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

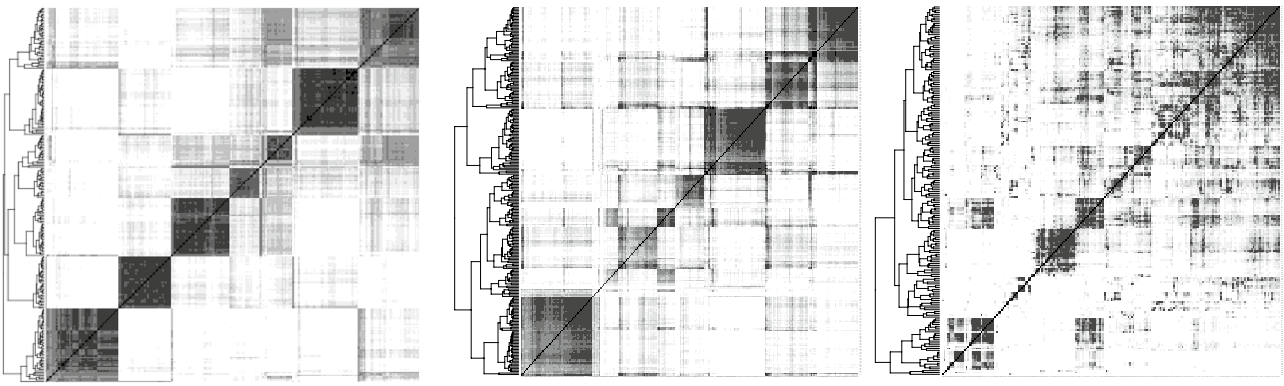

Figure 1. Clear-cut results of RACCROCHE across monocot[8] (left) and eudicot[9] (center) orders, compared to noisy results for intra-ordinal analysis of Fagales on the right (See underlying phylogeny in the "Results" section). Heat maps compare an optimal clustering of ancestral chromosomal fragments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

