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; qxu062@uottawa.ca
[2] Department of Computer Science, University of Saskatchewan, Saskatoon, SK S7N 5C9, Canada; lingling.jin@cs.usask.ca
[3] Department of Plant Biology, University of Georgia, Athens, GA 30602, USA; jleebensmack@uga.edu
* 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

check for updates

Citation: Xu, Q.; Jin, L.; Leebens-Mack, J.H.; Sankoff, D. Validation of Automated Chromosome Recovery in the Reconstruction of Ancestral Gene Order. *Algorithms* **2021**, *14*, 160. https://doi.org/10.3390/a14060160

Academic Editors: Hélène Touzet and Aïda Ouangraoua

Received: 1 April 2021
Accepted: 18 May 2021
Published: 21 May 2021

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.

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.

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 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,

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

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).

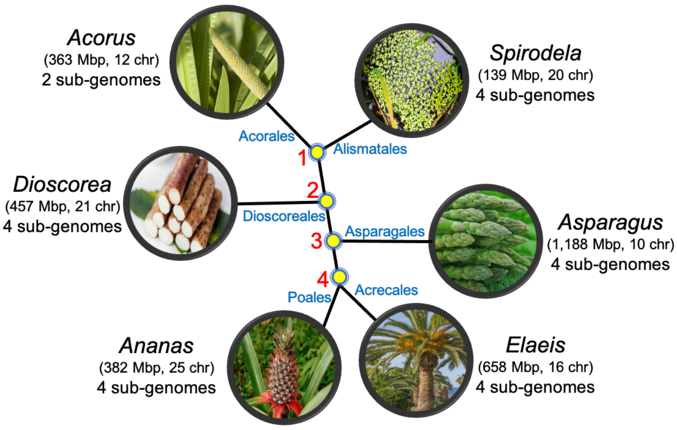
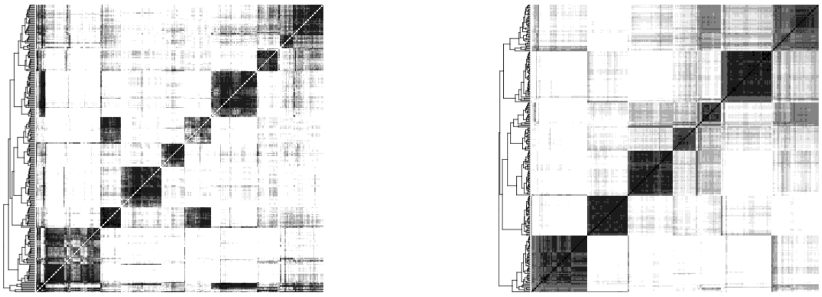


Figure 1. The phylogenetic relationship of the six extant monocot species according to APG IV is summarized in this phylogenetic tree, where ancestors (internal nodes) are labeled with numbers from 1 to 4. The root of this tree is between *Acorus* and Ancestor 1.

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.

