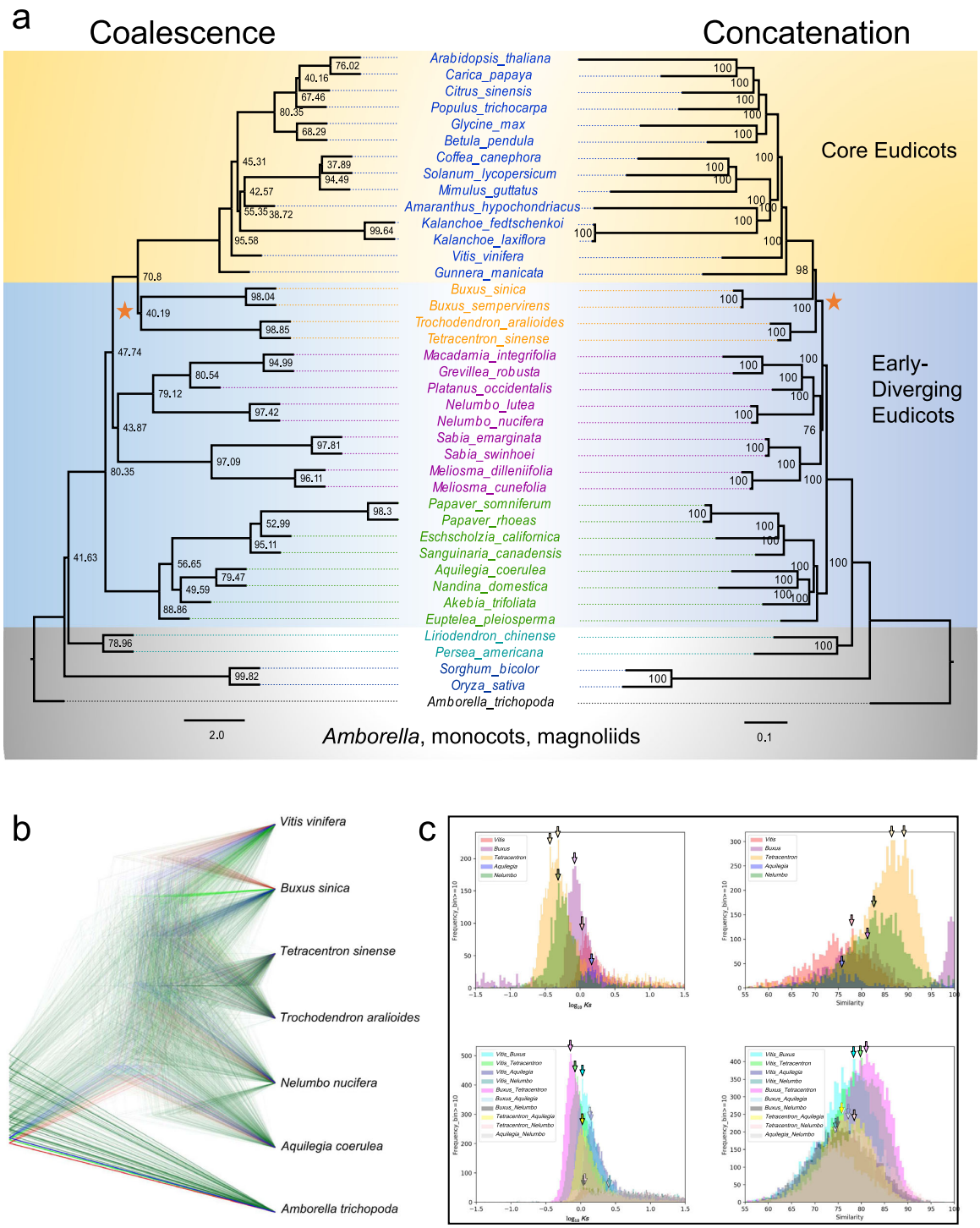


Fig. 2 Phylogenetic relations of *Buxus* and *Tetracentron*. **a** Phylograms depicting the coalescent solution of individual Maximum Likelihood (ML) gene trees (left) and partitioned ML analysis of a supermatrix of nucleotide sequence alignments (right). Node labels indicate quartet (coalescence) and bootstrap (supermatrix) support values, and orange stars highlight the positions of Buxales and Trochodendrales in the two trees. **b** Cloudogram of 763 SCN gene trees illustrating discordance surrounding the deep branches of eudicot phylogeny. The most frequent trees are blue, the next most frequent red, the third most frequent green, and the rest are dark green. **c** *Ks* (left) and Similarity (right) distributions showing peaks (arrows) that stem from WGD (top) and speciation events (bottom), respectively. Source data underlying Fig. 2b, c are provided as a Source data file.

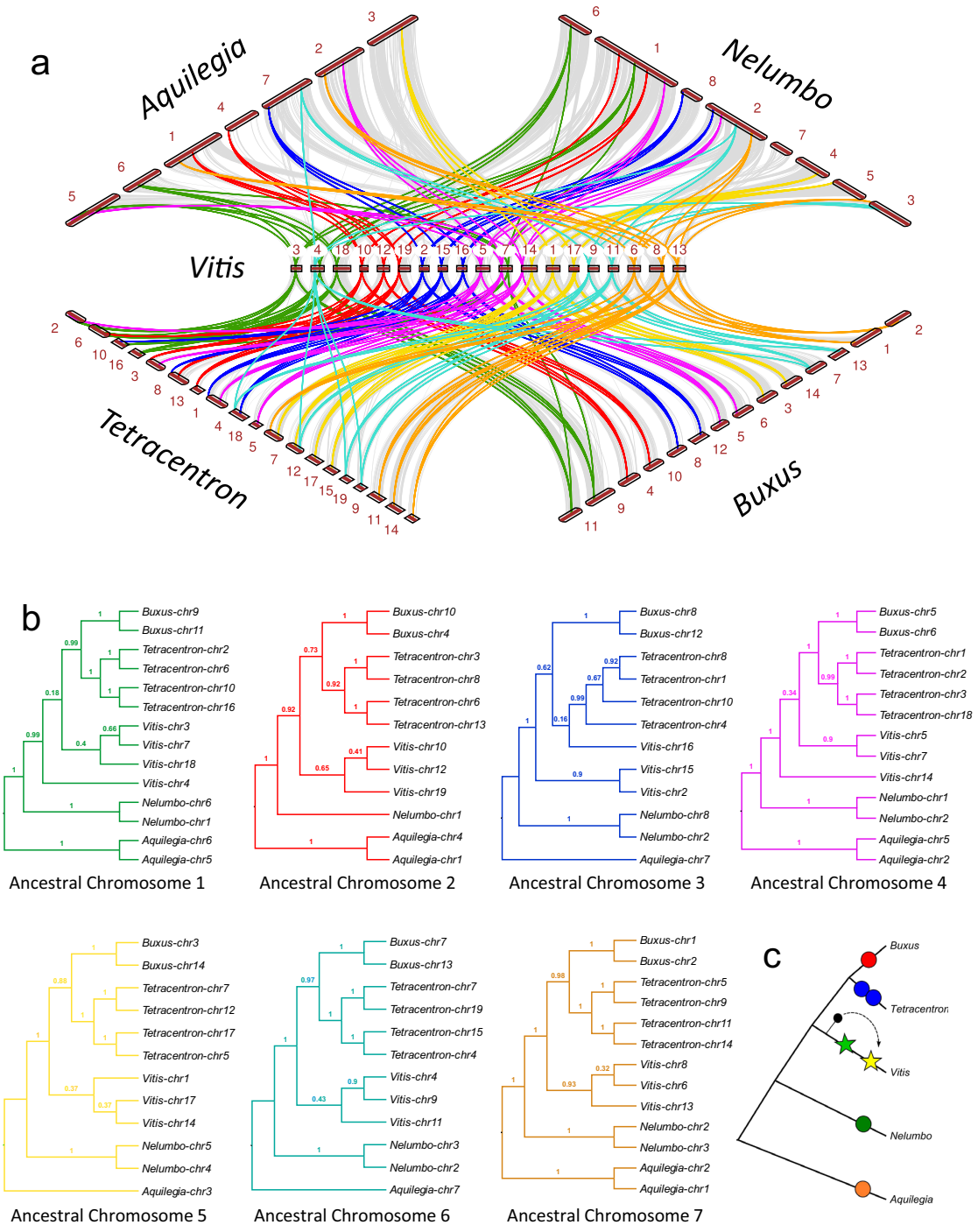


Fig. 3 Synteny and phylogenomics of eudicot subgenomes. a Macrosyntenic alignments of early-diverged eudicots against *Vitis* with tracking of genomic positions by color-coded syntenic blocks representing the seven ancestral eudicot chromosomes. **b** Coalescence-based phylogenies of syntelogs derived from duplication events affecting the seven ancestral eudicot chromosomes. Green, red, blue, purple, yellow, aqua, and brown tracks highlight positions of ancestral chromosomes 1 through 7, respectively. Branch labels are posterior probabilities. **c** Schematic reconstruction of ancient eudicot WGD history. Differently color-filled circles label putative independent duplication events and stars highlight the two *gamma* WGDs in which the third genome is donated to the initial tetraploid (green star) from an extinct lineage to form the hexaploid (yellow star). Source data underlying Fig. 3a are provided as a Source data file.

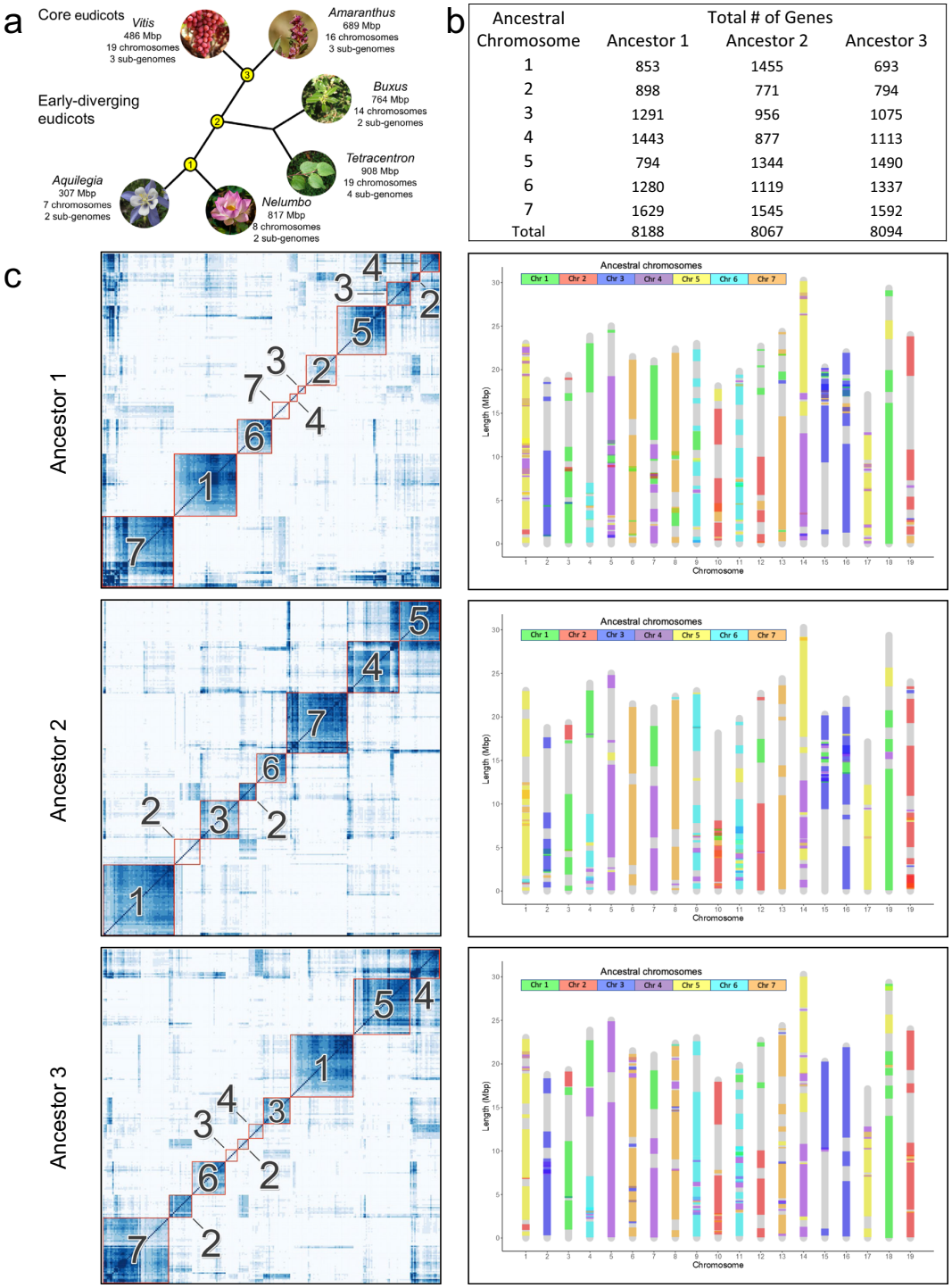
ancestral eudicot chromosomes (Fig. 3a, Supplementary Data 5, and see below), the evolutionary histories of which we have estimated through phylogenetic analyses of 1932 gene trees populated with 15872 genes (Fig. 3b). For example, syntenic blocks descended from ancestral chromosome 4 (purple tracks in Fig. 3a) occupy regions of *Vitis* chromosomes 5, 7, and 14, as well as portions of chromosomes 2 and 5 of *Aquilegia*, 1 and 2 of *Nelumbo*, 5 and 6 of *Buxus*, and 1, 2, 3, and 18 of *Tetracentron*. Microsynteny (gene level) alignments within these major synteny blocks comprise 235 homologous loci and a total of 1837 syntelogs (genes derived from the same ancestral genomic region) useful for inferring the evolutionary history of ancestral chromosome 4 (see Supplementary Data 5 for the modern distribution and representation of each ancestral chromosome). The coalescent solution of phylogenetic trees for these 235 loci shows that duplicated blocks of ancestral chromosome 4 now present in *Aquilegia*, *Nelumbo*, *Buxus*, and *Tetracentron* constitute lineage-specific clades (Fig. 3b), indicating that ancestral chromosome 4 was duplicated independently in each of the respective stem lineages of these four modern genomes. Indeed, the duplicated blocks of all seven ancestral chromosomes in *Aquilegia*, *Nelumbo*, *Buxus*, and *Tetracentron* constitute lineage-specific groupings (Fig. 3b), providing consensus that their respective WGDs are independent events and, importantly, exclusively involved genome donors that belonged to their respective clades, i.e., their stem lineage ancestors.

Phylogenetic alliances of the seven ancestral chromosomes occupying the modern, *gamma*-derived, *Vitis* genome are less clear. Of the three copies of ancestral chromosome 4, the syntenic blocks preserved on *Vitis* chromosome 5 and 7 form a well-supported sister group, but the block on *Vitis* chromosome 14 is placed as an earlier branch, albeit with low support. *Vitis*-specific clades were also not recovered for ancestral chromosomes 1 and 3, although again without high statistical support for non-monophyly. However, triplicated copies of ancestral chromosomes 2, 5, 6, and 7 in the *Vitis* genome group together as each other's closest relatives. Although clade support is strong only for the copies of ancestral chromosome 7 currently preserved on *Vitis* chromosomes 6, 8, and 13, the phylogenies of these four sets of genomic regions suggest they uniquely share a common ancestor, one that evolved separately from the other, earlier-diverged, eudicot lineages. Altogether, we recover *Vitis*-specific groupings for duplicates of four of the ancestral eudicot chromosomes, albeit as a well-supported clade only once. The relationships of the other three ancestral chromosomes may best be described as phylogenetically unresolved. Importantly, these findings are inconsistent with evolutionary scenarios of *gamma* formation through an extremely wide cross between a core eudicot and an early-diverging eudicot lineage, as has been previously proposed[22]. An initial tetraploidy event in the common ancestor of the eudicots[24] is also inconsistent with our finding that paralogous genomic blocks in *Aquilegia*, and all other basal eudicots, constitute lineage-specific clades. The only evolutionary scenario consistent with our analyses is one in which *gamma* hexaploidy exclusively involved stem lineage ancestors of extant core eudicot species as genome donors. As such, if hexaploidy was attained via a two-step process of sequential WGDs, the third of the *gamma* genomes must have been donated from a now extinct lineage that branched off the core eudicot ancestral line before the initial tetraploidy event (Fig. 3c).

Ancestral genomes. The independence of each of the WGD events associated with each of the early-diverging eudicot lineages implies unduplicated ancestral genomes leading all the way from the ancestral angiosperm up to *gamma* and the core eudicots. We explore this key inference through ancestral genome reconstruction. We reconstructed ancestral genomes at three nodes of the eudicot phylogeny (Fig. 4a): the common ancestor of the core eudicot clade (ancestor 3), two sequentially older nodes ancestral also to *Buxus* and *Tetracentron* (ancestor 2), and *Nelumbo* (ancestor 1).

All three of these ancestral genomes are reconstructed as seven putative protochromosomes, each with between 700 and 1600 protogenes, totaling more than 8000 protogenes, arranged in their ancestral order (Fig. 4b). Our ancestral genome reconstructions include ~2000 more (ca. 25%) ordered protogenes than previous reconstructions of an ancestral eudicot genome[49]. To understand the early evolution of eudicot genome structure, we partitioned the modern eudicot chromosomes into sets of syntenic regions and painted each of these according to its corresponding protochromosomes (Fig. 4c; Supplementary Fig. 10). These projections relate modern eudicot genomes to successive ancestral precursors and provide insights into the relative timing of any structural changes during eudicot genome evolution. Projections of the three ancestral genome reconstructions onto *Vitis* chromosomes (Fig. 4c) are globally similar, indicating genome structure remained relatively stable during early eudicot diversification. Inconsistent with the hypothesis of one ancestral eudicot tetraploidy[24], these projections indicate that fusion of the two ancestral chromosomes now combined in *Vitis* chromosome 7 and *Aquilegia* chromosome 5 (juxtaposed purple and green blocks in Fig. 4c and Supplementary Fig. 10, respectively) did not occur prior to the origin of the eudicot ancestor. Were this the case, both sections of these *Vitis* and *Aquilegia* chromosomes would be painted with a common color representing one ancestral chromosome whose 'chimeric' origin would be invisible to our methods. Instead, these, and other, chromosomal fusions appear to be independent, lineage-specific events that post-date ancestral genome arrangements. Several other genomic rearrangements, as measured by the 'choppiness' of chromosomal paintings (Supplementary Data 6), emerge from our reconstructions. In the case of *Vitis*, the modern genome has accumulated 41 inter-chromosomal exchanges relative to ancestors 1 and 2, and 31 after ancestor 3. The reduced number of inter-chromosomal exchanges indicates greater similarity of *Vitis* to the core eudicot ancestor (ancestor 3) relative to the more ancient ancestors 1 and 2. A similar reduction of inter-chromosomal exchanges, from 67 (relative to ancestor 2) to 56 (relative to ancestor 3), was also observed for *Amaranthus tuberculatus*, the other core eudicot genome in our analyses. As such, we can reject the occurrence of any single WGD in the eudicot stem lineage and instead firmly resolve independent WGDs in each modern eudicot lineage, including the core eudicots with their unique *gamma* hexaploid structure.