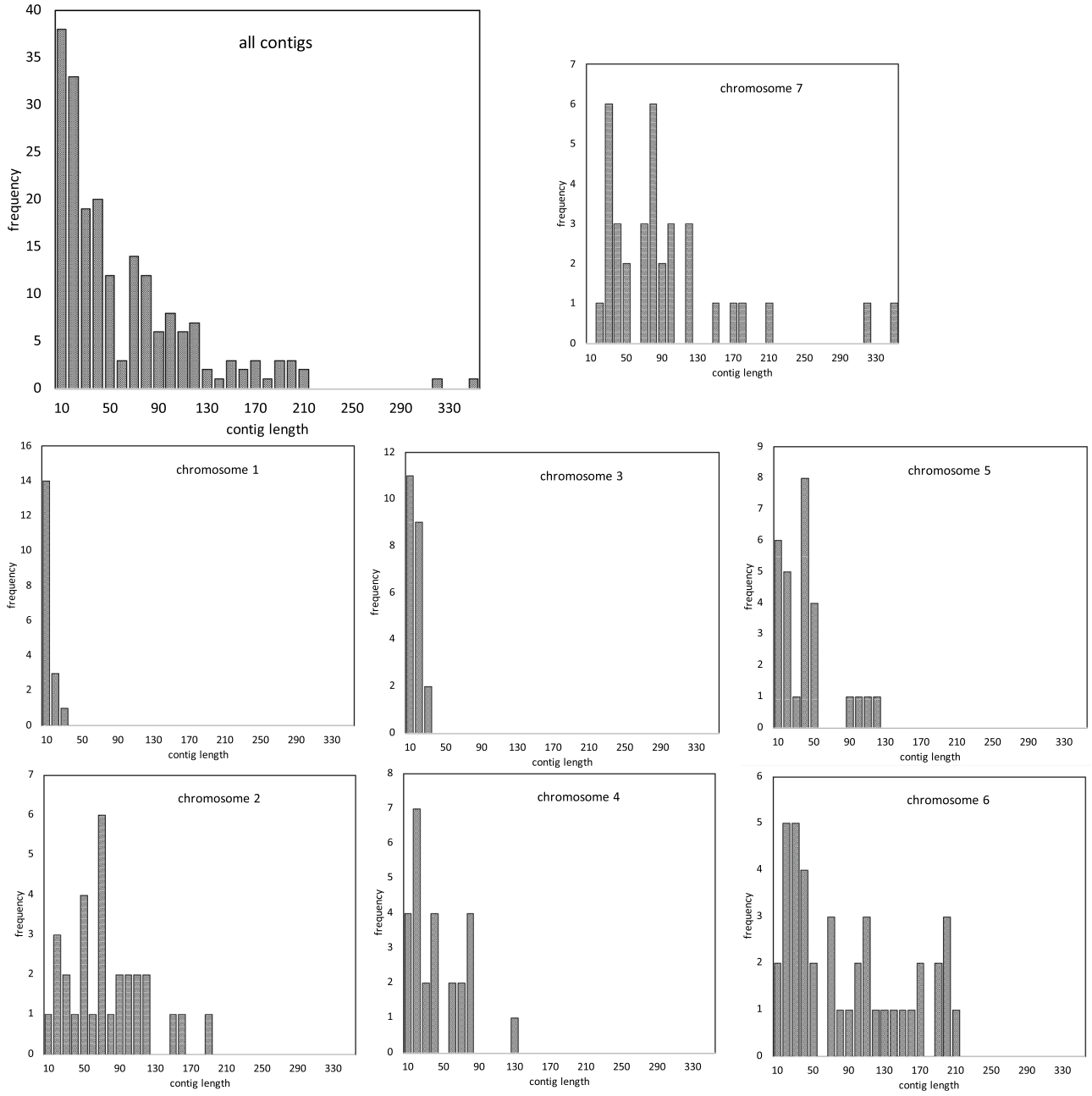

Figure 2. Contigs per chromosome, Fagales Ancestor 2 (which we use as an example throughout; see Fig. 8). N.B. Contig length is in number of genes. Frequency (raw number of occurrences) scale differs among chromosomes.

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

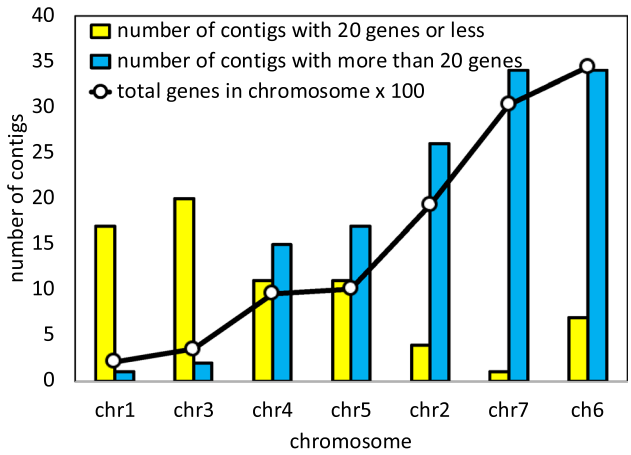

Figure 3. Statistics on the chromosomes in Fig. 2, showing bias in the assignment of long contigs.