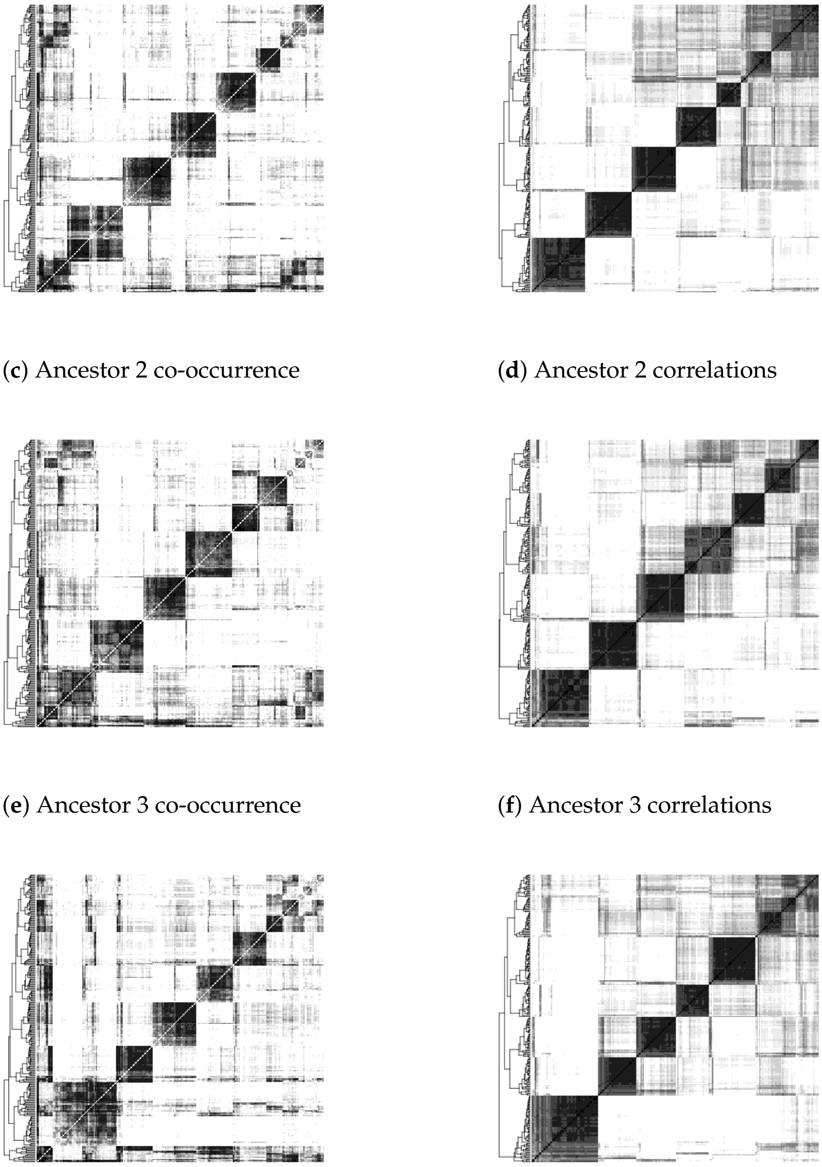
The most telling evidence of the credibility of a reconstruction is the final chromosome structure it produces. The heat map visualization of RACCROCHE applied to the monocot data in Figure 2 shows a completely unambiguous clustering of the contigs into seven chromosomes of comparable size, for all four 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.