Our *Buxus* and *Tetracentron* genome assemblies have facilitated rigorous assessments of alternative hypothesized scenarios for the origin of *gamma*, a key hexaploidy associated with a major event in the history of terrestrial life, the origin of core eudicots, which comprise the vast majority of flowering plants. We have presented and analyzed several lines of evidence, including *Ks* distributions, genomic synteny, fractionation bias, phylogenomics, and ancestral genome reconstruction, that bear relevance to the phylogenetic and WGD history of the early-diverging eudicot angiosperms. These analyses reconstruct the sequential branching order of the initial eudicot radiation and show that each of the early-diverging eudicot lineages is characterized by its own independent duplication event(s). We find no evidence to support hypotheses that a single polyploidy event might have been formative for eudicot diversification as a whole. Instead, our analyses place *gamma* hexaploidy on the stem lineage of core eudicots and rule out a role for other living early-diverging



Ancestral Chromosome	Total # of Genes Ancestor 1	Ancestor 2	Ancestor 3
1	853	1455	693
2	898	771	794
3	1291	956	1075
4	1443	877	1113
5	794	1344	1490
6	1280	1119	1337
7	1629	1545	1592
Total	8188	8067	8094

Fig. 4 Ancestral eudicot genomes. a Schematic phylogeny of the eudicot clade depicting extant and ancestral genomes (nodes 1–3) examined here. Images Credits: *Vitis*, Bob Nichols; *Amaranthus*, Patrick Alexander; *Buxus*, PiPi; *Tetracentron*, Daderot; *Nelumbo*, Engin Akyurt; *Aquilegia*, Ejohnsonboulder; all available in the Public Domain via Wikimedia Commons. **b** Protogene content of ancestral genomes. **c** Heatmaps of conserved synteny supporting the delimitation of seven protochromosomes in each ancestral genome (left panel), and *Vitis* chromosomes painted according to the protochromosomes of the three diploid ancestral genomes (right panel). Source data underlying Fig. 4c are provided as a Source data file.

eudicots as genome donors, a possibility that was consistent with the results of previous analyses[13–15,22]. Without a single, linking WGD common to all eudicots, an argument that one polyploidy event may have helped spur the massive eudicot diversification (via adaptive, alternative deployments of duplicate genes), even following a time lag, is not supported by our data. Instead, each independent WGD among the early-diverging eudicot lineages, other than *gamma*, underlies relatively species-poor lineages that show limited fossil or living evidence for extensive radiation. Thus, with the genomes of all living early-diverging eudicot lineages now examined for a possible genomic contribution to *gamma*, the origin of *gamma* remains another abominable angiosperm mystery despite intensive study.

Methods

DNA extraction, sequencing, and assembly. *Buxus sinica* and *Tetracentron sinense* tissues were obtained from individuals cultivated at the University of Wisconsin-Madison (accession no. UW 136) and the University of Washington Arboretum, Seattle (accession no. 385-62), respectively. Genome sizes for these accessions were estimated using flow cytometry with BD CellQuest Pro software (Supplementary Data 7) by the Benaroya Research Institute (Seattle, WA). High-molecular-weight genomic DNA was isolated from young leaf tissue using modified nuclei-preparation and cetyltrimethylammonium bromide (CTAB) DNA extraction methods. Briefly, leaf tissue was ground to a fine powder under liquid nitrogen and mixed with nuclear isolation buffer (15 mM Tris, 10 mM EDTA, 130 mM KCl, 20 mM NaCl, 1 mM Spermine, 1 mM Spermidine, 8% PVP-10, 0.1% Triton X-100, and 7.5% 2-mercaptoethanol), passed sequentially through 100 and 40 µm mesh filters, treated with 1% Triton X-100, and centrifuged at 2000 × *g* for 10 min at 4 °C to pellet the nuclei. The pellet resuspended for 1 h at 65 °C in lysis buffer (100 mM Tris-HCl, 100 mM NaCl, 50 mM EDTA, 2% CTAB, 1% PEG 6000), and high-molecular-weight DNA was isolated from the lysate via 24:1 chloroform/isoamyl alcohol and purified with the QIAGEN Genomic kit. SMRTbell 20-kb libraries were generated and sequenced on the PacBio RSII platform to ~160x genomic coverage. In addition, Hi-C libraries were prepared and sequenced to coverage depths of ~40x by Phase Genomics (Seattle, WA). PacBio reads were assembled using the pb-assembly suite of programs which includes the FALCON/FALCON-unzip assembly pipeline and performs contig phasing and polishing[35]. The polished assemblies were deduplicated with Purge Haplotigs[50] and scaffolded using Proximity Guided Assembly (PGA) and Hi-C reads by Phase Genomics (Seattle, WA).

RNA-seq data. Transcriptome assemblies were produced for *Buxus sinica* and *Tetracentron sinense* to aid annotation of their genome assemblies. We also produced transcriptome assemblies for six additional early-diverging eudicots (*Buxus sempervirens, Meliosma dillenifolia, Nelumbo lutea, Sabia emarginata, Sabia swinhonei*, *Trochodendron ralioides*), as well as the core eudicot (*Gunnera manicata*), to improve taxon sampling in phylogenetic analyses. Paired-end RNA-seq libraries were constructed from polyA selected total RNA extracted from floral and/or leaf tissues (Supplementary Data 8), and sequenced using the Illumina HiSeq 3000 system. Reads were trimmed with Trimmomatic[51] and assembled using Trinity[52]. Coding DNA (CDS) and protein sequences were predicted with TransDecoder (http://transdecoder.github.io).

Annotation. Genomes were annotated using the MAKER pipeline[53]. De novo transcriptome assemblies for *Buxus* and *Tetracentron*, along with proteomes for four publicly available eudicot genomes—*Arabidopsis thaliana*, *Aquilegia coerulea*, *Nelumbo nucifera*, and *Vitis vinifera* (Supplementary Data 8)—were provided as evidence. Custom repeat libraries for genome masking were produced according to the MAKER-P advanced protocol[54] using LTRharvest[55], LTRdigest[56], MITE-Hunter[57], RepeatModeler[58], and RepeatMasker[59]. Gene models were predicted from the masked assemblies using the SNAP[60] and Augustus[61] ab initio predictors after three rounds of training on interim high-quality (AED <= 0.25; length >= 50 amino acids) and BUSCO gene models, respectively.

Phylogenetic analyses. Three phylogenetic data sets were compiled from translated transcriptomes or genome-annotated proteomes for 40 angiosperms (Supplementary Data 8). Conserved single-copy land plant genes were identified by BUSCO[43] analyses with the embryophyta_odb10 data set, orthologs of the Angiosperms353 loci[44] were collected by BLAST searches seeded with *Amborella trichopoda* proteins, and orthogroups were circumscribed by Orthofinder[45]. For all data sets, protein sequences were aligned using MAFFT[62] and converted to codon alignments using PAL2NAL[63], which were refined in three successive rounds of sequence filtering and trimming using trimAl[64]. Initially, sequences with less than 50% residue overlap over >70% of their length were removed to discard any potentially spurious homologs. The passing sequences were next trimmed with trimAl's heuristic automatic method (-automated1) and filtered again as above to remove sequences that might contribute extensive missing data to the phylogenetic matrix. Alignments with fewer than 4 sequences, and missing representatives of either Buxales or Trochodendrales, were discarded. After all filtering steps, 1248 BUSCOs, 346 Angiosperms353 loci, and 2573 orthogroups were retained for phylogenetic analyses. Maximum likelihood (ML) trees for the single-copy data sets were inferred from alignments of individual loci as well as concatenations of these, produced with FASconCAT[65], using RAxML[66] with the GTR + gamma model of nucleotide evolution and 1000 bootstrap replicates. Concatenated alignments were analyzed using a partition scheme that defines individual genes as units for parameter optimization. Partitioned Bayesian Inference analyses were run with MrBayes with the GTR + I + G model for all partitions. Two independent parallel runs of four Metropolis-coupled Monte Carlo Markov Chains were run for 10 million generations with sampling every 1000 generations. Majority rule consensus trees and posterior probabilities of bipartitions were computed after discarding the first 25% of the sampled trees as burn-in. Orthogroup trees were inferred with IQ-Tree[67] with the best substitution model selected from among those implemented in RAxML and 1000 ultrafast bootstrap replicates. ASTRAL-III[68] and ASTRAL-Pro[69] were used to infer the species trees from single- and multi-copy gene trees, respectively, under the multi-species coalescent. DensiTree[70] was used for visualizations of discordance among a subset of single-copy gene trees without missing taxa.

Comparative genomics of polyploidy. CoGe's SynMap and FractBias programs were used to perform genome alignments and fractionation bias calculations. FractBias analyses were conducted using all genes in the target genomes and syntenic depth settings in accordance with ploidy levels of respective genomes, as revealed by SynMap plots. All analyses can be regenerated on the CoGe platform (see Code availability below). For synteny-guided phylogenomic analyses, inter-genomic alignments were produced and screened to identify all syntenic homologs (syntelogs) present in ratios of up to 3:2:2:2:4 in *Vitis, Aquilegia, Buxus, Nelumbo*, and *Tetracentron*, respectively, using MCscan[71]. This collection of syntenic homologs was divided into seven pools in accordance with the major synteny blocks conserved across these eudicot genomes (as identified by SynMap and FractBias mappings, and which correspond with ancestral eudicot chromosomes). Unique identifiers for individual loci were replaced by 'Species_chromosome' codes to create comparable phylogenetic matrices and trees for coalescence-based phylogenetic analyses as outlined above.

Ancestral genomes. To build the three ancestral genomes indicated in Fig. 4, we use the *RACCROCHE* pipeline[34]. Briefly, *RACCROCHE* uses all the syntenically validated homolog pairs generated by SynMap and builds disjoint gene families based on the principle that a gene homologous (orthologous or paralogous) with any gene in a family must also be a member of that family. For each genome, *RACCROCHE* extracts a set of 'generalized' adjacencies, namely all oriented pairs of genes within the same window containing seven consecutive genes. The pairs are represented by the non-adjacent ends of the two genes. The genes in these pairs are then labeled according to the gene families to which they belong. Each ancestor node has three incident branches, partitioning the tree into three subtrees defined by the one incoming edge (its ancestor) and two outgoing edges (its descendants). If an adjacency is found anywhere in any of the genomes in two or three of these subtrees, it is considered a candidate adjacency. With candidate adjacencies weighted as 2 or 3 according to the number of occurrences in subtrees, a maximum weight matching (MWM) of gene ends constructs the highest weight sets of compatible contiguous adjacencies (ancestral contigs). A gene end can only be matched to one end of another gene, so that these ancestral contigs are guaranteed to be linearly, or very occasionally circularly, ordered. Inversions with breakpoints within windows of seven consecutive genes will preserve common adjacencies between two genomes, but not reading directions within the window. Common adjacencies are our primary concern, so we do not use reading direction information in MWM. Circular contigs were linearized by breaking an adjacency of lowest weight. The ancestral contigs from MWM solutions were then aligned to chromosomes of modern genomes, and co-occurring contigs were clustered to assemble ancestral chromosomes. A complete-linkage clustering was applied to the correlations of contigs' co-occurrence to assemble ancestral chromosomes[72]. To aid in future studies of the genomic organization of gene function, a GO-term enrichment analysis of the members of each gene family was implemented to produce a functional annotation for the inferred ancestral genes. The functional annotations of ancestral genomes can be downloaded from https://git.cs.usask.ca/buxus/buxus-tetra.

Reporting summary. Further information on research design is available in the Nature Research Reporting Summary linked to this article.

Data availability

All raw sequence reads used in this study have been deposited in NCBI under the BioProject accession numbers PRJNA549075, PRJNA547721, and PRJNA548936. In addition, the *Buxus* and *Tetracentron* genome assemblies, associated annotation files, and predicted CDS and protein sequences, along with all phylogenetic data sets analyzed here, and ancestral genome reconstructions have been deposited in the Dryad Digital

Repository [https://doi.org/10.5061/dryad.cjsxksn6d][73]. Source data are provided with this paper.

Code availability

Custom scripts and command-line arguments have been deposited in GitHub [https://github.com/andrechanderbali/Buxus-Tetracentron-Genomes].

Received: 29 June 2021; Accepted: 14 January 2022;
Published online: 02 February 2022

References

1. Govaerts, R. How many species of seed plants are there? *TAXON* **50**, 1085–1090 (2001).
2. Friis, E. M., Pedersen, K. R. & Crane, P. R. Cretaceous angiosperm flowers: Innovation and evolution in plant reproduction. *Palaeogeogr., Palaeoclimatol., Palaeoecol.* **232**, 251–293 (2006).
3. Magallón, S., Gómez-Acevedo, S., Sánchez-Reyes, L. L. & Hernández-Hernández, T. A metacalibrated time-tree documents the early rise of flowering plant phylogenetic diversity. *N. Phytol.* **207**, 437–453 (2015).
4. Jiao, Y. et al. Ancestral polyploidy in seed plants and angiosperms. *Nature* **473**, 97–100 (2011).
5. Amborella Genome Project. The *Amborella* genome and the evolution of flowering plants. *Science* **342**, 1241089 (2013).
6. Schranz, M. E., Mohammadin, S. & Edger, P. P. Ancient whole genome duplications, novelty and diversification: the WGD Radiation Lag-Time Model. *Curr. Opin. Plant Biol.* **15**, 147–153 (2012).
7. Vanneste, K., Maere, S. & Peer, deY. V. Tangled up in two: a burst of genome duplications at the end of the Cretaceous and the consequences for plant evolution. *Philos. Trans. R. Soc. B* **369**, 20130353 (2014).
8. Tank, D. C. et al. Nested radiations and the pulse of angiosperm diversification: increased diversification rates often follow whole genome duplications. *N. Phytologist* **207**, 454–467 (2015).
9. Soltis, P. S. & Soltis, D. E. Ancient WGD events as drivers of key innovations in angiosperms. *Curr. Opin. Plant Biol.* **30**, 159–165 (2016).
10. Landis, J. B. et al. Impact of whole-genome duplication events on diversification rates in angiosperms. *Am. J. Bot.* **105**, 348–363 (2018).
11. Cantino, P. D. et al. Towards a phylogenetic nomenclature of Tracheophyta. *Taxon* **56**, 1E–44E (2007).
12. Jaillon, O. et al. The grapevine genome sequence suggests ancestral hexaploidization in major angiosperm phyla. *Nature* **449**, 463–467 (2007).
13. Jiao, Y. et al. A genome triplication associated with early diversification of the core eudicots. *Genome Biol.* **13**, R3 (2012).
14. Vekemans, D. et al. Gamma paleohexaploidy in the stem-lineage of core eudicots: significance for MADS-box gene and species diversification. *Mol. Biol. Evol.* https://doi.org/10.1093/molbev/mss183 (2012).
15. Chanderbali, A. S., Berger, B. A., Howarth, D. G., Soltis, D. E. & Soltis, P. S. Evolution of floral diversity: genomics, genes and gamma. *Philos. Trans. R. Soc. Lond., B, Biol. Sci.* **372**, 20150509 (2017).
16. Soltis, D. E. et al. Gunnerales are sister to other core eudicots: implications for the evolution of pentamery. *Am. J. Bot.* **90**, 461–470 (2003).
17. Endress, P. K. In *Advances in Botanical Research 44: Developmental Genetics of the Flower* (eds. Soltis, D. E., Leebens-Mack, J. H. & Soltis, P. S.) 1–61 (Elsevier, 2006).
18. Endress, P. K. Flower structure and trends of evolution in eudicots and their major subclades. *Ann. Mo. Botanical Gard.* **97**, 541–583 (2010).
19. Soltis, D. E. et al. Angiosperm phylogeny: 17 genes, 640 taxa. *Am. J. Bot.* **98**, 704–730 (2011).
20. Ohno, S. *Evolution by Gene Duplication* (Springer-Verlag, 1970).
21. Lyons, E., Pedersen, B., Kane, J. & Freeling, M. The value of nonmodel genomes and an example using SynMap within CoGe to dissect the hexaploidy that predates the rosids. *Tropical Plant Biol.* **1**, 181–190 (2008).
22. Ming, R. et al. Genome of the long-living sacred lotus (Nelumbo nucifera Gaertn.). *Genome Biol.* **14**, R41 (2013).
23. Tang, H. et al. Unraveling ancient hexaploidy through multiply-aligned angiosperm gene maps. *Genome Res.* **18**, 1944–1954 (2008).
24. Aköz, G. & Nordborg, M. The Aquilegia genome reveals a hybrid origin of core eudicots. *Genome Biol.* **20**, 256 (2019).
25. Velasco, R. et al. A high quality draft consensus sequence of the genome of a heterozygous grapevine variety. *PLoS ONE* **2**, e1326 (2007).
26. Worberg, A. et al. Phylogeny of basal eudicots: insights from non-coding and rapidly evolving DNA. *Org. Diversity Evolution* **7**, 55–77 (2007).
27. Ruhfel, B. R., Gitzendanner, M. A., Soltis, P. S., Soltis, D. E. & Burleigh, J. G. From algae to angiosperms–inferring the phylogeny of green plants (Viridiplantae) from 360 plastid genomes. *BMC Evolut. Biol.* **14**, 23 (2014).
28. Moore, M. J., Soltis, P. S., Bell, C. D., Burleigh, J. G. & Soltis, D. E. Phylogenetic analysis of 83 plastid genes further resolves the early diversification of eudicots. *Proc. Natl Acad. Sci. USA* **107**, 4623–4628 (2010).
29. Sun, Y. et al. Phylogenomic and structural analyses of 18 complete plastomes across nearly all families of early-diverging eudicots, including an angiosperm-wide analysis of IR gene content evolution. *Mol. Phylogenet. Evol.* **96**, 93–101 (2016).
30. Leebens-Mack, J. H. et al. One thousand plant transcriptomes and the phylogenomics of green plants. *Nature* **574**, 679–685 (2019).
31. Filiault, D. L. et al. The Aquilegia genome provides insight into adaptive radiation and reveals an extraordinarily polymorphic chromosome with a unique history. *Elife* **7**, e36426 (2018).
32. Strijk, J. S., Hinsinger, D. D., Zhang, F. & Cao, K. *Trochodendron aralioides*, the first chromosome-level draft genome in Trochodendrales and a valuable resource for basal eudicot research. *Gigascience* **8**, giz136 (2019).
33. Liu, P.-L. et al. The Tetracentron genome provides insight into the early evolution of eudicots and the formation of vessel elements. *Genome Biol.* **21**, 291 (2020).
34. Xu, Q., Jin, L., Zheng, C., Leebens-Mack, J. & Sankoff, D. in *Lecture Notes in Bioinformatics* Vol. 12686 (2021).
35. Chin, C.-S. et al. Phased diploid genome assembly with single-molecule real-time sequencing. *Nat. Methods* **13**, 1050–1054 (2016).
36. Ghurye, J. et al. Integrating Hi-C links with assembly graphs for chromosome-scale assembly. *PLoS Comput. Biol.* **15**, e1007273 (2019).
37. Yang, X., Lu, S. & Peng, H. Cytological studies on the eastern Asian family Trochodendraceae. *Bot. J. Linn. Soc.* **158**, 332–335 (2008).
38. Van Laere, K., Hermans, D., Leus, L. & Van Huylenbroeck, J. Genetic relationships in European and Asiatic Buxus species based on AFLP markers, genome sizes and chromosome numbers. *Plant Syst. Evolution* **293**, 1–11 (2011).
39. Simão, F. A., Waterhouse, R. M., Ioannidis, P., Kriventseva, E. V. & Zdobnov, E. M. BUSCO: assessing genome assembly and annotation completeness with single-copy orthologs. *Bioinformatics* **31**, 3210–3212 (2015).
40. Seppey, M., Manni, M. & Zdobnov, E. M. in *Gene Prediction: Methods and Protocols* (ed. Kollmar, M.) 227–245 (Springer, 2019).
41. Ratter, J. A. & Milne, C. in *Notes from the Royal Botanic Garden Edinburgh (UK)* (1976).
42. Dodsworth, S., Chase, M. W. & Leitch, A. R. Is post-polyploidization diploidization the key to the evolutionary success of angiosperms?. *Botanical J. Linn. Soc.* **180**, 1–5 (2016).
43. Waterhouse, R. M. et al. BUSCO applications from quality assessments to gene prediction and phylogenomics. *Mol. Biol. Evol.* **35**, 543–548 (2018).
44. Johnson, M. G. et al. A universal probe set for targeted sequencing of 353 nuclear genes from any flowering plant designed using k-medoids clustering. *Syst. Biol.* **68**, 594–606 (2019).
45. Emms, D. M. & Kelly, S. OrthoFinder: phylogenetic orthology inference for comparative genomics. *Genome Biol.* **20**, 238 (2019).
46. Mirarab, S. & Warnow, T. ASTRAL-II: coalescent-based species tree estimation with many hundreds of taxa and thousands of genes. *Bioinformatics* **31**, i44–i52 (2015).
47. Sankoff, D., Zheng, C., Lyons, E. & Tang, H. in *Algorithms for Computational Biology* (eds. Botón-Fernández, M., Martín-Vide, C., Santander-Jiménez, S. & Vega-Rodríguez, M. A.) 3–14 (Springer International Publishing, 2016).
48. Sankoff, D. & Zheng, C. in *Comparative Genomics: Methods and Protocols* (eds. Setubal, J. C., Stoye, J. & Stadler, P. F.) 291–315 (Springer, 2018).
49. Murat, F., Armero, A., Pont, C., Klopp, C. & Salse, J. Reconstructing the genome of the most recent common ancestor of flowering plants. *Nat. Genet* **49**, 490–496 (2017).
50. Roach, M. J., Schmidt, S. A. & Borneman, A. R. Purge Haplotigs: allelic contig reassignment for third-gen diploid genome assemblies. *BMC Bioinforma.* **19**, 460 (2018).
51. Bolger, A. M., Lohse, M. & Usadel, B. Trimmomatic: a flexible trimmer for Illumina sequence data. *Bioinformatics* **30**, 2114–2120 (2014).
52. Grabherr, M. G. et al. Trinity: reconstructing a full-length transcriptome without a genome from RNA-Seq data. *Nat. Biotechnol.* **29**, 644–652 (2011).
53. Holt, C. & Yandell, M. MAKER2: an annotation pipeline and genome-database management tool for second-generation genome projects. *BMC Bioinforma.* **12**, 491 (2011).
54. Campbell, M. S. et al. MAKER-P: a tool kit for the rapid creation, management, and quality control of plant genome annotations. *Plant Physiol.* **164**, 513–524 (2014).
55. Ellinghaus, D., Kurtz, S. & Willhoeft, U. LTRharvest, an efficient and flexible software for de novo detection of LTR retrotransposons. *BMC Bioinforma.* **9**, 18 (2008).
56. Steinbiss, S., Willhoeft, U., Gremme, G. & Kurtz, S. Fine-grained annotation and classification of de novo predicted LTR retrotransposons. *Nucleic Acids Res.* **37**, 7002–7013 (2009).
57. Han, Y. & Wessler, S. R. MITE-Hunter: a program for discovering miniature inverted-repeat transposable elements from genomic sequences. *Nucleic Acids Res.* **38**, e199 (2010).