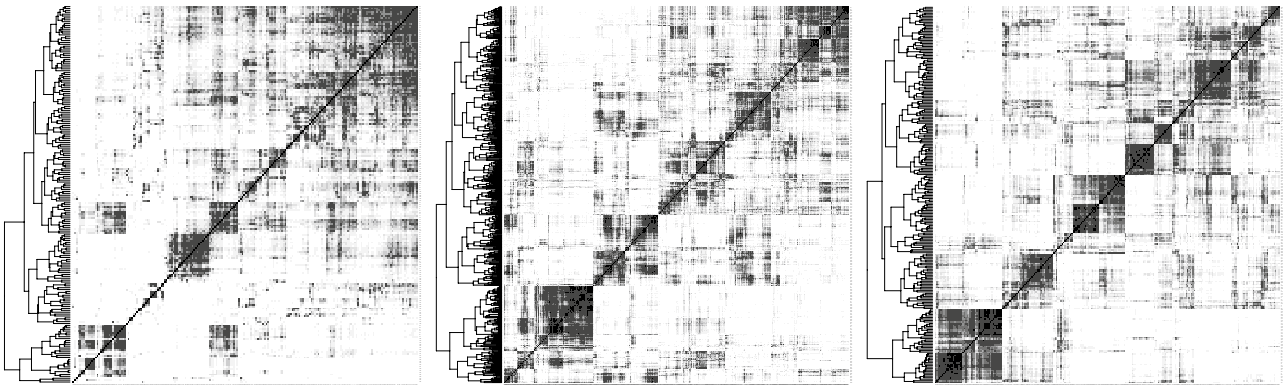

Figure 4. Heat map of Fagales Ancestor 2 based on full contigs (left) and on 20-mers (center) and 40-mers (right). The hierarchical cluster for each heat map depicted at the left-hand side. Shades of grey in cells, representing frequency of chromosomal contig co-occurrence in extant genomes, controlled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

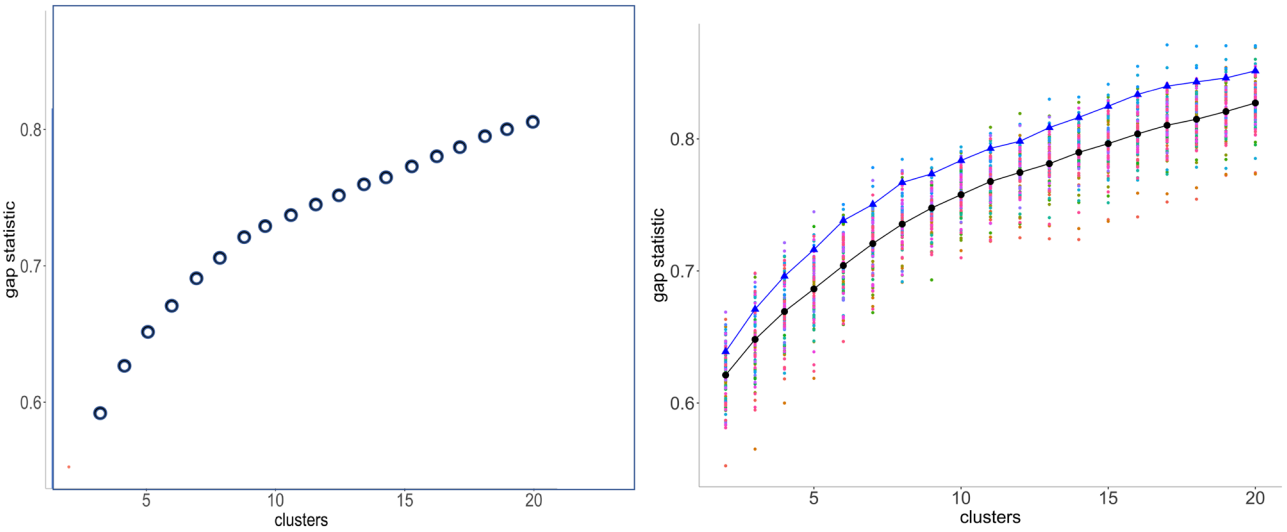

Figure 5. Left: Typical gap statistic for a single MWM sample. Right: Gap statistics for 100 samples with Fagales Ancestor 2. Black: means. Blue: means of 10 "best (as described in text).

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),

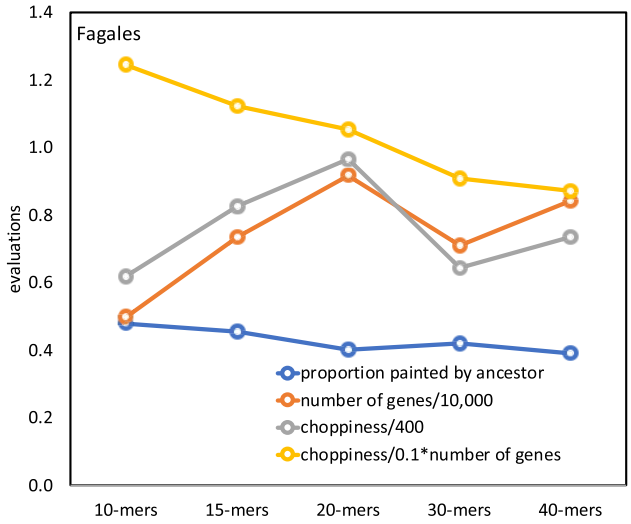

Figure 6. Reconstruction quality as a function of *g*. Averaged over all extant Fagales genomes. Dip at $g = 30$ only reflects a computation cost-imposed switch from 500 to 250 contigs in the clustering.