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;
```

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, 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 $N(k, j)$.

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

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

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

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
         $G_{1\ldots H}$, 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 : $G'_{1\ldots H}$, simulated gene-order genomes related by $Tr$
1  for ancestor $i \leftarrow 1$ to $I$ do
2      if $i == 1$ then
3          initialize ancestor 1 with genes in $N(1 \times J)$ randomly distributed in 7
             chromosomes;
4      else
5          construct ancestor $i$ by doubling or equal $N((i-1) \times J)$ 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.

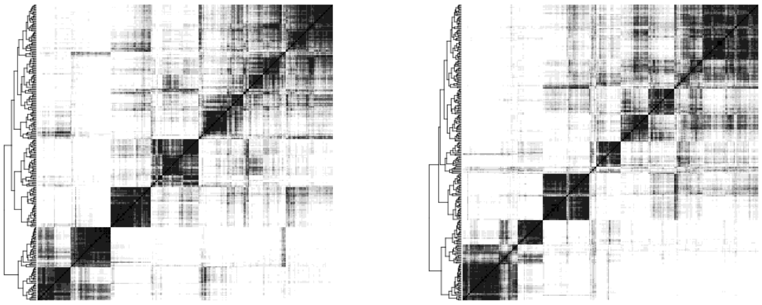
The number of chromosomes in each descendant is then adjusted by fusions of the shortest chromosomes to form a new longer one.

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

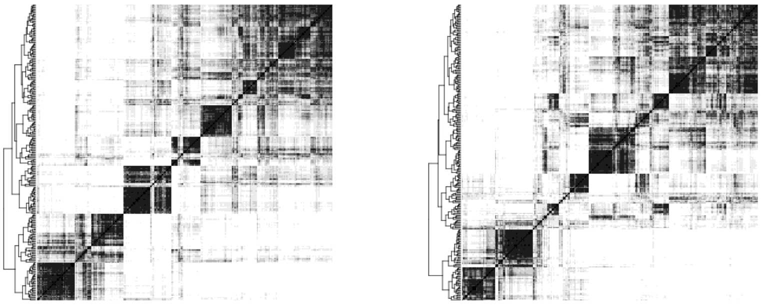
The output of one simulation is summarized by the heat maps in Figure 3. A clear clustering pattern is evident, reminiscent of that obtained for the real data. There is somewhat more noise, suggesting that our simulation is more disruptive to gene order than is natural evolution, either because of the 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.

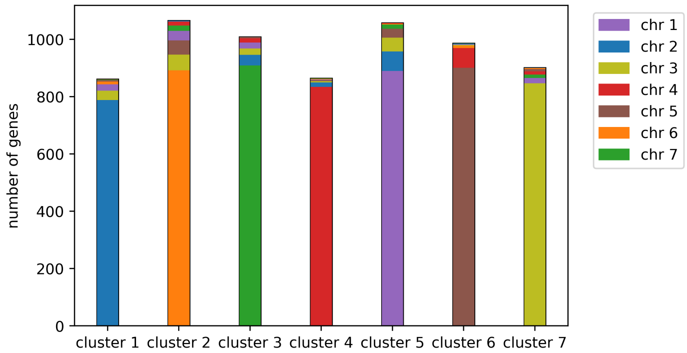
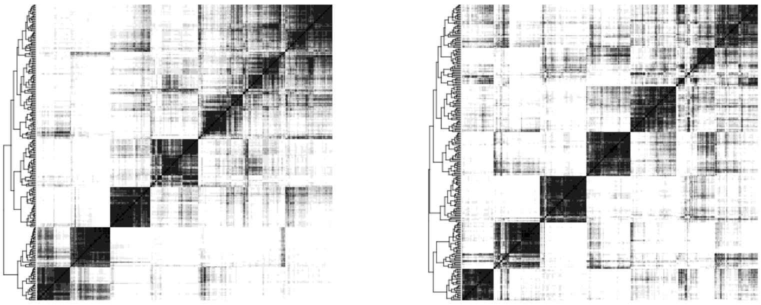


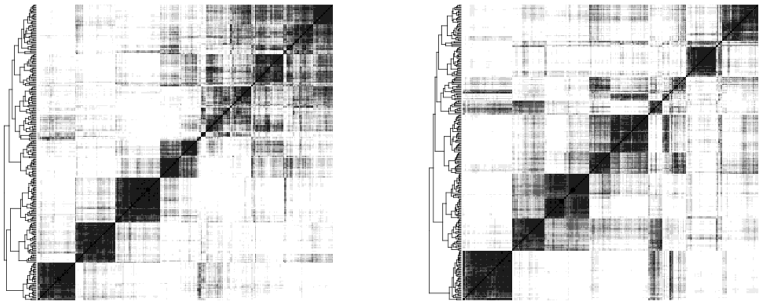
Figure 4. Projection of simulated ancestral chromosomes (colored bars) on version reconstructed by RACCROCHE (outlined bars).

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.H.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/ (accessed on 25 March 2021)) and Phytozome (https://phytozome.jgi.doe.gov/pz/portal.html (accessed on 25 March 2021)). The pipeline is available at https://github.com/jin-repo/RACCROCHE (accessed on 25 March 2021).

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. [CrossRef] [PubMed]
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. [CrossRef] [PubMed]
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. [CrossRef] [PubMed]
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. [CrossRef] [PubMed]
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. In Proceedings of the 10th International Conference on Computational Advances in Bio and Medical Sciences, Virtual, 10 December–12 December, 2020; 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. [CrossRef]
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. [CrossRef]
8. Yang, Z.; Sankoff, D. Natural parameter values for generalized gene adjacency. *J. Comput. Biol.* **2010**, *17*, 1113–1128. [CrossRef] [PubMed]
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. Available online: https://hal.archives-ouvertes.fr/hal-02535466/ (accessed on 25 March 2021).
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. [CrossRef] [PubMed]
12. Shalev-Shwartz, S.; Ben-David, S. *Understanding Machine Learning: From Theory to Algorithms*; Cambridge University Press: Cambridge, UK, 2014.
13. Chanderbali, A. *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. [CrossRef] [PubMed]