58. Hubley, R. & Smit, A. RepeatModeler Open-1.0. http://www.repeatmasker.org/RepeatModeler/ (2008).
59. Smit, A. F., Hubley, R. & Green, P. *RepeatMasker* (2013).
60. Korf, I. Gene finding in novel genomes. *BMC Bioinforma.* **5**, 59 (2004).
61. Stanke, M. & Waack, S. Gene prediction with a hidden Markov model and a new intron submodel. *Bioinformatics* **19**, ii215–ii225 (2003).
62. Katoh, K. & Standley, D. M. MAFFT multiple sequence alignment software version 7: improvements in performance and usability. *Mol. Biol. Evolution* **30**, 772–780 (2013).
63. Suyama, M., Torrents, D. & Bork, P. PAL2NAL: Robust conversion of protein sequence alignments into the corresponding codon alignments. *Nucleic Acids Res.* **34**, W609–W612 (2006).
64. Capella-Gutiérrez, S., Silla-Martínez, J. M. & Gabaldón, T. trimAl: a tool for automated alignment trimming in large-scale phylogenetic analyses. *Bioinformatics* **25**, 1972–1973 (2009).
65. Kück, P. & Meusemann, K. FASconCAT: convenient handling of data matrices. *Mol. Phylogenet. Evol.* **56**, 1115–1118 (2010).
66. Stamatakis, A. RAxML version 8: a tool for phylogenetic analysis and post-analysis of large phylogenies. *Bioinformatics* **30**, 1312–1313 (2014).
67. Nguyen, L.-T., Schmidt, H. A., von Haeseler, A. & Minh, B. Q. IQ-TREE: a fast and effective stochastic algorithm for estimating maximum-likelihood phylogenies. *Mol. Biol. Evol.* **32**, 268–274 (2015).
68. Zhang, C., Rabiee, M., Sayyari, E. & Mirarab, S. ASTRAL-III: polynomial time species tree reconstruction from partially resolved gene trees. *BMC Bioinforma.* **19**, 153 (2018).
69. Zhang, C., Scornavacca, C., Molloy, E. K. & Mirarab, S. ASTRAL-Pro: quartet-based species-tree inference despite paralogy. *Mol. Biol. Evolution* **37**, 3292–3307 (2020).
70. Bouckaert, R. R. DensiTree: making sense of sets of phylogenetic trees. *Bioinformatics* **26**, 1372–1373 (2010).
71. Tang, H. et al. Synteny and collinearity in plant genomes. *Science* **320**, 486–488 (2008).
72. Xu, Q., Jin, L., Leebens-Mack, J. & Sankoff, D. Validation of automated chromosome recovery in the reconstruction of ancestral gene order. *Algorithms* **14**, 160 (2021).
73. Chanderbali, A. S. et al. Data from: Buxus and Tetracentron genomes. *Dryad, Dataset*. https://doi.org/10.5061/dryad.cjsxksn6d (2021).

Acknowledgements

This work was supported by National Science Foundation grants DEB-1455601 to A.S.C., DEB-1457440 to D.G.H., DEB-2030871 to V.A.A., and Discovery grants to L.J. and to D.S. from the Natural Sciences and Engineering Research Council of Canada. D.S. holds the Canada Research Chair in Mathematical Genomics. We thank Brent Berger, Ray Larson, and Veronica Di Stilio for aid in plant collection and DNA extraction and Hanqi Ye for bioinformatics discussion.

Author contributions

A.S.C., D.E.S., D.G.H., D.S., P.S.S., and V.A.A. conceived and designed the study. A.S.C. generated the whole-genome and transcriptome assemblies, performed phylogenetic and comparative genomics analyses, and drafted the primary manuscript. D.S., L.J., and Q.X. generated and analyzed the ancestral genome reconstructions. Y.Z., E.C., and V.A.A. analyzed data. S.J. provided data. Additional text and discussion were provided by D.E.S., D.G.H., D.S., L.J., J.Z., P.S.S., and V.A.A. All authors approved the final manuscript.

Competing interests

The authors declare no competing interests.

Additional information

Supplementary information The online version contains supplementary material available at https://doi.org/10.1038/s41467-022-28312-w.

Correspondence and requests for materials should be addressed to Andre S. Chanderbali.

Peer review information *Nature Communications* thanks the anonymous reviewers for their contribution to the peer review of this work. Peer reviewer reports are available.

Reprints and permission information is available at http://www.nature.com/reprints

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Chapter 5

Expansion of the Pipeline and an Extensive Application

Qiaoji Xu, Lingling Jin, Chunfang Zheng, Xiaomeng Zhang, James Leebens-Mack and David Sankoff. From comparative gene content and gene order to ancestral contigs, chromosomes and karyotypes. Sci Rep, 13:6095, 2023.

scientific reports

Check for updates

OPEN

From comparative gene content and gene order to ancestral contigs, chromosomes and karyotypes

Qiaoji Xu[1], Lingling Jin[2], Chunfang Zheng[1], Xiaomeng Zhang[1], James Leebens-Mack[3] & David Sankoff[1]✉

To reconstruct the ancestral genome of a set of phylogenetically related descendant species, we use the RACCROCHE pipeline for organizing a large number of generalized gene adjacencies into contigs and then into chromosomes. Separate reconstructions are carried out for each ancestral node of the phylogenetic tree for focal taxa. The ancestral reconstructions are monoploids; they each contain at most one member of each gene family constructed from descendants, ordered along the chromosomes. We design and implement a new computational technique for solving the problem of estimating the ancestral monoploid number of chromosomes x. This involves a "g-mer" analysis to resolve a bias due long contigs, and gap statistics to estimate x. We find that the monoploid number of all the rosid and asterid orders is $x = 9$. We show that this is not an artifact of our method by deriving $x \approx 20$ for the metazoan ancestor.

Evolutionary inference on a set of species in a biological family, order or higher grouping, implies the reconstruction of ancestral phenotypes or genotypes. Phenotypic reconstruction, essentially genome-free, can be derived from comparative macroscopic or microscopic evidence from extant forms or fossils, while inference of genotypes is based on the genome sequences. In this work, we focus on analyses of annotated genes and the chromosomal ordering of these genes in the genomes of extant organisms.

The *genome-free* inference of the basic (or monoploid) ancestral chromosome number x, based on the values of x for a very large number of extant species, has a long history in plant evolutionary biology, exemplified in Grant's ground-breaking 1963 work[1], (pp. 483–487). More recently, sophisticated combinatorial optimization techniques and Bayesian inference approaches have been developed to infer ancestral chromosome numbers[2–4], but these approaches aim to elucidate neither the genetic composition nor the chromosomal structure of the ancestral species. In contrast, the present *genome-based* study, accessing all the common gene adjacencies (including "gapped" adjacencies) among species within a phylogenetic context, seeks to recover the largest possible consistent subset of these adjacencies, organized into hypothetical ancestral chromosomes of a monoploid ancestor. One such set of chromosomes is constructed for each ancestral node of the phylogenetic tree describing the relationship among the analyzed species.

Inference about ancestral genome structure is difficult in the plant kingdom (as reviewed in[5]). Adjacencies are disrupted in plant genomes by whole genome duplication followed by random deletion of duplicate genes ("fractionation"), in addition to niche-specific expansion and contraction of gene families, chromosomal rearrangements, fissions and fusions, by rampant invasions and culling of transposons, which typically comprise the majority of the genome, and other processes. Much of the work on reconstruction, e.g.,[5,6], relies on a bottom-up, greedy stepwise inference of "contiguous ancestral regions", incorporating external information, for example known whole genome duplication events, without particular attention to the number and nature of individual chromosomes.

In contrast, the focus on monoploidy in our method permits a single-step reconstruction of ancestral chromosomal fragments, *contigs*, without any recourse to information external to the given set of phylogenetically-related annotated genome sequences. The maximum weight matching (MWM) algorithm embedded in the RACCROCHE pipeline[7,8] assures a robust monoploid reconstruction; each ancestor contains at most one representative of

[1]Department of Mathematics and Statistics, University of Ottawa, Ottawa, Ontario K1N 6N5, Canada. [2]Department of Computer Science, University of Saskatchewan, Saskatoon, Saskatchewan S7N 5C9, Canada. [3]Department of Plant Biology, University of Georgia, Athens, Georgia 30602, USA. ✉email: sankoff@uottawa.ca

each gene family. These gene family representatives are organized into a number x of ancestral chromosomes, the "basic number", and ordered along the chromosomes in a way most consistent with the gene order in their extant descendants.

Somewhat unexpectedly, as illustrated in Fig. 1, the method produces more clear-cut inferences on clades of more remotely related species, such as six genomes each from a different monocot order[8], or six eudicot species from different orders[9], than sampling of lineages within orders (here Fagales) or families, where the results are degraded by high levels of noise[10].

In this paper, we trace the origin of this noise to a severe bias arising during the analysis of a clade of closely related genomes. We devise a method to eliminate this bias and thus mitigate the resulting noise. In addition we introduce methods to determine the statistically optimal number of chromosomes and reconstruct these chromosomes automatically.

We use our pipeline to determine the monoploid number of the common ancestor of all sampled species for each order as well as ancestors represented by internal nodes for each phylogeny. For each of six rosid orders—Fagales, Cucurbitales, Malpighiales, Myrtales, Malvales and Sapindales, and five asterid orders—Asterales, Gentianales, Lamiales, Solanales and Ericales, species were chosen based largely on the availability of annotated chromosome-level genome sequences representing many or most of the families in each order.

In the "Methodology" section, we present the motivations for each of the tools we introduce, with examples illustrating the problems addressed by these tools. This includes a "g-mer" technique for overcoming the contig length biaa, the sampling of 100 solutions of each MWM problem to take account of the non-uniquesness of MWM solutions, and the gap statistic approach to identify the best clustering trajectories and their inflection points.

The "Results" section summarizes our findings from applying our methods to a total of 71 plant genomes in eleven orders, producing 49 ancestral genomes in all. With 100 MWM samples for each ancestor, this required almost 5000 runs of the computationally costly MWM procedure. The full results are reported in the Supplementary Materials. The "Results" section proper contains a comparison of the remarkably parallel gap statistic trajectories of the eleven orders, all with inflection points at $x = 9$. Thus a major result of our genome-based method is that the monoploid number of the rosid and asterid orders is determined to be $x = 9$, compared to the $x = 7$ or $x = 8$ estimated from a recent genome-free study[4].

Methods

Generating sets of long contigs. To infer gene content and gene order for each chromosome in each ancestral genome in a phylogeny, we identify a large number of generalized[11] (or "gapped"[12]) gene adjacencies, allowing for example, up to 7 spacer genes between the two considered adjacent, from all chromosomes in the set of input genomes and then infer adjacencies for each ancestral node in the species phylogeny. To do this, for each ancestor, graphs generated with all phylogenetically informative generalized adjacencies as vertices and edges joining any two adjacencies that each contain one of the 5' and 3' ends of the same gene, are analyzed using the MWM algorithm. This outputs inferred linear ancestral "contigs", each containing up to several hundred genes. Figure 2 shows a typical distribution of contig content for one ancestor genome, using the methodology available preceding the innovations to be described in this section.

The data used for this work are annotated, chromosome-level or other high-quality genome sequences, accessible on the COGE[13,14] platform, or uploaded to a dedicated repertoire on this platform from public sources, as well as phylogenies for each of the orders studied, as extracted from recent literature and databases[15,16]. The only pre-processing software required was the SYNMAP[13,14] comparative genomics package, also on the COGE platform, which produces syntenically validated homology identification between genomes (orthologs) and within single genomes (paralogs). The term "gene" here is used broadly to refer to gene families, or sets of homologous genes in the extant genomes as well as the hypothetical ancestral genes inferred by our procedures.

An important observation is that the lengths of the contigs constructed from the MWM output are highly variable, ranging from a single adjacency to several hundred in some cases. The lengths of the longest few contigs

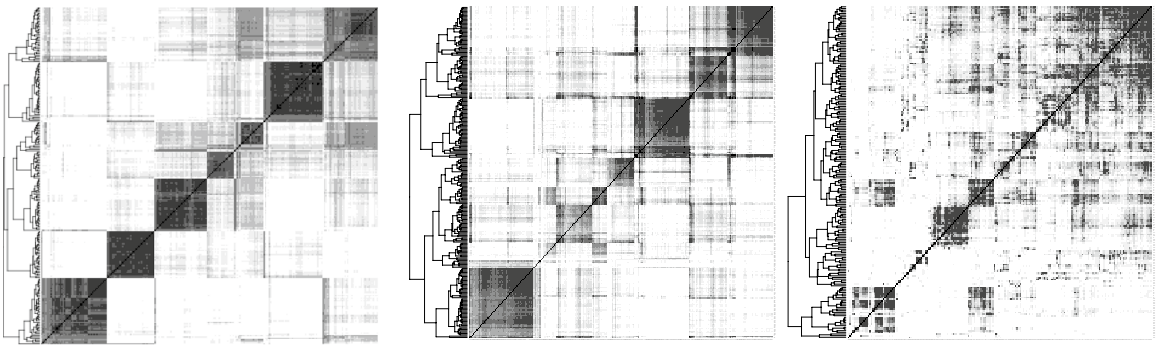


Figure 1. Clear-cut results of RACCROCHE across monocot[8] (left) and eudicot[9] (center) orders, compared to noisy results for intra-ordinal analysis of Fagales on the right (See underlying phylogeny in the "Results" section). Heat maps compare an optimal clustering of ancestral chromosomal fragments with itself, with dark cells representing two fragments which co-occur on the same chromosome in several extant genomes.

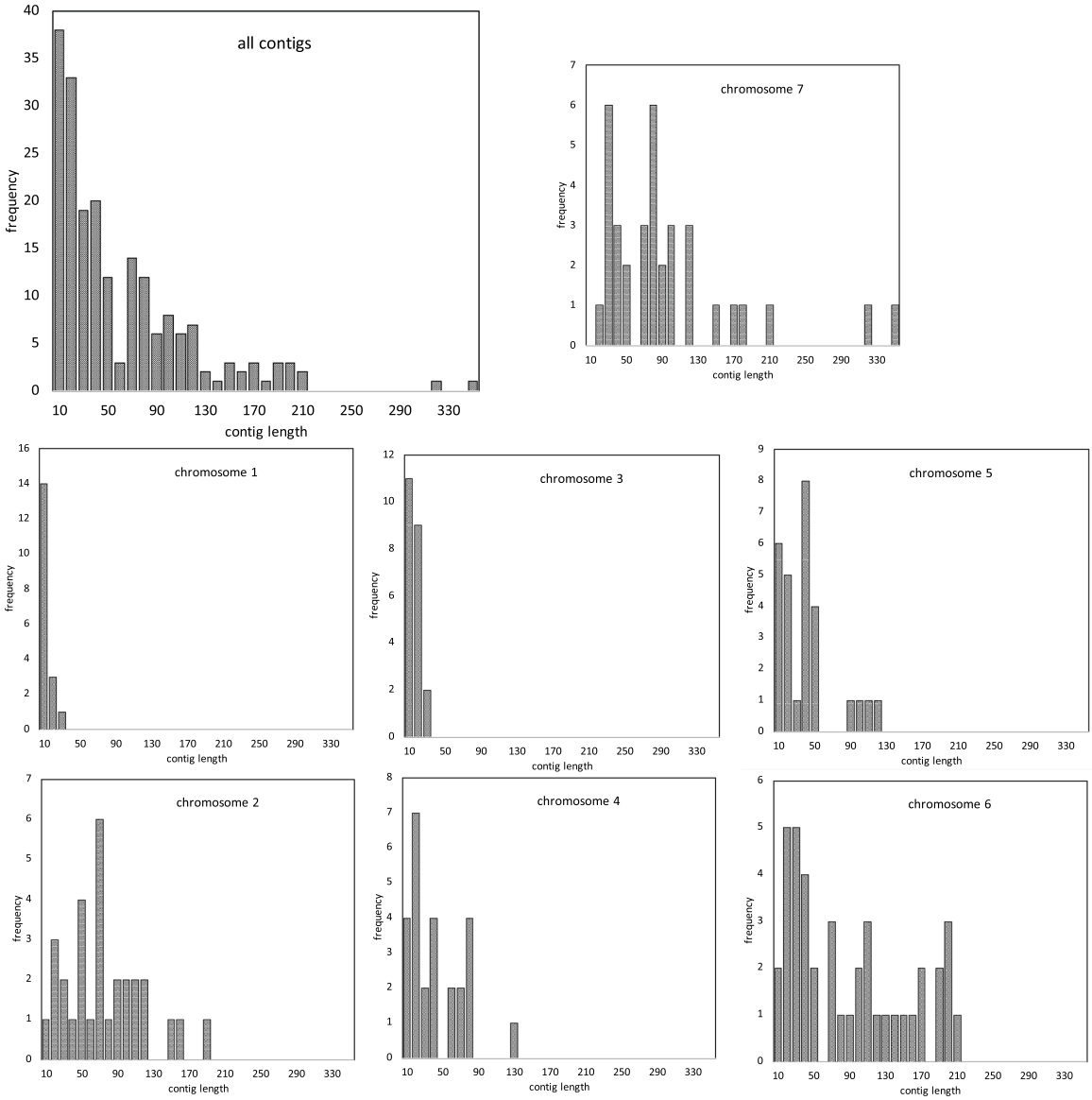


Figure 2. Contigs per chromosome, Fagales Ancestor 2 (which we use as an example throughout; see Fig. 8). N.B. Contig length is in number of genes. Frequency (raw number of occurrences) scale differs among chromosomes.

provide a measure of the conservation of gene order among the extant genomes, up to several hundred when reconstructing the ancestor of a plant family or order, compared to less than a hundred for analyses of more distantly related species in a more inclusive taxon such as the monocot or eudicot clades (on the left and center, respectively, of Fig. 1). On the other hand, the extreme variability of contig lengths encountered at the family or order level can lead to ambiguous or distorted clustering of contigs into inferred ancestral chromosomes.