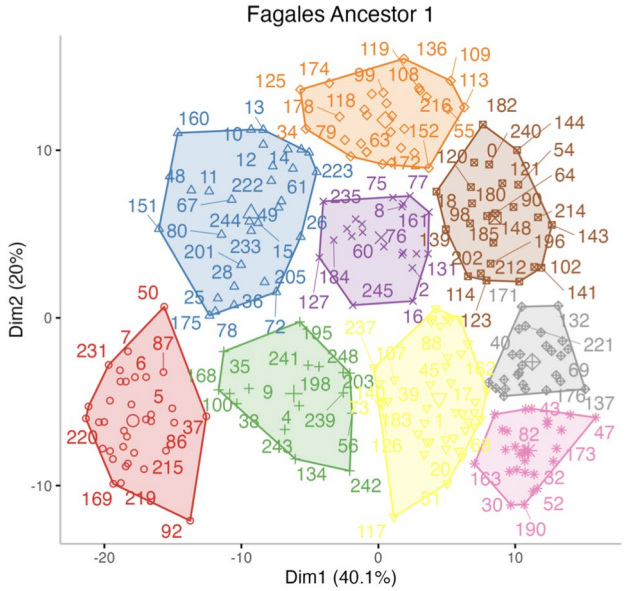

Figure 7. Nine clusters in first two principal components for Fagales Ancestor 1 20-mer data.

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.