Clustering the chromosomal co-occurrence matrix and the long-contig bias. To group contigs into clusters reflecting ancient chromosomes, we match each contig against the chromosomes of the extant genomes, and count the number of times any two contigs match the same chromosome, taking account of their ordering, possibly twice or more within a single genome. The resulting co-occurrence matrix, smoothed by a correlation analysis of pairs of contigs[8], is then submitted to a complete-link clustering analysis to distinguish the contigs, and hence the gene content, appropriate to each hypothetical ancestral chromosome. Once contig content of each chromosome is posited, the data on relative order of each pair of contigs on a chromosome is

submitted to a Linear Ordering Problem routine to locate them along the chromosome. It is at this point that the variability of contig length leads to serious biases, due to the longest contigs tending to group together to produce unrealistically large clusters, as illustrated in Figs. 2 and 3 for a 7-chromosome analysis of a putative Fagales ancestor. The reason for this lies in the gene families with more than one (but ≤ 10) representatives in some extant genomes. Our focus on small gene families (larger families are excluded from the analysis) rather than inferred orthologs for our ancestral contig reconstructions avoids error in orthology assignment, such as those due to widespread whole genome duplication events in plant lineages, while at the same time increasing overlap in "gene" adjacencies among analyzed genomes. This inclusion, however, allows the MWM to join distantly homologous or non-homologous generalized adjacencies when assembling the ancestral contigs. This may result in splicing of two, three or more part-contigs from different chromosomes. Thus the various long contigs tend to involve many genes in common, deriving from several chromosomes in the extant genomes. For shorter contigs, this can also happen, but is rare.

The effect of this artifact is apparent not only in imbalances among the inferred chromosomes, as in Figs. 2 and 3 , but also very noisy heat maps as on the right of Fig. 1.

Introducing *g*-mers to remove bias and noise. As illustrated with the case of Fagales in Figs. 2 and 3 , the presence of extreme-length contigs produces biases in counting contig co-occurrences, leading to an unbalanced set of chromosomes. This would seem a severe problem with the RACCROCHE method, especially when applied to sets of closely related genomes. It is possible, however, to completely remove the length bias by simply cutting each contig of length L into approximately L/g contigs of length g, called g-mers, exempting of course for contigs where $L \leq g$ already. We can then carry out the cluster analysis based on the g-mers derived from all the contigs.

A clustering may be visualized by constructing a "heat map" comparing the cluster to itself, as in Fig. 4, which shows the improvement in distinctness and size balance of an ancestral genome reconstruction through the use of g-mers. The figure suggests that at least in this example, any choice of g results in a clear improvement.

Sampling of maximum weight matching solutions. The MWM algorithm that we invoked to find an optimal matching of the adjacencies does not return a unique solution. Indeed, given the massive number of generalized adjacencies in our analyses, there may be many thousands of equally optimal solutions, usually quite similar - 95% $\pm$ 3% - but exhibiting considerable amount of variation in gene content and gene order among the ancestral contigs.

The MWM algorithm constructs all these optimal sets of matchings without taking into account the properties of the contigs they determine or the clustering used to build chromosomes. In particular, whether they give rise to neat clusters in the complete-link analysis or not, does not influence the MWM constructions.

For this reason, we sample a number (100 or 50 in this study) of optimal MWM solutions. For each g, then, our problem becomes one of searching among these solutions for one that gives the clearest clustering pattern, towards which end we implement the following definition and analysis.

Gap statistics to determine the basic chromosome number *x*. To determine the number of chromosomes in an ancestral genome, we cut the hierarchical clustering at a series of levels, starting near the root and proceeding towards the leaves, at each step increasing the number of clusters k by 1.

The gap statistic method[17] tests the significance of the k-cluster analysis for $k = 2, 3, \cdots$ against a null hypothesis that there is no clustering, i.e., $k = 1$.

A plot of this gap statistic, as on the left in Fig. 5, for a k-chromosome analysis shows a rapid, though concave, rise for $k = 2, 3, \cdots$, representing real improvements in the explanatory power of larger k, until a point where the rate of the increase drops visibly, becoming a slow linear trend measuring non-explanatory overfitting by excessive chromosome numbers. The point where one trend gives way to the other may be taken as an estimate

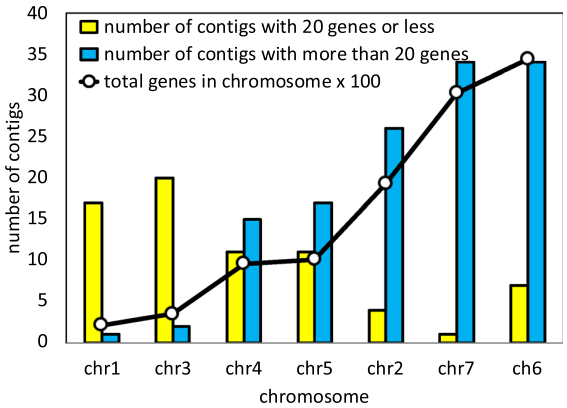


Figure 3. Statistics on the chromosomes in Fig. 2, showing bias in the assignment of long contigs.

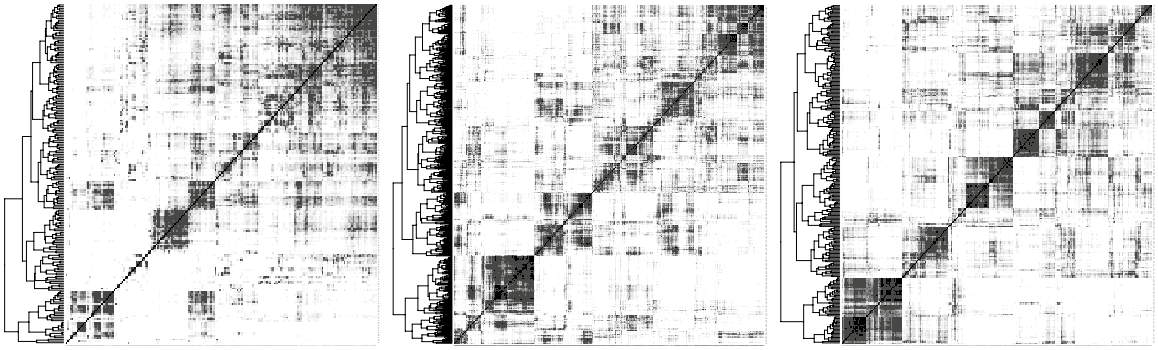


Figure 4. Heat map of Fagales Ancestor 2 based on full contigs (left) and on 20-mers (center) and 40-mers (right). The hierarchical cluster for each heat map depicted at the left-hand side. Shades of grey in cells, representing frequency of chromosomal contig co-occurrence in extant genomes, controlled to have equal darkness proportions across all heat maps.

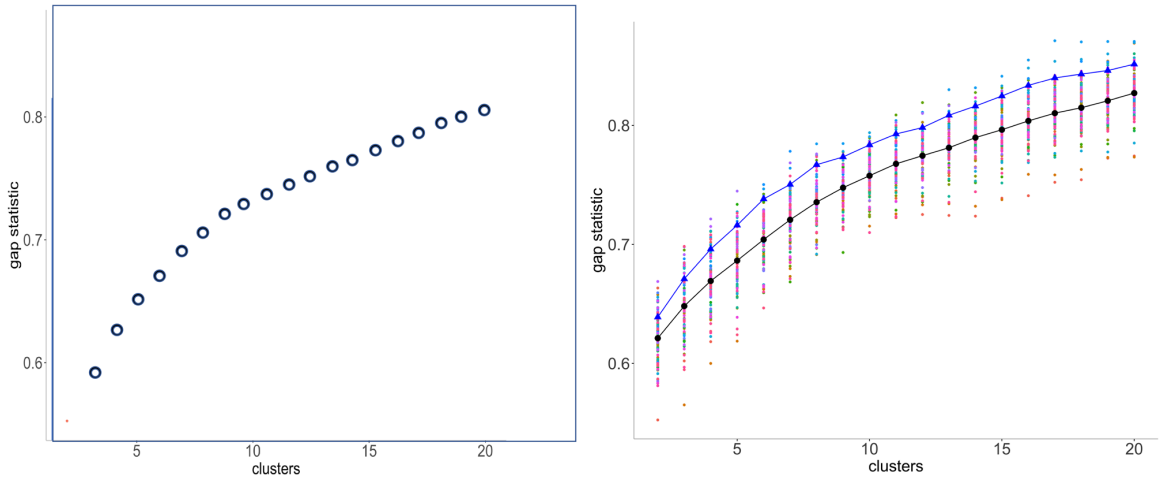


Figure 5. Left: Typical gap statistic for a single MWM sample. Right: Gap statistics for 100 samples with Fagales Ancestor 2. Black: means. Blue: means of 10 "best (as described in text).

of the basic chromosome number x. Since this typically varies among the MWM samples, we plot all the values, plus their mean (indicated by back dots in the display), on a single graph as a first step in the search for the best value of x. There are various methods for detecting the inflection point of a curve transitioning from one trend to another. The intersection of linear fits to the gap statistic for first few values of k, and for the last few k[10], proves to be unstable and misleading estimate, largely due to the variable concavity of the first trend. The method known as "kneedles"[18], based on finding the point of maximum curvature in the plot, is biased towards low values of k, also because of the concavity of the initial trend. And there are many other methods, but the most appropriate approach for our data is to fit least squares line to the noisy trend, based on $k = 12, \ldots, 20$ and to simply take x to be largest k for which the improvement in the gap statistic exceeds the prediction of this line.

As stressed above, since the MWM samples vary as to how clear a clustering they produce, we are not directly interested in the mean gap statistics, but seek instead the samples with the highest values of this statistic. Thus for each value of k, we note the 10 best values out of the hundred samples, and retain those samples that appear in the 10 best at least twice for $4 \leq k \leq 10$ In the applications to be described below, this generally resulted in a choice of 9–15 samples. (For the metazoan example introduced later, we surveyed k for $4 \leq k \leq 20$ to find the best MWM runs.) The mean gap statistics for these samples appear as blue dots in Fig. 5.

We consider the inflection point of the gap statistics in terms of the blue dots as the most pertinent to the estimate of x. And we consider the clustering with the highest score as the best choice to represent the ancestral monoploid. This clustering can be slightly adjusted, without changing k, using the Dynamic Cutting routine[19], which corrects for deeply nested subclustering as well as outlier contigs. Heat maps for the reconstructions in this paper are available in Supplementary Material A.

Choice of *g* for *g*-mers. For each ancestor, we first break down the contigs into *g*-mers as described above, calculate a new co-occurrence matrix, construct a hierarchical cluster and carry out the gap statistic analysis for each of several values of *g*. This involves choosing the best MWM sample and cluster number *k*. To see the effect of choice of *g* on the properties of the reconstructed ancestor, we can use the following statistics.

Coherence The coherence of the construction is reflected in the resemblance between each chromosome, in either an extant or ancestral descendant genome, and some chromosome of its immediate ancestor. Then for each chromosome we calculate the maximum proportion of its genes originating in any chromosome of its immediate ancestor. We average this over all chromosomes of the descendant genome. And take an overall average over all descendants in the phylogeny, separately for extants and ancestors.

Coverage This is simply the number of genes in the reconstructed genome.

Choppiness When painting an extant genome by the colors of the chromosomes of the nearest ancestor, as illustrated in Fig. 12, we define the choppiness by counting the number of single colour regions ($>$ 300 Kb) on all the extant chromosomes. This an indicator of how much genome rearrangement has intervened between the ancestor and its descendent.

Figure 6 suggests that for $g > 10$, there is little change in the quality of the reconstruction for *g* up to 40, at least. While the levels of the evaluation statistics vary from order to order, as seen in the Supplementary Material B, there is no systematic dependence on *g*.

Alternative clustering. Our final partition of the contigs into discrete clusters does not retain any inter-chromosomal relationships. However, the stepwise decomposition of the higher order links in the hierarchical clustering, as a "greedy" procedure, may lead to suboptimal results. This is mitigated by our use of dynamic tree cutting, which can redress inappropriate hierarchical constraints, and even more important, by the extensive sampling of MWM solutions, from which the most cleanly separated reconstructions are selected.

Approaches such as *k*-means are also possible, but this incurs stability issues with the co-occurrence matrix and does not possess the contig ordering properties of the hierarchical clustering.

Perhaps more important is that the matrix of contig co-occurrences, as well as the derived correlations between contigs, are situated in a very high dimensional space. With such data, much of the variance resides in relatively few dimensions. Principal Component Analysis (PCA) allows us to home in on these important dimensions, relegating the rest to noise. The clustering can then be done based on the coordinates of the contigs in these few dimensions only. Using the HCPC package for R, we produce the two-dimensional display for a Fagales Ancestor 1 in Fig. 7. Similar plots for the other ten orders are presented in Supplementary Material E.

Results

The Angiosperm Phylogeny Group circumscription of flowering plant orders and families, version IV[15], includes eight fabid and eight malvid orders (plus Vitales) as making up the rosids as well as seven campanulid and eight lamiid orders (plus Ericales) as constituting the asterids. We wished to include as many orders as possible in our study, ideally with access to at least six genomes with high quality, preferably chromosome-level, assemblies, distributed among at least three different families. At the time of data collection, we could obtain suitable data from three fabid orders, Fagales[20–27] (CoGe IDs: 28205, 35079, 51680, 60890-60894, 61298), Cucurbitales[28–33] (CoGe IDs: 51412, 52000, 52078, 52080, 52081, 52083, 52084) and Malpighiales[34–39] (CoGe IDs: 16772, 60439, 63100, 63108-63110); three malvid orders, Myrtales[40–44] (CoGe IDs: 35018, 63010, 63011, 63078, 63095),

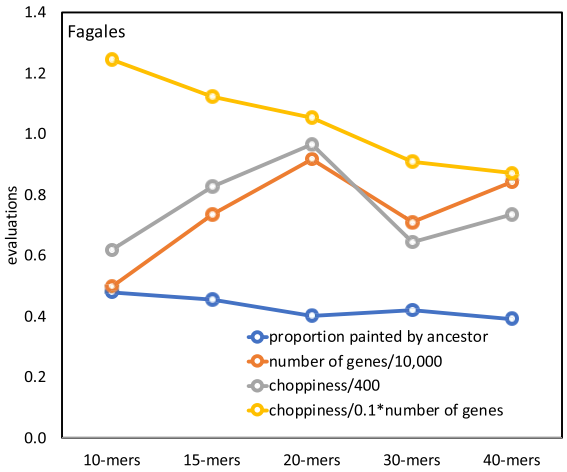


Figure 6. Reconstruction quality as a function of *g*. Averaged over all extant Fagales genomes. Dip at $g = 30$ only reflects a computation cost-imposed switch from 500 to 250 contigs in the clustering.

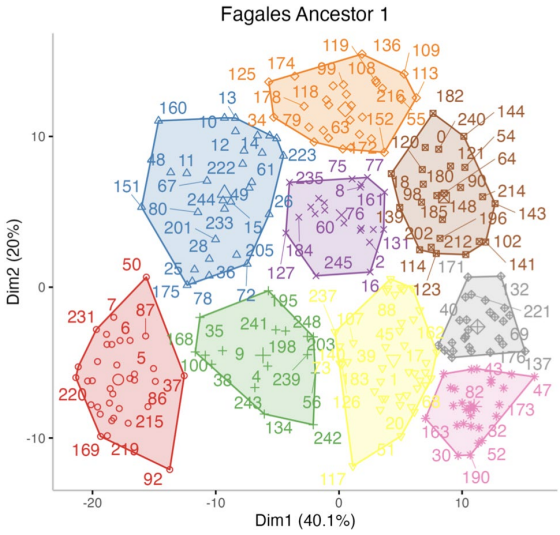


Figure 7. Nine clusters in first two principal components for Fagales Ancestor 1 20-mer data.

Malvales[45–50] (CoGe IDs: 10997, 51247, 51249, 51764, 51857, 57762) and Sapindales[51–61] (CoGe IDs: 53702, 60071, 60073-60076, 60708, 61333); one campanulid order, Asterales[62–68] (CoGe IDs: 28333, 63635, 63704, 63706, 63708, 63722); and three lamiid orders, Gentianales[69–75] (CoGe IDs: 36623, 54651, 62692, 63659, 63600, 63659), Lamiales[76–82] (CoGe IDs: 55705, 55706, 61332, 62516, 63702, 63658) and Solanales[83–88] (CoGe IDs: 52600, 54650, 54663, 57792, 61620, 63661); plus Ericales[89–94] (CoGe IDs: 60226, 61151, 62508, 62516, 62597, 63696). The phylogenies we used for the rosid orders appear in Fig. 8, those for the asterids in Fig. 9. Other orders with sufficient genomes available were not selected, such as Fabales, because representative genomes from only one or two families were available, or Brassicales, which is the subject of a separate manuscript.

In the case of our data, there is some uncertainty about locating the inflection point within ± 2 for any particular ancestor and any particular g-mer analysis. But the inflection point does not display any sensitivity to g in the range, say, from 15-50, although the overall gap score plot may be shifted upwards or downwards for different g, as in Fig. 10. Further there does not seem to be any tendency for the estimate of x to vary from ancestor to ancestor within an order; which is understandable as the basic chromosome number would tend to be the same across a single order. More complete data appears in Supplementary Materials C.

Though the gap statistics curves all display similar shapes, the improvement, namely increment in the significance level from $k-1$ to k is subject to considerable statistical fluctuation, which is the reason for the uncertainty in determining x, even for the best MWM samples. To attack this problem, under the hypotheses that the choice of g in the range from 10–15 to 40–50 is of little consequence, and that all the ancestors in an order have the same monoploid number, we take the average of the gap statistic across all these ancestors and all the g as most likely to reveal the trends in the order.

In addition, to amplify the visual impression of the tendencies in gap statistic, we can plot the improvement, i.e., the increment of this quantity from $k-1$ to k, instead of the statistic itself. We display this increment in Fig. 11.

Once the ancestral reconstructions are completed, we can visualize the evolution of an extant genome from its most recent ancestor. We assign a colour to each chromosome in this ancestor, and then assign that colour to any region of an extant chromosome that matches with a contig in the ancestral chromosome. Examples are shown in Fig. 12. Further examples are in the Supplementary Materials D.

The reconstruction analysis within each of the 11 core eudicot orders was carried out independently of the other orders, including the identification of the gene families. To what extent do the ancestors of the various orders resemble each other? To answer this for a particular pair of orders, we first have to determine which gene families in one order correspond to a gene family in the other. This can be done by finding pairs of genes in extant genomes that are orthologous. Once these are identified we can determine the co-occurrence of gene families across the chromosomes of the ancestral genomes in the two orders. Figure 13 gives the results of this for the Fagales and Mapighiales orders. It can be seen that for the most part, we can identify corresponding chromosomes for the two orders.

Another aspect of the consistency of our reconstruction is a comparison with the PCA-based reconstruction. Figure 14 shows that although there are many genes that do not fit the general pattern, we can still identify, in most cases a 1-1 correspondence between the two sets of chromosomes. In the figure, 7 out of 9 chromosomes correspond in this way.

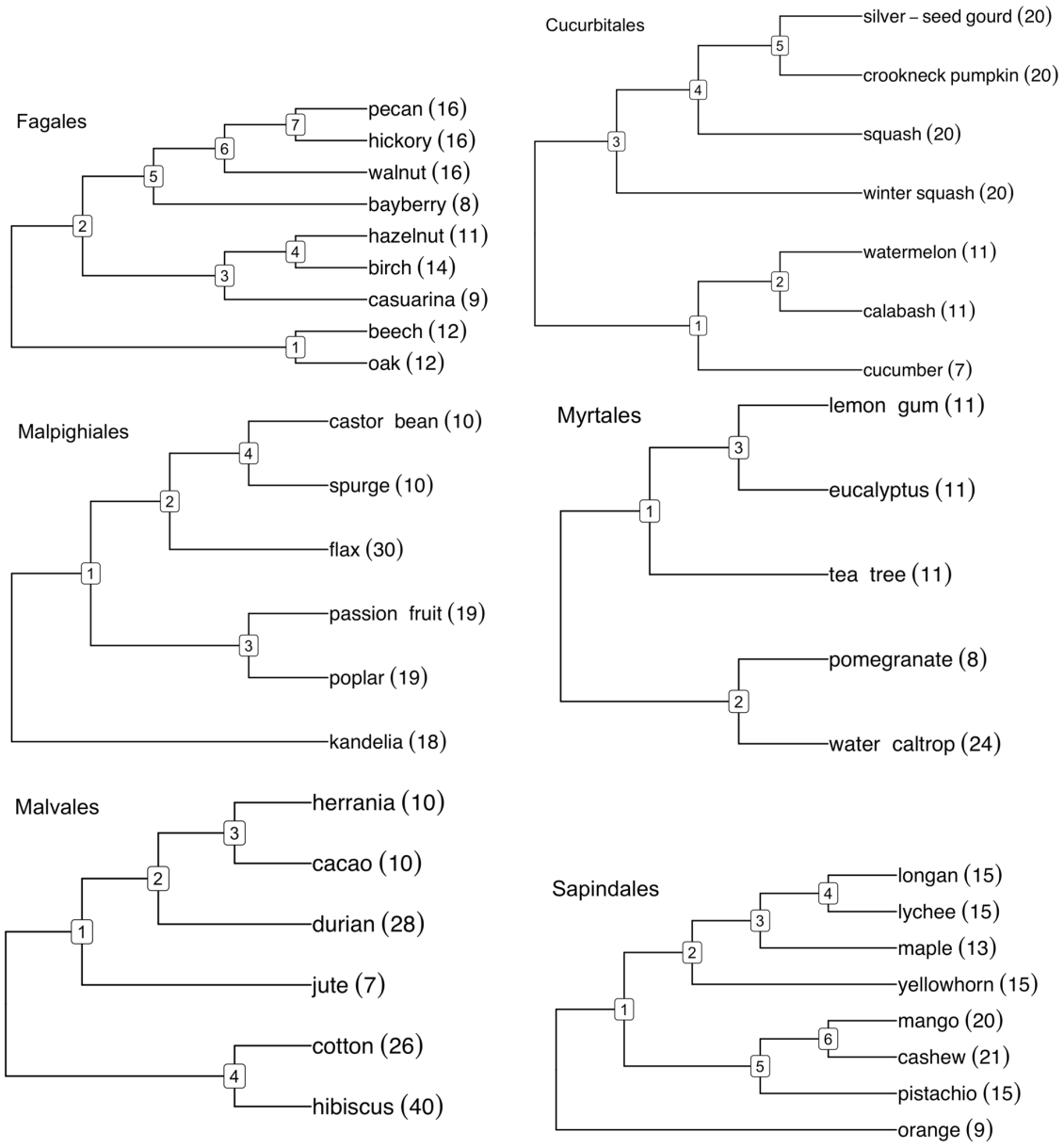


Figure 8. Phylogenies of rosid orders with haploid numbers of chromosomes.

No upper limit on *x*

Our reconstruction of monoploid ancestors of rosid ancestors, not only at the highest nodes, but for somewhat more recent ancestors, all seem to have monoploid number $k \leq 9$. These results are eminently plausible, but still may provoke the question of whether RACCROCHE would even be able to detect a higher k for an ancestor if this were warranted.

Lacking any knowledge of ground truth about ancient plant karyotypes, we could have recourse to simulations, and simulation protocols have been used successfully in studying plant evolution[8]. But simulations of plant evolution starting with an $x = 20$, say, ancestor, and known parameters for chromosome fusion, whole genome duplication, and other processes, could only be very speculative, generating unrealistic versions of extant genomes to test the RACCROCHE reconstruction.

Instead, we venture outside the plant world to an evolutionary domain where the monoploid ancestor is agreed to have x around 20, namely the animals, or metazoans[95]. We used two out of the five genomes from those studied in[95], namely the lancelet *Branchiostoma floridea* (CoGe ID 63435), representing the deuterostomes,

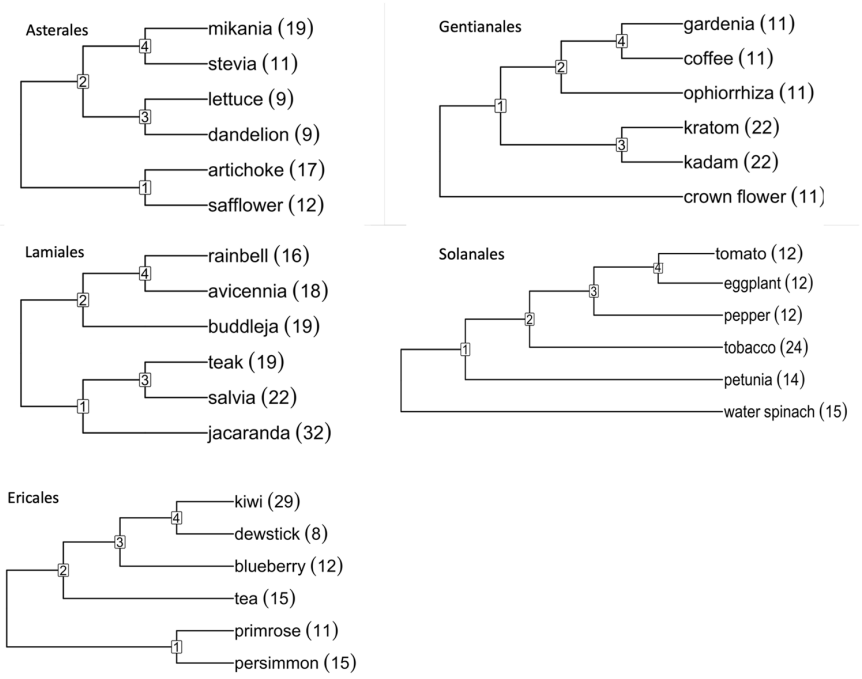


Figure 9. Phylogenies of asterid orders with haploid numbers of chromosomes.

and the sponge *Ephydatia muelleri*[96] (CoGe ID 63175). The annotated genome files from the other three species in[95] being publicly unavailable, we substituted *Octopus sinensis*[97] (CoGe ID 63434) from the phylum Mollusca as a representative of the protostomes, the cnidarian *Acropora millepora*[98] (CoGe ID 63395) and the placozoan *Trichoplax adhaerans*[99] (CoGe ID 63410). These five species represent major branches of the animal kingdom, including the subkingdoms Porifera (sponges) and Eumetazoa, the latter branching into the placozoan and cnidarian phyla and the bilaterians - protostomes and deuterostomes, as in Fig. 15. We note that the time scale is 6 to 10 times as long as that of the plant orders we have focused on. Not surprisingly, given the well-known lack of conserved gene order among early metazoan lineages[100], RACCROCHE produces relatively short contigs with these data.

The results of our analysis is summarized in the significance increment graph Fig. 16. Here the monoploid number appears to be between 15 and 25, and certainly not in the range from 7 to 9.

Discussion

Grant's visionary work on basic chromosome number of ancestral plants[1] predated genomics by several decades, but made use of data on many thousands of species to produce excellent estimates. Modern genome-free approaches[2–4] use sophisticated statistical methodology on greatly expanded data sets to improve and automate this line of research.

With the rise of molecular approaches to evolution and genomics, however, it behooves us to investigate whether the gene order on the chromosomes of a set of related extant genomes carry a signal about the basic chromosome number.

Despite their demonstrated ability to estimate gene content and to some extent gene order in reconstructed ancient genomes, the problem of delimiting chromosomes in an automated way has proved difficult[5]. Our approach differs from previous methods in that it focuses solely on monoploid reconstructions, whether or not this corresponds to the ploidy of the hypothesized ancestors. This is done in a purely automated way, given the gene orders on the chromosomes or scaffolds of extant genomes, as well as their phylogenetic relationships, without taking into account supplementary information or hypotheses in the process. The use of MWM, *g*-mer decomposition, chromosomal co-occurrence matrices and gap statistics to achieve the monoploid reconstruction is entirely novel.

The result of applying our method to eleven rosid and asterid orders, without directly referencing chromosome numbers of the extant genomes, is that the basic chromosome number of these core eudicots is nine. This is somewhat higher than the value of eight recently obtained by genome-free methods[4] using chromosome numbers of many thousands of extant species, but not at all inconsistent with Grant's original assessment[1], [p. 486].

Genome-free analysis and genome-based reconstruction both aim to infer the basic chromosome number x of entire orders, but their data and algorithmic approaches are completely different. Genome-free analyses are basically improvements on Grant's ideas from six decades ago. The basic data are just the chromosome number

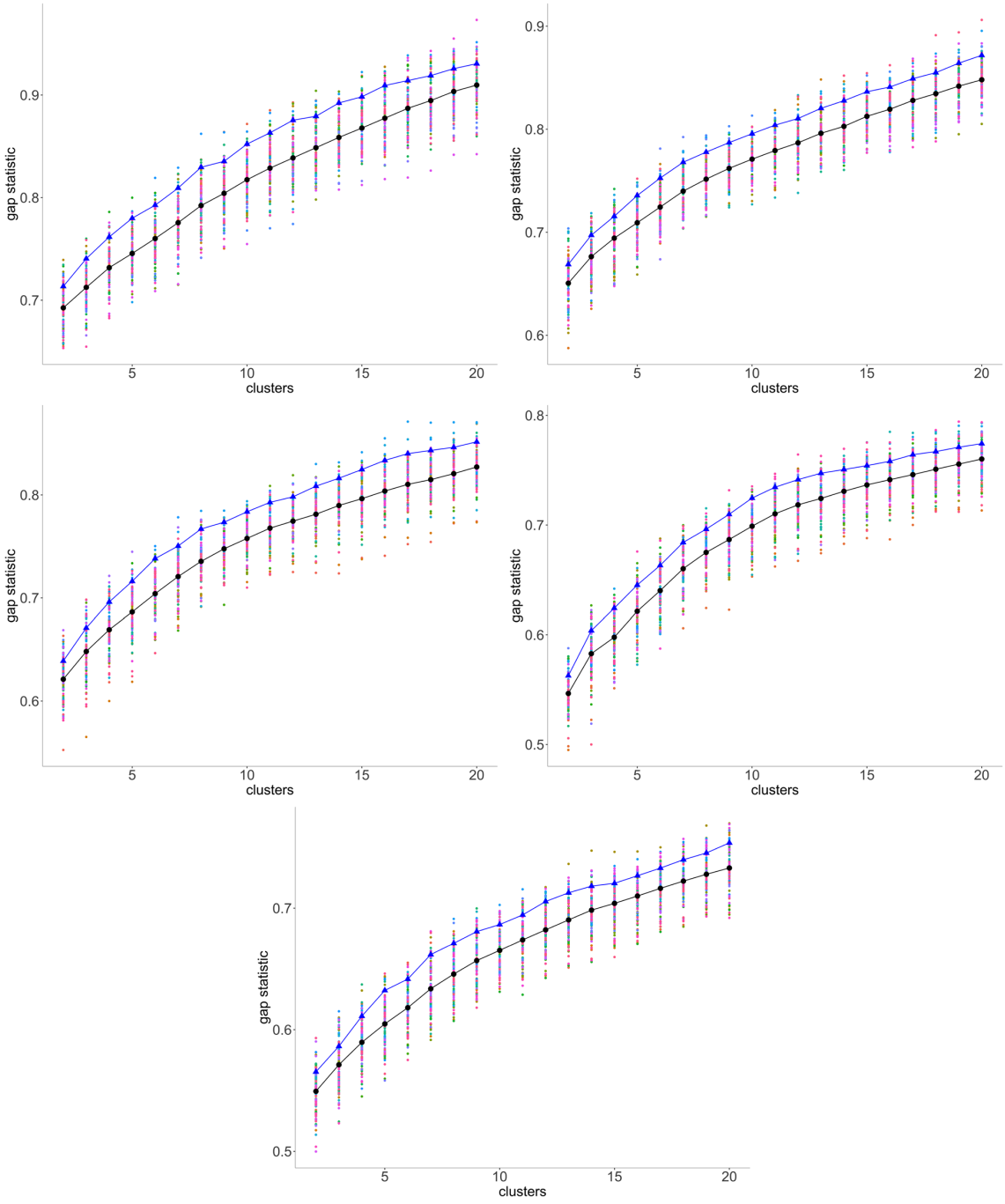


Figure 10. Gap statistics for five Cucurbitales ancestors for 100 samples; means and means of 10 best (in blue). Results for 20-mers. From left to right, and from top to bottom, panels display results for Ancestors 1, 2, 3 , 4 and 5, respectively.

from each genome. Traditionally this was derived from cytology, long before any genomes were sequenced. The algorithms derive more ancestral chromosome numbers from those of more recent ancestors or extant species. There is no reference to genes. The genome-based approach on the other hand, does not make use of the chromosome number of the extant genomes nor does the inferred chromosome number of one ancestor depend on that of another ancestor. The data are all derived from the tens of thousands gene adjacencies in each extant genome.

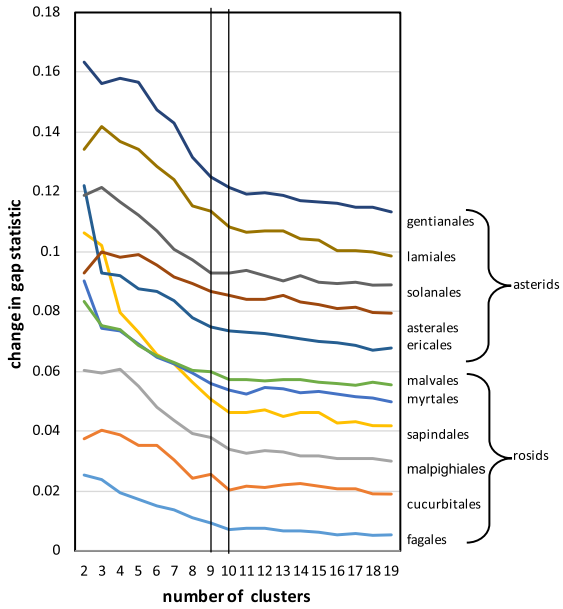


Figure 11. Transition between meaningful and noisy phases of increase in gap statistic for 11 core eudicot orders. The y-axis values are displaced + 0.01 for each order in the list above the previous item. Vertical lines at $k = 9$ and $k = 10$ highlight that for most orders, the increment at $k = 10$ is in line with the trend of uninformative additional clustering at higher values of k while for $k = 9$ the increment exceeds this trend.

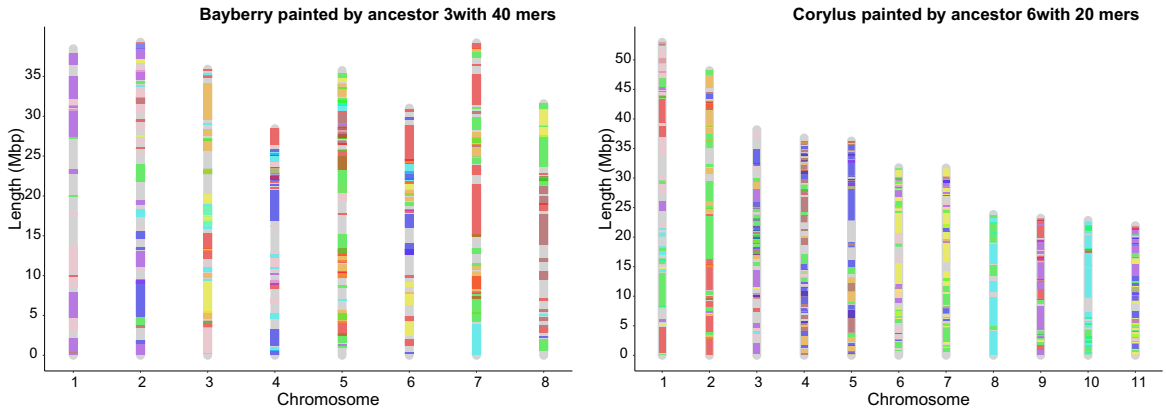


Figure 12. Painting the chromosomes of the extant genomes using colors corresponding to the ancestral chromosomes. Example of ancestral Fagales colors on the oak and the bayberry genomes.

And the algorithms are combinatorial optimization in nature, handling all the adjacencies simultaneously. In short the genome-free and genome-based approaches have nothing in common except their final inferences.

Implicit in the notion of the basic chromosome number of an order is the idea that this pertains to a monoploid ancestor. Our approach is unique in that it is the only one that is designed to infer monoploid ancestors, as achieved through the MWM algorithm.