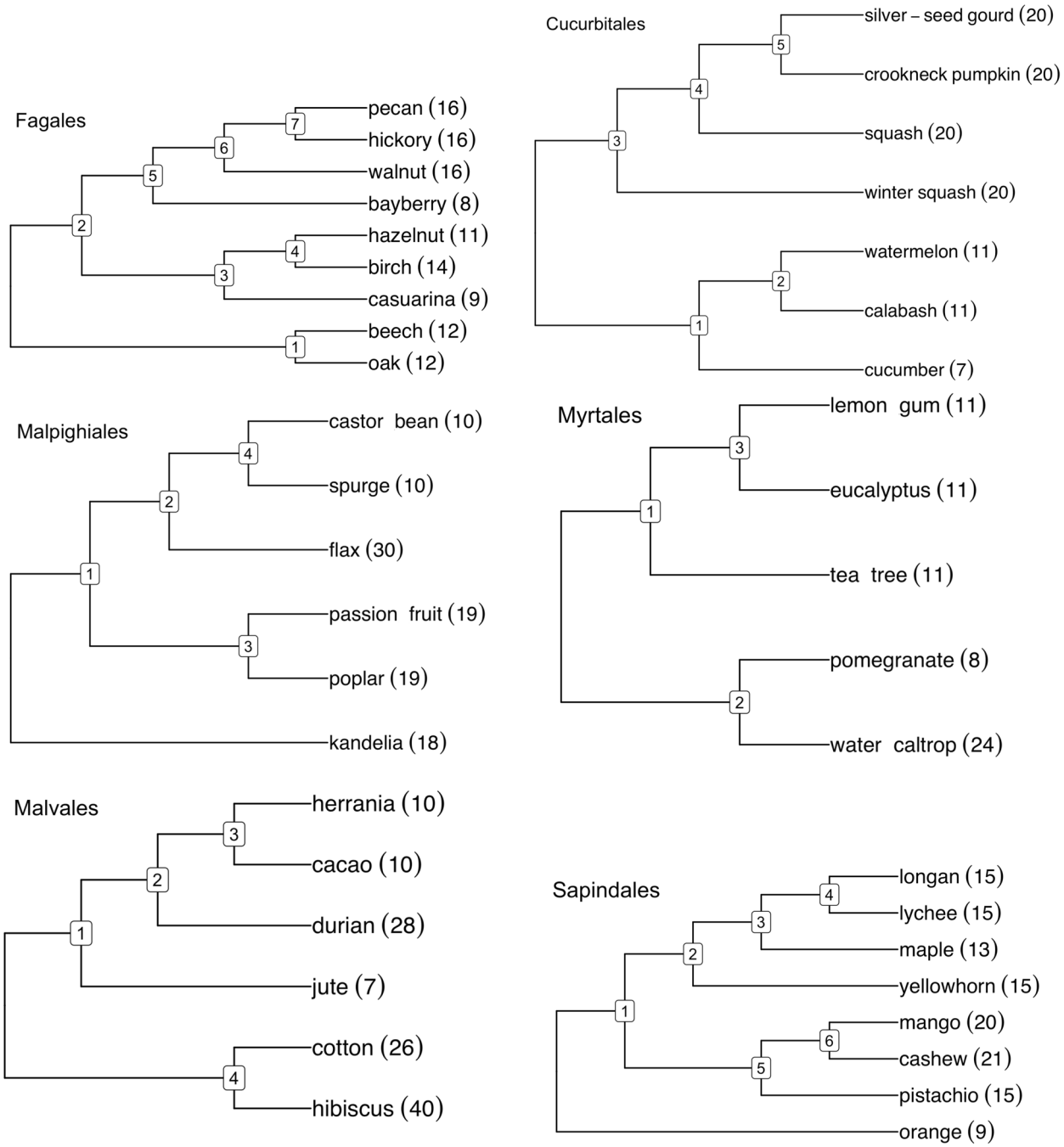

Figure 8. Phylogenies of rosid orders with haploid numbers of chromosomes.

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,

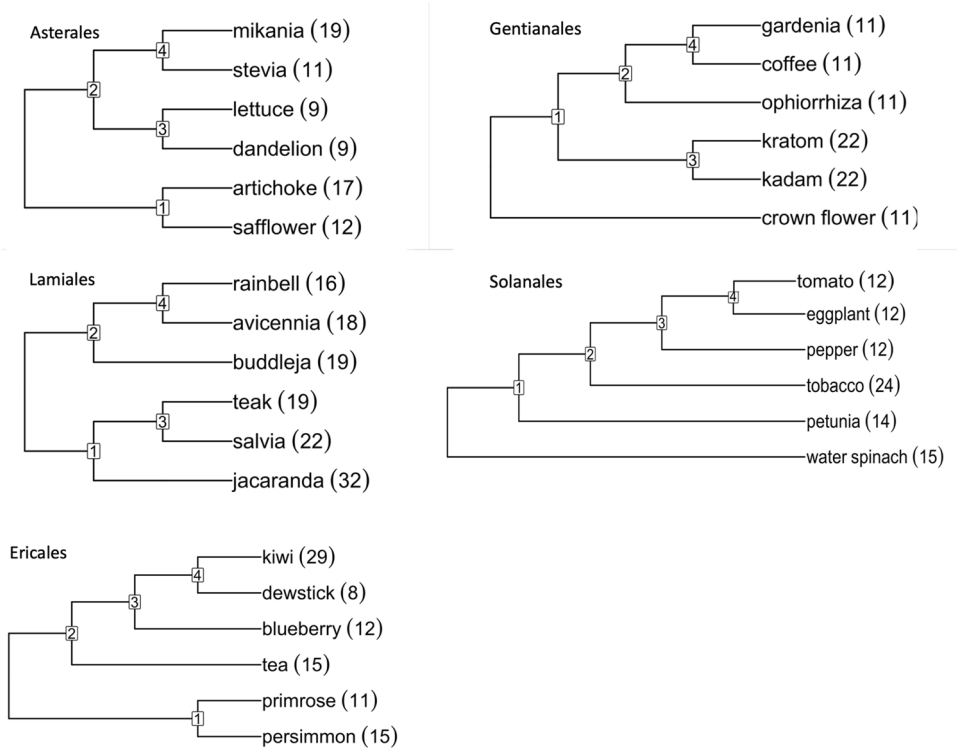

Figure 9. Phylogenies of asterid orders with haploid numbers of chromosomes.