Our reconstructions must be considered preliminary. First, we only recover around 10,000 gene families, less than half of what we expect from plant genomes, even those which have not undergone whole genome duplication. Second the assignment of these "genes" to chromosomes, and their ordering along the chromosomes varies somewhat from one MWM sample to another, even among those resulting in the clearest clustering. Nevertheless, every stage of our pipeline, which is not influenced by any information or data outside the input genomes, produces global or locally optimal results. Moreover, we have several indications of consistency, including chromosome-by-chromosome correspondences among the ancestors from different core eudicot orders, which are constructed independently from entirely different genomes. This is also clear from the parallel plots of gap statistics increments and the switch between meaningful increase and noisy increase. We also

Fagales chromosome

Malpighiales chromosome	1	2	3	4	5	6	7	8	9	total
1	195	138	117	185	60	36	30	44	45	850
2	162	139	257	426	58	53	65	277	26	1463
3	14	44	53	29	382	35	41	25	226	849
4	124	119	24	28	42	268	15	12	13	645
5	47	189	29	27	36	30	84	15	36	493
6	25	43	32	24	42	81	220	13	10	490
7	11	21	198	65	16	17	6	151	8	493
8	60	254	27	32	12	46	49	11	12	503
9	18	30	135	36	43	14	33	21	25	355
total	656	977	872	852	691	580	543	569	401	6141

Figure 13. Gene families shared between Fagales and Malpighiales ancestors. For each Malpighiales ancestral chromosome, the yellow or green cell indicates the Fagales chromosome that shares a maximum number of gene families. For each Fagales ancestral chromosome, the blue or green cell indicates the Malpighiales chromosome that shares a maximum number of gene families. There are six green cells indicating closely related chromosomes in the two independently calculated ancestors. A total of 8419 gene families were reconstructed in the Malpighiales ancestor, of which 2278 were not recovered in Fagales. A total of 9424 gene families were reconstructed in the Fagales ancestor, of which 3283 were not recovered in Malpighiales.

RACCROCHE:	1	2	3	4	5	6	7	8	9
PCA:									
1	0	0	1226	0	0	0	0	0	0
2	0	20	0	0	0	781	0	92	0
3	0	0	0	0	187	0	628	0	0
4	140	333	0	0	0	0	0	441	151
5	0	0	0	0	607	0	220	17	287
6	0	60	0	360	0	180	0	0	0
7	745	204	0	0	200	0	0	60	269
8	685	590	0	0	0	0	0	70	0
9	140	73	0	699	0	0	0	60	0

Figure 14. Gene families shared between hierarchical clustering-based chromosomes and PCA-based chromosomes. A green cell indicates that a chromosome from one method shares a maximum number of gene families with a single chromosome from the other. method.

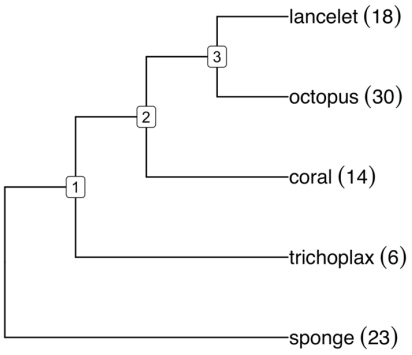


Figure 15. Metazoan phylogeny with haploid numbers of chromosomes. We can consider ancestral genome 1 to represent either the eumatazoan ancestor, a sister group to the sponges (porifera), or the more recent parahoxoan ancestor, giving rise to the placazoans like trichoplax, the cnidarians like coral and the bilaterians. Ancestral genome 2 represents the common ancestor of the cnidarians, such as coral, and the bilateria. Ancestral genome 3 is the ancestor of the bilateria, including the protostomia and the deuterostomia.

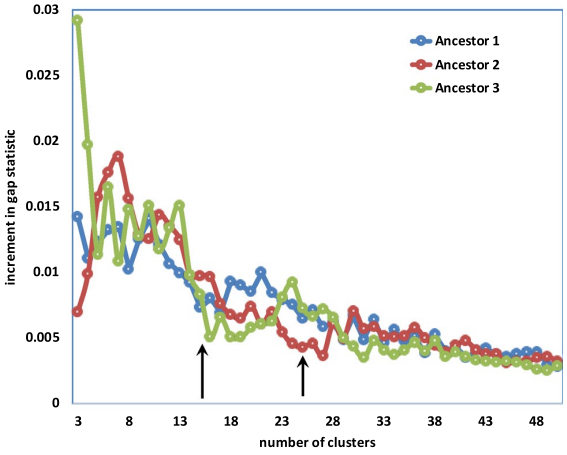


Figure 16. Significance increment graph for cluster-based metazoan karyotypes.

have correspondences between the results from different clustering methods. Furthermore we know that these results are not an artifact of some limitation in the detection power of our method; it successfully estimated an x value twice as large as the core eudicots for the ancestral metazoan genomes. This work opens up new directions for research into the evolution of the chromosomal structures of plants and other organisms.

Data availability

The assembled and annotated genomes analysed in the current study are publicly available in the CoGe platform, https://genomevolution.org/coge/. Unique CoGe ID numbers for the genomes in each order (and for the metazoans) are given in the above text.

Received: 29 October 2022; Accepted: 6 April 2023
Published online: 13 April 2023

References

1. Grant, V. *The Origin of Adaptations* (Columbia University Press, 1963).
2. Glick, L. & Mayrose, I. ChromEvol: Assessing the pattern of chromosome number evolution and the inference of polyploidy along a phylogeny. *Mol. Biol. Evol.* **31**(7), 1914–1922. https://doi.org/10.1093/molbev/msu122 (2014).
3. Goldberg, E. E. & Igić, B. Tempo and mode in plant breeding system evolution. *Evolution* **66**, 3701–3709. https://doi.org/10.1111/j.1558-5646.2012.01730.x (2012).
4. Carta, A., Bedini, G. & Peruzzi, L. A deep dive into the ancestral chromosome number and genome size of flowering plants. *New Phytol.* **228**, 1097–1106 (2020).
5. Anselmetti, Y., Luhmann, N., Bérard, S., Tannier, E. & Chauve, C. Comparative methods for reconstructing ancient genome organization. *Comp. Genom.* **1704**, 343–362. https://doi.org/10.1007/978-1-4939-7463-4_13 (2018).
6. Murat, F. *et al.* Reconstructing the genome of the most recent common ancestor of flowering plants. *Nat. Genet.* **49**, 490–496. https://doi.org/10.1038/ng.3813 (2017).
7. Xu, Q., Jin, L., Zheng, C., Leebens-Mack, J. H. & Sankoff, D. Raccroche: Ancestral flowering plant chromosomes and gene orders based on generalized adjacencies and chromosomal gene co-occurrences. *Lect. Notes Comput. Sci.* **12686**, 97–115 (2021).
8. Xu, Q., Jin, L., Leebens-Mack, J. H. & Sankoff, D. Validation of automated chromosome recovery in the reconstruction of ancestral gene order. *Algorithms* **14**(6), 160 (2021).
9. Chanderbali, A. S. *et al.* Buxus and Tetracentron genomes help resolve eudicot genome history. *Nat. Commun.* **13**, 643. https://doi.org/10.1038/s41467-022-28312-w (2022).
10. Xu, Q. *et al.* Ancestral flowering plant chromosomes and gene orders based on generalized adjacencies and chromosomal gene co-occurrences. *J. Comput. Biol.* **28**(11), 1156–79 (2021).
11. Yang, Z. & Sankoff, D. Natural parameter values for generalized gene adjacency. *J. Comput. Biol.* **17**(9), 1113–1128 (2010).
12. Gagnon, Y., Blanchette, M. & El-Mabrouk, N. A flexible ancestral genome reconstruction method based on gapped adjacencies. *BMC Bioinform.* **13**, 4 (2012).
13. Lyons, E. & Freeling, M. How to usefully compare homologous plant genes and chromosomes as DNA sequences. *Plant J.* **53**, 661–673 (2008).
14. Lyons, E. *et al.* Finding and comparing syntenic regions among Arabidopsis and the outgroups papaya, poplar and grape: CoGe with rosids. *Plant Physiol.* **148**, 1772–1781 (2008).
15. Chase, M. W. *et al.* An update of the Angiosperm Phylogeny Group classification for the orders and families of flowering plants: APG IV. *Bot. J. Linnean Soc.* **181**, 1–20 (2016).
16. Stevens, P. F. *Angiosperm Phylogeny Website. Version 14.* http://www.mobot.org/MOBOT/research/APweb/ (2017).
17. Hastie, T., Tibshirani, R. & Walther, G. Estimating the number of clusters in a data set via the gap statistic. *J. R. Stat. Soc. B* **63**, 411–423 (2001).
18. Satopää, V., Albrecht, J., Irwin, D. & Raghavan, B. Finding a "Kneedle" in a haystack: detecting knee points in system behavior. *31st International Conference on Distributed Computing Systems Workshops* 166–171. https://doi.org/10.1109/ICDCSW.2011.20 (2011).
19. Langfelder, P., Zhang, B. & Horvath, S. Defining clusters from a hierarchical cluster tree: The Dynamic Tree Cut package for R. *Bioinformatics* **24**(5), 719–720 (2007).

Fagales

20. Mishra, B. *et al.* A reference genome of the European beech (*Fagus sylvatica* L.). *GigaScience* **7**(6), 1–8. https://doi.org/10.1093/gigascience/giy063 (2018).
21. Plomion, C. *et al.* Oak genome reveals facets of long lifespan. *Nat. Plants* **4**, 440–452. https://doi.org/10.1038/s41477-018-0172-3 (2018).
22. Chen, S. *et al.* Genome sequence and evolution of Betula platyphylla. *Hortic. Res.* **8**, 21037. https://doi.org/10.1038/s41438-021-00481-7 (2021).
23. Li, Y. *et al.* The Corylus mandshurica genome provides insights into the evolution of Betulaceae genomes and hazelnut breeding. *Hortic. Res.* **8**, 54. https://doi.org/10.1038/s41438-021-00495-1 (2021).
24. Ye, G. *et al.* De novo genome assembly of the stress tolerant forest species *Casuarina equisetifolia* provides insight into secondary growth. *Plant J.* **97**, 779–794. https://doi.org/10.1111/tpj.14159 (2019).
25. Jia, H. M. *et al.* The red bayberry genome and genetic basis of sex determination. *Plant Biotechnol. J.* **17**, 397–409. https://doi.org/10.1111/pbi.12985 (2019).
26. Marrano, A. *et al.* High-quality chromosome-scale assembly of the walnut (*Juglans regia* L.) reference genome. *Gigascience* **9**(5), giaa050. https://doi.org/10.1093/gigascience/giaa050 (2020).
27. Huang, Y. *et al.* The genomes of pecan and Chinese hickory provide insights into Carya evolution and nut nutrition. *Gigascience* **8**(5), giz036. https://doi.org/10.1093/gigascience/giz036 (2019).

Cucurbitales

28. Sun, H. *et al.* Karyotype stability and unbiased fractionation in the paleo-allotetraploid Cucurbita genomes. *Mol. Plant* **10**(10), 1293–1306. https://doi.org/10.1016/j.molp.2017.09.003 (2017).
29. Barrera-Redondo, J. *et al.* The genome of Cucurbita argyrosperma (silver-seed gourd) reveals faster rates of protein-coding gene and long noncoding RNA turnover and neofunctionalization within Cucurbita. *Mol. Plant* **12**(4), 506–520. https://doi.org/10.1016/j.molp.2018.12.023 (2019).
30. Levi, A. *et al.* Sequencing the genome of the heirloom watermelon cultivar Charleston Gray. *Plant and Animal Genome Conference* 2018 (2011) .
31. Li, Z. *et al.* RNA-Seq improves annotation of protein-coding genes in the cucumber genome. *BMC Genom.* **12**, 540. https://doi.org/10.1186/1471-2164-12-540 (2011).
32. Montero-Pau, J. *et al.* De novo assembly of the zucchini genome reveals a whole-genome duplication associated with the origin of the Cucurbita genus. *Plant Biotechnol. J.* **16**(6), 1161–1171. https://doi.org/10.1111/pbi.12860 (2018).
33. Wu, S. *et al.* The bottle gourd genome provides insights into Cucurbitaceae evolution and facilitates mapping of a Papaya ring-spot virus resistance locus. *Plant J.* **92**(5), 963–975. https://doi.org/10.1111/tpj.13722 (2017).

Malpighiales

34. Xia, Z. *et al.* Chromosome-scale genome assembly provides insights into the evolution and flavor synthesis of passion fruit (Passiflora edulis Sims). *Hortic Res.* **8**(1), 14. https://doi.org/10.1038/s41438-020-00455-1 (2021).
35. Zhang, J. *et al.* Genomic comparison and population diversity analysis provide insights into the domestication and improvement of flax. *iScience* **23**(4), 100967. https://doi.org/10.1016/j.isci.2020.100967 (2020).
36. Hu, M. J. *et al.* Chromosome-scale assembly of the Kandelia obovata genome. *Hortic. Res.* **7**(1), 75. https://doi.org/10.1038/s41438-020-0300-x (2020).
37. ...An, X. *et al.* High quality haplotype-resolved genome assemblies of *Populus tomentosa* Carr., a stabilized interspecific hybrid species widespread in Asia. *Mol. Ecol. Resour.* **22**(2), 786–802. https://doi.org/10.1111/1755-0998.13507 (2022).
38. Wang, M., Gu, Z., Fu, Z. & Jiang, D. High-quality genome assembly of an important biodiesel plant, *Euphorbia lathyris* L. *DNA Res.* **28**(6), dsa022. https://doi.org/10.1093/dnares/dsab022 (2021).
39. Lu, J. *et al.* A chromosome-level assembly of a wild castor genome provides new insights into the adaptive evolution in a tropical desert. *Genom. Proteomics Bioinform.* **S1672–0229**(21), 00162–5. https://doi.org/10.1016/j.gpb.2021.04.003 (2021).

Myrtales

40. ...Healey, A. L. *et al.* Pests, diseases, and aridity have shaped the genome of *Corymbia citriodora*. *Commun. Biol.* **4**(1), 537. https://doi.org/10.1038/s42003-021-02009-0 (2021).
41. Julia, V., Mervyn, S. & Ramil, M. A high-quality draft genome for Melaleuca alternifolia (tea tree): A new platform for evolutionary genomics of myrtaceous terpene-rich species. *Gigabyte*https://doi.org/10.46471/gigabyte.28 *(2021).*
42. ...Myburg, A. A. *et al.* The genome of Eucalyptus grandis. *Nature* **510**(7505), 356–62. https://doi.org/10.1038/nature13308 (2014).
43. Luo, X. *et al.* The pomegranate (*Punica granatum* L.) draft genome dissects genetic divergence between soft- and hard-seeded cultivars. *Plant Biotechnol. J.* **18**(4), 955–968. https://doi.org/10.1111/pbi.13260 (2020).
44. Lu, R. S. *et al.* Genome sequencing and transcriptome analyses provide insights into the origin and domestication of water caltrop (Trapa spp Lythraceae). *Plant Biotechnol. J.*https://doi.org/10.1111/pbi.13758 *(2021).*

Malvales

45. Kim, Y. M. *et al.* Genome analysis of Hibiscus syriacus provides insights of polyploidization and indeterminate flowering in woody plants. *DNA Res.* **24**, 71–80 (2017).
46. Li, F. *et al.* Genome sequence of cultivated Upland cotton (*Gossypium hirsutum* TM-1) provides insights into genome evolution. *Nat. Biotechnol.* **33**, 524–30 (2015).
47. Teh, B. T. *et al.* The draft genome of tropical fruit durian (*Durio zibethinus*). *Nat. Genet.* **49**, 1633–1641 (2017).
48. Argout, X. *et al.* The genome of Theobroma cacao. *Nat. Genet.* **43**, 101–8 (2011).
49. NCBI. *Herrania umbratica* Annotation Release 100. https://www.ncbi.nlm.nih.gov/genome/annotationeuk/Herraniaumbratica/100/ (2017).
50. Islam, M. S. *et al.* Comparative genomics of two jute species and insight into fibre biogenesis. *Nat. Plants* **3**, 16223 (2017).

Sapindales

51. Lin, Y. *et al.* Supporting data for "Genome-wide sequencing of longan (*Dimocarpus longan* Lour.) provides insights into molecular basis of its polyphenol-rich characteristics." *GigaScience Data* (2017).
52. Hu, G. *et al.* Two divergent haplotypes from a highly heterozygous lychee genome point to independent domestication events for early and late-maturing cultivars. *Nat. Genet.* **54**, 73–83 (2022).
53. Yang, J. *et al.* Supporting data for "De novo genome assembly of the endangered Acer yangbiense, a plant species with extremely small populations endemic to Yunnan of China.". *GigaScience Database*https://doi.org/10.5524/100610 *(2019).*

54. Liang, Q. *et al.* Supporting data for "The genome assembly and annotation of yellowhorn (*Xanthoceras sorbifolium* Bunge)" *GigaScience Database*. https://doi.org/10.5524/100589(2019).
55. Li, W. *et al.* SMRT sequencing generates the chromosome-scale reference genome of tropical fruit mango, *Mangifera indica*. *bioRxiv* **15**, 4. https://doi.org/10.1101/2020.02.22.960880 (2020).
56. Grattapaglia, D. & Silva, O. *Anacardium occidentale* v0.9. Phytozome 13, https://phytozome-next.jgi.doe.gov/info/Aoccidentale v0.9 (2021).
57. Zeng, L. *et al.* Whole genomes and transcriptomes reveal adaptation and domestication of pistachio. *Genome Biol.* **20**(1), 79. https://doi.org/10.1186/s13059-019-1686-3 (2019).
58. Xu, Q. *et al.* The draft genome of sweet orange (*Citrus sinensis*). *Nat. Genet.* **45**, 59–66. https://doi.org/10.1038/ng.2472 (2013).
59. Wu, G. A. *et al.* Sequencing of diverse mandarin, pummelo and orange genomes reveals complex history of admixture during citrus domestication. *Nat. Biotechnol.* **32**, 656–662. https://doi.org/10.1038/nbt.2906 (2014).
60. Muellner-Riehl, A. N. *et al.* Molecular phylogenetics and molecular clock dating of Sapindales based on plastid rbcL, atpB and trnL-trnF DNA sequences. *Taxon* **65**, 1019–1036. https://doi.org/10.12705/655.5 (2016).
61. Wannan, B. S. Analysis of generic relationships in Anacardiaceae. *Blumea* **51**, 165–195 (2006).

Asterales

62. Wu, Z. *et al.* The chromosome-scale reference genome of safflower (Carthamus tinctorius) provides insights into linoleic acid and flavonoid biosynthesis. *Plant Biotechnol. J.* https://safflower.scuec.edu.cn/download.html (2021).
63. Wen, X. *et al.* The Chrysanthemum lavandulifolium genome and the molecular mechanism underlying diverse capitulum types. *Hortic. Res.* **18**, uhab022 (2022).
64. Kim, J. *et al.* Whole-genome, transcriptome, and methylome analyses provide insights into the evolution of platycoside biosynthesis in Platycodon grandiflorus a medicinal plant. *Hortic. Res.* **7**, 112 (2020).
65. Bellinger, R. M. Beggartick: A genome for Bidens hawaiensis: A member of a hexaploid Hawaiian plant adaptive radiation. *J. Hered.* **113**, 205–214 (2022).
66. Reyes-Chin-Wo, S. *et al.* Lettuce: Genome assembly with in vitro proximity ligation data and whole-genome triplication in lettuce. *Nat. Commun.* **8**, 14953 (2017).
67. Lin, T. *et al.* Genome ID 28333 Dandelion: Extensive sequence divergence between the reference genomes of Taraxacum kok-saghyz and Taraxacum mongolicum. *Sci. China Life Sci.* **65**, 515–528 (2021).
68. Xu, X. *et al.* The chromosome-level Stevia genome provides insights into steviol glycoside biosynthesis. *Hortic. Res.* **8**, 129 (2021).

Gentianales

69. Arabica Genome
70. Hoopes, G. M. *et al.* Genome assembly and annotation of the medicinal plant Calotropis gigantea, a producer of anti-cancer and anti-malarial cardenolides. *G3* **8**(2), 385–391 (2018).
71. Hao, X. *et al.* Chromosome-level assembly of Neolamarckia cadamba genome provides insights into the evolution of cadambine biosynthesis. *Plant J.* **109**, 891–908 (2021).
72. Brose, J. *et al.* The Mitragyna speciosa (Kratom) Genome: A resource for data-mining potent pharmaceuticals that impact human health. *G3* **2**, 058 (2021).
73. Liu, Y. *et al.* Whole-genome sequencing and analysis of the Chinese herbal plant Gelsemium elegans. *Acta Pharm. Sin. B* **10**, 374–382 (2019).
74. Rai, A. *et al.* Chromosome-level genome assembly of Ophiorrhiza pumila reveals the evolution of camptothecin biosynthesis. *Nat. Commun.* **12**(1), 405 (2021).
75. Xu, Z. *et al.* Tandem gene duplications drive divergent evolution of caffeine and crocin biosynthetic pathways in plants. *BMC Biol.* **18**, 63 (2020).

Lamiales

76. Zhao, D. *et al.* A chromosomal-scale genome assembly of Tectona grandis reveals the importance of tandem gene duplication and enables discovery of genes in natural product biosynthetic pathways. *Gigascience* **8**(3), 55706 (2019).
77. Jia, K. H. *et al.* Chromosome-scale assembly and evolution of the tetraploid Salvia splendens (Lamiaceae) genome. *Hortic. Res.* **8**(1), 177 (2021).
78. Hu, Y. *et al.* High-quality genome of the medicinal plant Strobilanthes cusia provides insights into the biosynthesis of indole alkaloids. *Front. Plant Sci.* **12**, 7424240 (2021).
79. Natarajan, P. *et al.* A reference-grade genome identifies salt-tolerance genes from the salt-secreting mangrove species Avicennia marina. *Commun. Biol.* **4**(1), 851 (2021).
80. Wang, M. *et al.* Chromosomal-level reference genome of the neotropical tree Jacaranda mimosifolia D. Don. *Genome Biol. Evol.* **3**, evab094 (2021).
81. Ma, Y. P. *et al.* Genome-wide analysis of butterfly bush (Buddleja alternifolia) in three uplands provides insights into biogeography, demography and speciation. *New Phytol.* **232**, 1463–1476 (2021).
82. Yang, X. *et al.* The chromosome-level quality genome provides insights into the evolution of the biosynthesis genes for aroma compounds of Osmanthus fragrans. *Hortic. Res.* **5**, 72 (2018).

Solanales

83. Su, X. *et al.* A high-continuity and annotated tomato reference genome. *BMC Genom.* **22**(1), 898 (2021).
84. Wei, Q. *et al.* A high-quality chromosome-level genome assembly reveals genetics for important traits in eggplant. *Hortic Res.* **7**, 153 (2020).
85. Kim, S. *et al.* New reference genome sequences of hot pepper reveal the massive evolution of plant disease-resistance genes by retroduplication. *Genome Biol.* **18**(1), 210 (2017).
86. Sierro, N. *et al.* The impact of genome evolution on the allotetraploid Nicotiana rustica: An intriguing story of enhanced alkaloid production. *BMC Genom.* **19**(1), 855 (2018).
87. Bombarely, A. *et al.* Insight into the evolution of the Solanaceae from the parental genomes of Petunia hybrida. *Nat. Plants* **2**, 16074 (2016).
88. Hao, Y. *et al.* The chromosome-based genome provides insights into the evolution in water spinach. *Sci. Hortic.* **289**, 110501 (2021).

Ericales

89. Wu, H. *et al.* A chromosome-level genome assembly for the wild kiwifruit Actinidia kolomikta provides insights into canker resistance and fruit development. *Plant Biotechnol. J.* **2021**, 13748 (2021).
90. Xia, E. *et al.* The tea plant reference genome and improved gene annotation using long-read and paired-end sequencing data. *Sci. Data* **6**(1), 122 (2019).

91. Kawash, J. *et al.* Contrasting a reference cranberry genome to a crop wild relative provides insights into adaptation, domestication, and breeding. *PLoS ONE* **17**(3), e0264966 (2022).
92. Suo, Y. *et al.* A high-quality chromosomal genome assembly of Diospyros oleifera Cheng. *Gigascience* **9**(1), 62597 (2020).
93. Potent, G. *et al.* Comparative genomics elucidates the origin of a supergene controlling floral heteromorphism. *Mol. Biol. Evol.* **39**, msac035 (2022).
94. Hartmann, S. *et al.* Annotated genome sequences of the carnivorous plant Roridula gorgonias and a non-carnivorous relative, Clethra arborea. *BMC Res. Notes* **13**(1), 426 (2020).

Metazoa

95. Simakov, O. *et al.* Deeply conserved synteny resolves early events in vertebrate evolution. *Nat. Ecol. Evol.* **4**(6), 820–830. https://doi.org/10.1038/s41559-020-1156-z (2020).
96. Kenny, N. J. *et al.* Tracing animal genomic evolution with the chromosomal-level assembly of the freshwater sponge Ephydatia muelleri. *Nat. Commun.* **11**(1), 3676. https://doi.org/10.1038/s41467-020-17397-w (2020).
97. Li, F. *et al.* Chromosome-level genome assembly of the East Asian common octopus (Octopus sinensis) using PacBio sequencing and Hi-C technology. *Mol. Ecol. Resour.* **20**(6), 1572–1582. https://doi.org/10.1111/1755-0998.13216 (2020).
98. Fuller, Z. L. *et al.* Population genetics of the coral Acropora millepora: Toward genomic prediction of bleaching. *Science* **369**(6501), eaba4674 (2020).
99. ...Srivastava, M. *et al.* The Trichoplax genome and the nature of placozoans. *Nature* **454**(7207), 955–60. https://doi.org/10.1038/nature07191 (2008).
100. Simakov, O. *et al.* Deeply conserved synteny and the evolution of metazoan chromosomes. *Sci. Adv.* **8**(5), eabi5884. https://doi.org/10.1126/sciadv.abi5884 (2022).

Author contributions

Q.X., L.J., J.L. and D.S. conceived the research, Q.X. carried out the computations, C.Z. obtained and curated the genome data, X.Z. validated the phylogenies, and Q.X., L.J., J.L. and D.S. wrote the paper. All authors reviewed the manuscript.

Competing interests

The authors declare no competing interests.

Additional information

Supplementary Information The online version contains supplementary material available at https://doi.org/10.1038/s41598-023-33029-x.

Correspondence and requests for materials should be addressed to D.S.

Reprints and permissions information is available at www.nature.com/reprints.

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Chapter 6

A New Direction: the Large Phylogeny Problem

Qiaoji Xu and David Sankoff. Gene order phylogeny via ancestral genome reconstruction under dollo. Editors: Katharina Jahn, Tomas Vinar, 2023.

Gene order phylogeny via ancestral genome reconstruction under Dollo

Qiaoji Xu[1] and David Sankoff[1]

Abstract. We present a proof of principle for an new kind of stepwise algorithm for unrooted binary gene-order phylogenies. This method incorporates a simple look-ahead inspired by Dollo's law, while simultaneously reconstructing each ancestor (HTU). We first present a generic version of the algorithm illustrating a necessary consequence of Dollo characters. In a concrete application we use generalized oriented gene adjacencies and maximum weight matching (MWM) to reconstruct fragments of monoploid ancestral genomes as HTUs. This is applied to three flowering plant orders while estimating phylogenies for these orders in the process. We discuss how to improve on the extensive computing times that would be necessary for this method to handle larger trees.

Keywords: large phylogeny problem · inferring gene-order HTUs · Dollo's law · gene proximities · maximum weight matching · plant orders.

1 Introduction

In formal phylogenetics, Dollo models postulate that a character can only appear once in the course of evolution, although it may disappear from several descendent lineages. This idea has a number of combinatorial consequences, and suggests an algorithmic inference of phylogeny that differs significantly from standard approaches. In this paper, we present a proof of principle for such an algorithm in the context of unrooted binary trees.

Our approach is basically hierarchical, agglomerative, combining pairs of input taxa or already constructed subtrees but, crucially, does make use, via Dollo, of a limited amount of information from outside these pairs.

The **input** to the generic version of our algorithm consists of n taxa, "OTUs", corresponding to the terminal vertices of the phylogeny to be constructed, each of which represented by a set of distinct characters, where the sets may overlap with each other to varying degrees. The strategy is to construct $n-2$ **output** sets, each containing some of the same characters, corresponding to the non-terminal vertices, "HTUs", of a binary tree constructed at the same time, such that the Dollo condition, or at least some of its important consequences, is maintained.

In contrast to the generic version of the algorithm, for any specific formulation of a Dollo model, the algorithm must be modified so that the output sets conform to the requirements of the problem at hand. These modifications generally mean that Dollo is no longer a sufficient condition, but remains a necessary condition for a character to be included in an output set.

In the main part of this paper, the elements of an input set are all the proximities between oriented genes in one genome. The elements of each output set are chosen from a very large number of proximities, namely those satisfying the Dollo condition, but narrowed down to include only those consistent with an optimal linear ordering, i.e. fragments of chromosome. We have previously invoked these concepts in studying the "small phylogeny problem", where the tree topology is given (e.g, [1–5]), but here we concentrate on the "large phylogeny problem", where we actually construct this phylogeny.

We apply our method to three plant orders to compare the results with known phylogenies, with almost total agreement.

Finally we discuss possible improvements in efficiency ad extensions beyond unrooted binary branching phylogenies.

2 Dollo's law in the context of unrooted binary trees

The idea that a character is gained only one time and can never be regained if it is lost is realized in an unrooted tree by the property that the set of vertices containing the character are connected. This is a necessary and sufficient condition, valid both for terminal vertices (or degree 1) and internal (ancestral) ones (degree 3 in an unrooted binary tree).

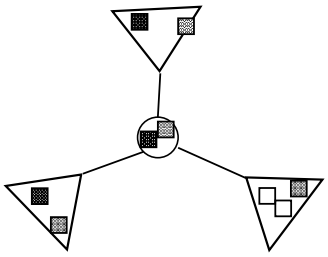


Fig. 1. Necessary condition for characters to appear at internal vertex of binary branching tree. Dark shaded character (small square) appears in all three trees (triangles) subtended by the internal vertex (circle). Light shaded character appears in only two of the trees. Unshaded character appears in only one subtree so does not affect internal vertex.

The connectedness condition can be satisfied by a set of non-terminal nodes of a tree, but for phylogenetic inference, we require that a character be present in the input set of at least two terminal vertices; otherwise it could not be necessary for any of the output sets at the non-terminal vertices.

In an unrooted binary branching tree, each non-terminal vertex subtends three subtrees, as in Figure 1. For a data set to be used in constructing a phylogenetic tree and the output sets, clearly each character must be present in at

least two of the three subtrees, as illustrated in the figure. More precisely, each character must be present at least in one terminal vertax set in at least two of the three subtrees.

3 Generic Algorithm

1. input n sets of characters
2. $i = 1$
3. all n sets are "eligible" vertices
4. while $i \leq n - 3$
   (a) *for each pair (G, H) of the $m = n - i + 1$ eligible vertices calculate a potential ancestral vertex A containing all phylogenetically validated characters (in both G and H, or in one of G or H plus any other eligible vertex.
   (b) pick the pair (G', H') with maximum total number of the characters in their potential ancestor A.
   (c) ancestor vertex A becomes eligible, and the other two, G' and H', become ineligible
   (d) two edges of the tree are defined: AG' and AH'
   (e) $i = i + 1$
5. Now $i = n - 3$, so that there are three eligible vertices G', H', K'. Define three edges of the tree : AG', AH' and AK'
6. calculate ancestral vertex A containing all phylogenetically validated characters (in any two or all three of G', H' and K')
7. output all $2n - 3$ tree edges and all $n - 2$ selected output (ancestral, or "HTU") sets.

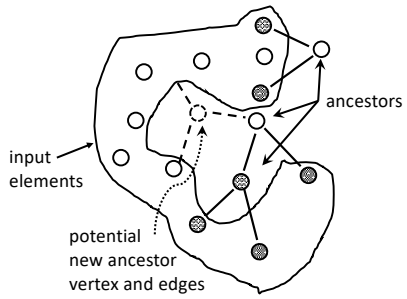


Fig. 2. Ineligible vertices shaded.

The asterisked step 4a may be interpreted in at least two different ways. In requiring that a character be present in a subtree, we may mean that that

character be present in some input set associated with a terminal vertex of that subtree, or we may require something stronger, that the character be present in the eligible (ancestral) vertex of that subtree.

In constructing a potential ancestor A of two input sets G and H, we define a **Dollo** edge for a character to be one in either G or H, but not both, as well as in some eligible vertex other than G or H.

Our sketch of the generic algorithm is meant to illustrate how the Dollo principle allows a kind of look-ahead in the hierarchical construction of the phylogeny. Without some modification, however, the algorithm can result in some counter-intuitive results.

Consider the input sets (1,2),(3,4),(1,3),(2,4), where all four characters are also in other sets. The largest potential ancestor (1,2,3,4) is constructed by pairing (1,2) with (3,4). Note that this ancestor is based entirely on Dollo edges.

The grouping of (1,2) with (3,4) in this construction is not intuitively satisfying from the phylogenetic viewpoint since these two input sets have nothing in common. This suggests down-weighting the Dollo edges in favour of the potential edges when analyzing a character. For example, if we assign weight 1/2 to Dollo edges compared to weight 2 to a character in both G and H, the potential ancestor of (1,2) and (3,4) only has weight 2, while the ancestor of (1,2) and (1,3) has weight 3. It is the latter that will be chosen by the algorithm if we replace "maximium number" by "maximum weight".

A principled way of assigning weights might be $2 + \alpha q/m$ for characters in both G and H plus any other eligible vertex, and $1 + \beta q/m$ for a character in one of G or H plus any other eligible vertex, where q is the number of other eligible vertices out of m containing the character, and $2\beta < \alpha < 1$.

4 Genomics case: generalized adjacencies of genes

4.1 Algorithm: ancestor via Maximum Weight Matching

1. input n extant genomes
2. $i = 1$
3. all n genomes are "eligible"
4. while $i \leq n - 3$
   (a) for all pairs (G, H) of the $m - i + 1$ eligible vertices find potential ancestral genome A as the Maximum Weight Matching of phylogenetically validated generalized adjacencies (in both G and H, or in one of G or H plus any other eligible vertex).
   (b) pick the pair (G', H') with the highest Maximum Weight Matching score.
   (c) ancestor genome A becomes eligible, and the other two, G' and H' become ineligible.
   (d) two edges of the tree are defined: AG' and AH'
   (e) $i = i + 1$
5. Now $i = n - 3$, so that there are three eligible vertices G', H', K'. Define three edges of the tree : AG', AH' and AK'

6. calculate ancestral vertex A containing all phylogenetically validated adjacencies (in any two or all three of G', H' and K')

A down-weighting scheme for Dollo edges may also be adopted here, although we do not consider that further here.

5 Application to phylogenies of three plant orders

Detailed references to all the genomes mentioned here are given in reference [5], including access codes for the CDS files of the genomes we use in the CoGe platform [6, 7].

The phylogenies that serve as validation of our constructs are by and large uncontroversial, based on up-to-date sources, mainly [8–10]

5.1 Asterales

From the family Asteraceae, we used the published genomes of safflower (*Carthamus tinctorius*) and artichoke (*Cynara cardunculu*) from the subfamily Carduoideae, lettuce (*Lactuca sativa*) and dandelion (*Taraxacum mongolicum*) from the subfamily Cichorioideae, and *Mikania micrantha* and *Stevia rebaudian* from the subfamily Asteroideae.

Figure 3 and Table 1 show that our method partitioned the six genomes correctly into three groups.

Fig. 3. Partial phylogeny of the order Asterales (family Asteraceae)as correctly reconstructed by our algorithm, with haploid numbers of chromosomes. Labels on interior nodes indicate the algorithm steps at which they were created (cf Table 1).

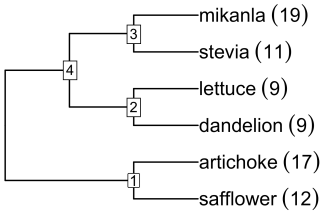


Table 1. Steps in searching for Asterales ancestors

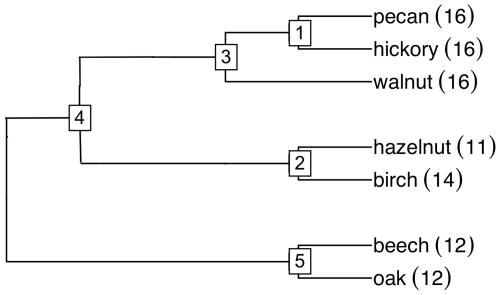
potential sisters	MWM for potential anc 1	MWM for pot. anc 2	MWM for potential. anc 3	anc 4
Safflower Artichoke	**1844325**			
Lettuce Dandelion	1767015	**1711720**		
Stevia Mikania	1677330	1648609	**1537870**	
anc3_anc2_anc1				**1322071**
Lettuce Safflower	1752682			
Lettuce Stevia	1663257	1646505		
Lettuce Artichoke	1814764			
Lettuce Mikania	1588424	1576347		
Safflower Stevia	1631338			
Safflower Dandelion	1683778			
Safflower Mikania	1565038			
Stevia Dandelion	1590905	1581354		
Stevia Artichoke	1685538			
Dandelion Artichoke	1741613			
Dandelion Mikania	1525367	1518558		
Artichoke Mikania	1612480			
anc1_Lettuce		1644223		
anc1_Stevia		1494957	1467125	
anc1_Dandelion		1573684		
anc1_Mikania		1414983	1399206	
anc2_Stevia			1468319	
anc2_Mikania			1395159	
anc2_anc1			1374650	

5.2 Fagales

From the order Fagales, we used the published genomes of oak (*Quercus robur*) and beech (*Fagus sylvatica*) from the family Fagaceae, birch (*Betula platyphylla*) and hazelnut (*Corylus mandshurica*) from the family Betulaceae, and walnut (*Juglans regia*), pecan (*Carya illinoinensis*) and hickory (*Carya cathayensis*) from the family Juglandaceae.

Figure 4 and Table 2 show that our method partitioned the seven genomes correctly into three families.

Fig. 4. Partial phylogeny of the order Fagales, as correctly reconstructed by our algorithm, with haploid numbers of chromosomes. Labels on interior nodes indicate the algorithm steps at which they were created (cf Table 2).



5.3 Sapindales

While the results on the Asterales and Fagales are heartening, we cannot expect the method to always perform as well. This is illustrated by an analysis of the order Sapindales. From this order, we used the published genomes of cashew (*Anacardium occidentale*), mango (*Mangifera indica*) and pistachio (*Pistacia vera*) from the family Anacardiaceae and maple (*Acer catalpifolium*), longan (*Dimocarpus longan*), lychee (*Litchi chinensis*) and yellowhorn (*Xanthoceras sorbifoli*) from the family Sapindaceae.
As seen in Figure 5, except for the incorrect placement of yellowthorn, the method separates the two families as expected. Were this genome to be removed, the reconstructed tree would be identical to the known tree, aside from a permutation of the species within the Sapindaceae. Indeed, Table 3 shows that at the fourth step in the algorithm, yellowthorn was almost assigned to join the other Sapindaceae.

Fig. 5. Partial phylogeny of the order Sapindaless with haploid numbers of chromosomes. Left is the known phylogeny and right is the reconstructed one. Labels on interior nodes at right indicate the algorithm steps at which they were created.

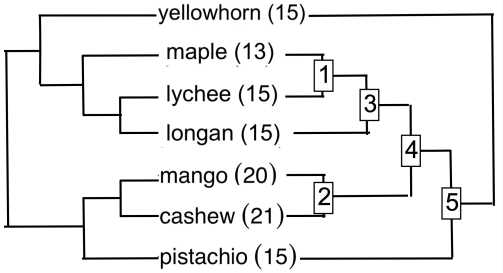


Table 2. Steps in searching for Fagales ancestors

potential sisters	MWM for potential anc 1	MWM for pot. anc 2	MWM for pot. anc 3	MWM for pot.anc 4	ancestor 5
Hickory Pecan	**2280458**				
Birch Hazelnut	2133010	**2130194**			
anc1_Walnut		1938190	**1929418**		
anc3_anc2				**1744607**	
anc4_Oak_Beech					**1502714**
Walnut Birch	1780904	1777540			
Walnut Oak	1432598	1414818	1399566		
Walnut Hickory	2138055				
Walnut Pecan	2048931				
Walnut Beech	1662931	1654141	1623485		
Walnut Hazelnut	2026641	2013614			
Birch Oak	1474374	1481635			
Birch Hickory	1980129				
Birch Pecan	1892718				
Birch Beech	1735741	1744305			
Oak Hickory	1612423				
Oak Pecan	1534218				
Oak Beech	1390113	1394735	1366291	1332320	
Oak Hazelnut	1675695	1677963			
Hickory Beech	1851957				
Hickory Hazelnut	2244383				
Pecan Beech	1772551				
Pecan Hazelnut	2148460				
Beech Hazelnut	1951082	1952202			
anc1_Oak		1443626	1460361		
anc1_Beech		1704015	1711078		
anc1_Hazelnut		2108715			
anc1_Birch		1844003			
anc2_Walnut			1817364		
anc2_Oak			1473713	1487666	
anc2_Beech			1743668	1737292	
anc2_anc1			1849853		
anc3_Oak				1410577	
anc3_Beech				1650427	

Table 3. Steps in searching for Sapindales ancestors

potential sisters	MWM for potential anc 1	MWM for pot. anc 2	MWM for pot. anc 3	MWM for pot.anc 4	ancestor 5
Cashew Mango	**2114258**				
Lychee Maple	1909424	**1915717**			
anc2_Longan			**1881830**		
anc3_anc1				**1812017**	
anc4_Yellowhorn_Pistachio					**177012**
Lychee Longan	1839623	1847180			
Lychee Yellowhorn	1792086	1718022			
Lychee Cashew	1715727				
Lychee Pistachio	1881466	1885124			
Lychee Mango	1899616				
Longan Maple	1741981	1750100			
Longan Yellowhorn	1775908	1568607	1561594		
Longan Cashew	1566047				
Longan Pistachio	1707939	1713967	1699756		
Longan Mango	1729214				
Maple Yellowhorn	1892491	1652072			
Maple Cashew	1652307				
Maple Pistachio	1821498	1820282			
Maple Mango	1847530				
Yellowhorn Cashew	1840431				
Yellowhorn Pistachio	1966295	1689715	1694495	1511822	
Yellowhorn Mango	1945493				
Cashew Pistachio	1702324				
Pistachio Mango	1920295				
anc1_Lychee		1795812			
anc1_Longan		1621378	1634641		
anc1_Maple		1734318			
anc1_Yellowhorn		1651337	1678486	1636597	
anc1_Pistachio		1814383	1830387	1656675	
anc2_Yellowhorn			1696079		
anc2_Pistachio			1803140		
anc2_anc1			1739109		
anc3_Yellowhorn				1802826	
anc3_Pistachio				1754820	

6 Discussion and Conclusions

The algorithmic reconstruction of ancient gene orders, and associated phylogenies, has a history of over three decades (e.g., [11–14]). The present work differs from all these in that it is situated in a paradigm [1–5] where only the strictly monoploid ancestors in a phylogeny are reconstructed, or strictly linear chromosomal fragments, whether or not such genomes actually existed or are just

inherent in the basic chromosomal organization within the possibly re-occurring polyploid history of the group.

Although our approach builds a hierarchy by combining pairs of OTUs or already constructed subtrees, it is unusual that it does make use, via Dollo, of a limited amount of information from outside these pairs. This is in effect a very partial look-ahead.

The goal of this paper was to present evidence that our method can construct accurate or plausible phylogenies, based entirely on comparative gene order. We were not preoccupied with questions of computing time. Indeed, since our reconstruction is based on maximum weight matching (MWM) software on very large graphs, it is bound to be computationally expensive. Moreover, since the current experimental version uses MWM to exhaustively evaluate all possibilities separately at each step in building a hierarchy, only a moderate number of OTUs can be input. There are, however, many possibilities to improving the efficiency, by constraining the search space, by branch and bound tehniques, by saving a certain number of partial solutions in parallel, and other techniques.

References

1. Xu Q, Jin L, Zheng C, Leebens-Mack JH, Sankoff D (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.
2. Xu Q, Jin L, Leebens-Mack JH, Sankoff D. (2021) Validation of automated chromosome recovery in the reconstruction of ancestral gene order. *Algorithms* **14**(6):160.
3. Chanderbali, A.S., Jin, L., Xu, Q. et al. Buxus and Tetracentron genomes help resolve eudicot genome history. *Nature Commununications* 13, 643 (2022). https://doi.org/10.1038/s41467-022-28312-w
4. Xu Q, Jin L, Zhang Y, Zhang X, Zheng C, Leebens-Mack JH, Sankoff D. Ancestral flowering plant chromosomes and gene orders based on generalized adjacencies and chromosomal gene co-occurrences. *Journal of Computational Biology* 2021 Nov 1;28(11):1156-79.
5. Xu Q, Jin L, Zheng C, Zhang X, Leebens-Mack J, Sankoff D. (2022) From comparative gene content and gene order to ancestral contigs, chromosomes and karyotypes. *bioRxiv* 2022.09.28.509880.
6. Lyons E and Freeling M (2008) How to usefully compare homologous plant genes and chromosomes as DNA sequences. *The Plant Journal* **53**: 661–673.
7. Lyons E, Pedersen B, Kane J, Alam M, Ming R, Tang H, Wang X, Bowers J, Paterson A, Lisch D, Freeling M (2008) Finding and comparing syntenic regions among *Arabidopsis* and the outgroups papaya, poplar and grape: CoGe with rosids. *Plant Physiology* **148**: 1772–1781.
8. Published Plant Genomes (2022) Usadel lab, Forschungszentrum Jülich, Heinrich Heine University., Düsseldorf. https://www.plabipd.de/
9. Stevens PF (2017) Angiosperm Phylogeny Website. Version 14. http://www.mobot.org/MOBOT/research/APweb/.
10. Chase MW, Christenhusz MJM, Fay MF, Byng JW, Judd WS, Soltis DE, Mabberley DE, Sennikov AN, Soltis PS, Stevens PF (2016) An update of the Angiosperm Phylogeny Group classification for the orders and families of flowering plants: APG IV. *Botanical Journal of the Linnean Society* **181**:1–20.

11. Sankoff D, Leduc G, Antoine N, Paquin B, Lang BF, Cedergren R (1992) Gene order comparisons for phylogenetic inference: evolution of the mitochondrial genome. *Proceedings of the National Academy of Sciences* **15**:6575-9.
12. Moret BM, Warnow T (2005) Advances in phylogeny reconstruction from gene order and content data. *Methods in enzymology* **395**:673-700.
13. Hu F, Lin Y, Tang J (2014) MLGO: phylogeny reconstruction and ancestral inference from gene-order data. *BMC bioinformatics* **15**:1-6.
14. Perrin A, Varré JS, Blanquart S, Ouangraoua A (2015) ProCARs: Progressive Reconstruction of Ancestral Gene Orders *BMC Genomics* **16** S5:S6.

Chapter 7

Conclusion

As mentioned in the Introduction, at the outset of the research presented in this thesis, it was not clear that there would be sufficient evolutionary signal in the extant genomes to make plausible reconstructions of the ancestors. This worry was assuaged as the work progressed. Several results lend confidence that we were on the right track:

- the output from the MWM took the form of substantial linear contigs rather many isolated adjacencies that did not fit together,
- the clusters underlying the final ordered chromosomes in the *Buxus* and monocot studies were extraordinarily well-defined with very little overlap, as was immediately visible from the heat maps,
- the results from the simulation studies reported in Chapter 3 were clearly confirmatory of the accuracy of the method,
- the remarkably consistent results on base chromosome number in Chapter 5 showed that the evolution of the plant orders represented nature's replicate experiment eleven times.

Each chapter in this thesis has both technical and biological interest. In Chapter 2, aside its main contribution of laying out the pipeline Raccroche for the first time in some detail, we were able to locate the important "tau" duplication in the phylogeny, with the help of the MCScan visualization [28]

In Chapter 3 we simulated the multiple tetraploidizations we previously inferred throughout the monocot phylogeny, as well as the extensive chromosomal rearrangements, massive deletion of duplicate genes, and numerous insertions, with randomization introduced at each step. Applying Raccroche to this very close simulation of the evolution of the extant genomes proved that the simulated version of the original

ancestor could be reconstructed chromosome by chromosome, with impressive accuracy at the level of gene content. This chapter also showed that RACCROCHE was superior to proCARS [26], at least with respect to total gene content recovered.

In Chapter 4, along with our results on the common origin of the *Buxus* and *Tetracentron* orders, and other biologically important results, we showed how to assign functional categories to the genes (actually gene labels) and these are accessible from the paper.

The preceding three chapters were each inspired by data from genomes dispersed across several orders. Further work within individual orders or families, however, showed that while MWM produced longer contigs, the clustering steps produced noisy heat maps and ambiguity about the number clusters/chromosomes in the ancestors. Chapter 5 details new techniques to overcome these problems:

- 100 samples of MWM solutions to search for the clearest cluster analysis,
- the “g-mer" method to remove the long contig bias,
- gap statistics to evaluate the significance of each cluster, and
- techniques for identifying the gap statistic inflection point.

As a result we were able to reconstruct ancestors for 11 of the rosid/asterid orders based on six or more genomes in each spread across several families. Each of these had the same basic chromosome number (nine). Reconstructing a common chromosome number must be considered a validation, since it is the only reconstruction method to even attempt to estimate this number for any ancestor. Non-genomic methods have previously estimated this number as seven or eight.

The RACCROCHE approach was designed for the small phylogeny problem. It may not have seemed a good way to approach the large phylogeny problem using a hierarchical clustering algorithm, since the reconstruction targets each ancestor separately based on the entire set of extant genomes. There is no other built-in connection between ancestor and immediate descendants or between sister descendants. Nevertheless, in Chapter 6, we found a way to relax the notion of phylogenetic validity, which enabled a tree toplogy to be built using a hierarchy-like algorithm. It was not obvious that this would work, given some small counter-examples, but when applied to three plant orders, it worked very well.

Taken together, these five chapters present innovative methods for genome reconstruction, ancestral genome annotation, and phylogenetic analysis. They demonstrate the importance of considering the unique challenges posed by plant comparative genomics, and how computational approaches can be tailored to address these challenges. The findings from these papers will pave the way for further research into the evolution of the chromosomal structures of plants and other organisms.

A major theme that recurs in all this work is the interpretation of the reconstructed monoploid genome. We did not realize the importance of this question until after publishing the monocot and *Buxus* work, and discussed it in another paper [10]. The latter paper is not included as a chapter in this thesis because of overlap with previous papers and Chapter 5 In many cases, like the *Buxus* ancestor in Chapter 4, the earliest monocot ancestor in Chapters 2 and 3, and the founding ancestor of each of the rosid and asterid orders in Chapter 5, the reconstructed monoploid can be understood as a close approximation of some real genome that existed at certain point of time. In other cases, we may know that a more recent ancestor must be polyploid, but our reconstructed monoploid retains the same basic structure as the original ancestor, except that duplicate chromosomal segments are simply collapsed together. Why the original chromosome number is maintained despite all the rearrangements, is a question for further research.

Availability The code for the RACCROCHE pipeline is available at github. Thanks to Alexander Liu for his hard work on the current version of our pipeline.

Bibliography

[1] Angiosperm Phylogeny Group, Mark W Chase, Maarten JM Christenhusz, Michael F Fay, JW Byng, WS Judd, DE Soltis, DJ Mabberley, AN Sennikov, PS Soltis, et al. An update of the angiosperm phylogeny group classification for the orders and families of flowering plants: APG IV. *Botanical journal of the Linnean Society*, 181(1):1–20, 2016.

[2] Robert R. Sokal and Peter H. A. Sneath. *Principles of Numerical Taxonomy*. W.H. Freeman, 1963.

[3] Farris JS. On the use of the parsimony criterion for inferring evolutionary trees. *Syst. Zool*, 22:250–256, 1973.

[4] E. Zuckerkandl and L Pauling. Evolutionary divergence and convergence in proteins. In *Evolving Genes and Proteins*, pages 97–166. 1965.

[5] John Yin, Chao Zhang, and Siavash Mirarab. ASTRAL-MP: scaling ASTRAL to very large datasets using randomization and parallelization. *Bioinformatics*, 35:3961–3969, 10 2019.

[6] Remco Bouckaert, Joseph Heled, Denise Kühnert, Timothy Vaughan, Chieh-Hsi Wu, Dong Xie, Marc A Suchard, Andrew Rambaut, and Alexei J Drummond. BEAST 2: a software platform for bayesian evolutionary analysis. *PLoS Comput Biol*, 10(4):e1003537, 2014.

[7] Eric Lyons and Michael Freeling. How to usefully compare homologous plant genes and chromosomes as dna sequences. *The Plant Journal*, 53:661–673, 2008.

[8] Chunfang Zheng, Eric Chen, Victor A Albert, Eric Lyons, and David Sankoff. Ancient eudicot hexaploidy meets ancestral eurosid gene order. *BMC genomics*, 14(S7):S3, 2013.

[9] V. Grant. *The origin of adaptations*. Columbia University Press, New York & London, 1963.

[10] Q Xu, L Jin, Y Zhang, X Zhang, C Zheng, JH Leebens-Mack, and D Sankoff. Ancestral flowering plant chromosomes and gene orders based on generalized adjacencies and chromosomal gene co-occurrences. *Lecture Notes in Computer Science*, 2020.

[11] Qiaoji Xu, Lingling Jin, Chunfang Zheng, James H. Leebens-Mack, and David Sankoff. Validation of automated chromosome recovery in the reconstruction of ancestral gene order. *Algorithms*, 14, 2021.

[12] Andre S. Chanderbali, Lingling Jin, Qiaoji Xu, and et al. Buxus and tetracentron genomes help resolve eudicot genome history. *Nat Commun*, 13:643, 2022.

[13] Qiaoji Xu, Lingling Jin, Chunfang Zheng, and et al. From comparative gene content and gene order to ancestral contigs, chromosomes and karyotypes. *Sci Rep*, 13:6095, 2023.

[14] Qiaoji Xu and David Sankoff. Gene order phylogeny via ancestral genome reconstruction under dollo. Editors: Katharina Jahn, Tomas Vinar, 2023. In press.

[15] T. Dobzhansky and A.H. Sturtevant. Inversions in the chromosomes of *Drosophila Pseudoobscura*. *Genetics*, 23(1):28–64, Jan 1938.

[16] F. Murat, A. Armero, C. Pont, C. Klopp, and J. Salse. Reconstructing the genome of the most recent common ancestor of flowering plants. *Nature Genetics*, 49:490–496, 2017.

[17] Hélène Badouin, Jérôme Gouzy, Christopher J Grassa, Florent Murat, S Evan Staton, Ludovic Cottret, Christine Lelandais-Brière, Gregory L Owens, Sébastien Carrère, Baptiste Mayjonade, et al. The sunflower genome provides insights into oil metabolism, flowering and asterid evolution. *Nature*, 546(7656):148–152, 2017.

[18] Diego P Rubert, Fábio V Martinez, Jens Stoye, and Daniel Doerr. Analysis of local genome rearrangement improves resolution of ancestral genomic maps in plants. *BMC genomics*, 21:1–11, 2020.

[19] Pavel Avdeyev, Shuai Jiang, Sergey Aganezov, Fei Hu, and Max A. Alekseyev. Reconstruction of ancestral genomes in presence of gene gain and loss. *Journal of Computational Biology*, 23(3):150–164, Mar 2016.

[20] Bernard M.E. Moret, Jijun Tang, Li-San Wang, and Tandy Warnow. Steps toward accurate reconstructions of phylogenies from gene-order data. *Journal of Computer and System Sciences*, 65(3):508–525, 2002.

[21] C. Zheng and D. Sankoff. On the pathgroups approach to rapid small phylogeny. *BMC Bioinformatics*, 12:4–10, 2011.

[22] Bernard M. E. Moret, Yu Lin, and Jijun Tang. Rearrangements in phylogenetic inference: Compare, model, or encode? In *Models and Algorithms for Genome Evolution*. Springer, 2013.

[23] Guénola Drillon, Raphaël Champeimont, Francesco Oteri, Gilles Fischer, and Alessandra Carbone. Phylogenetic reconstruction based on synteny block and gene adjacencies. *Molecular Biology and Evolution*, 37(9):2747–2762, 2020.

[24] Jian Ma, Louxin Zhang, Bernard B Suh, Brian J Raney, Richard C Burhans, W James Kent, Mathieu Blanchette, David Haussler, and Webb Miller. Reconstructing contiguous regions of an ancestral genome. *Genome research*, 16(12):1557–1565, 2006.

[25] Yoann Anselmetti, Nina Luhmann, Sèverine Bérard, Eric Tannier, and Cedric Chauve. Comparative methods for reconstructing ancient genome organization. In *Comparative Genomics*, pages 343–362. Springer, 2018.

[26] Amandine Perrin, Jean-Stéphane Varré, Samuel Blanquart, and Aïda Ouangraoua. ProCARs: Progressive reconstruction of ancestral gene orders. *BMC genomics*, 16(S5):S6, 2015.

[27] Matthieu Muffato, Alexandra Louis, Nga Thi Thuy Nguyen, and et al. Reconstruction of hundreds of reference ancestral genomes across the eukaryotic kingdom. *Nature Ecology & Evolution*, 7:355–366, 2023.

[28] Yupeng Wang, Haibao Tang, Jeremy D. DeBarry, Jingping Li Xu Tan, Xiyin Wang, Tae Ho Lee, Huizhe Jin, Barry Marler, Hui Guo, Jessica C. Kissinger, and Andrew H. Paterson. MCScanX: a toolkit for detection and evolutionary analysis of gene synteny and collinearity. *Nucleic Acids Research*, 40(7):e49, 2012.