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

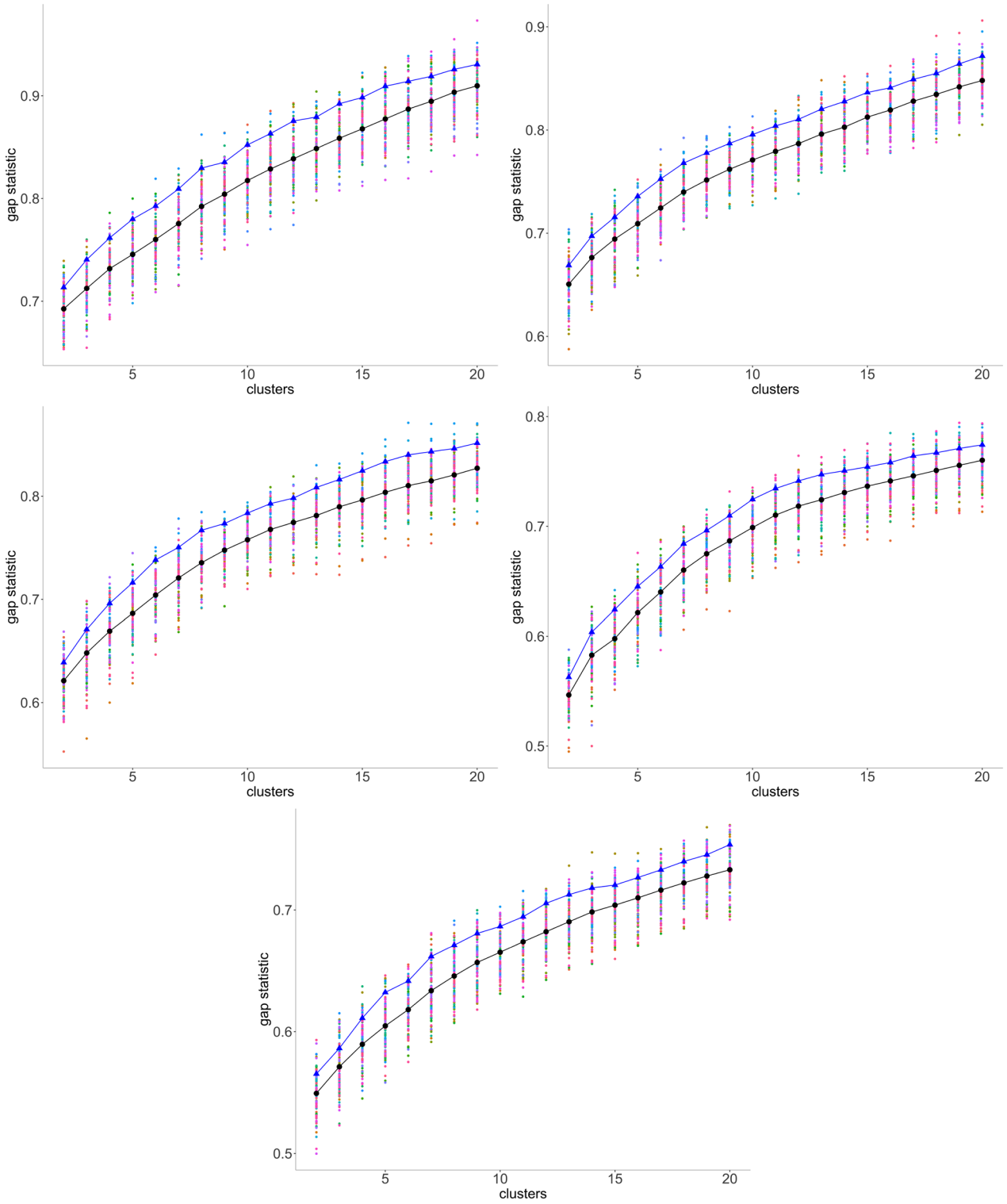

Figure 10. Gap statistics for five Cucurbitales ancestors for 100 samples; means and means of 10 best (in blue). Results for 20-mers. From left to right, and from top to bottom, panels display results for Ancestors 1, 2, 3 , 4 and 5, respectively.

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.

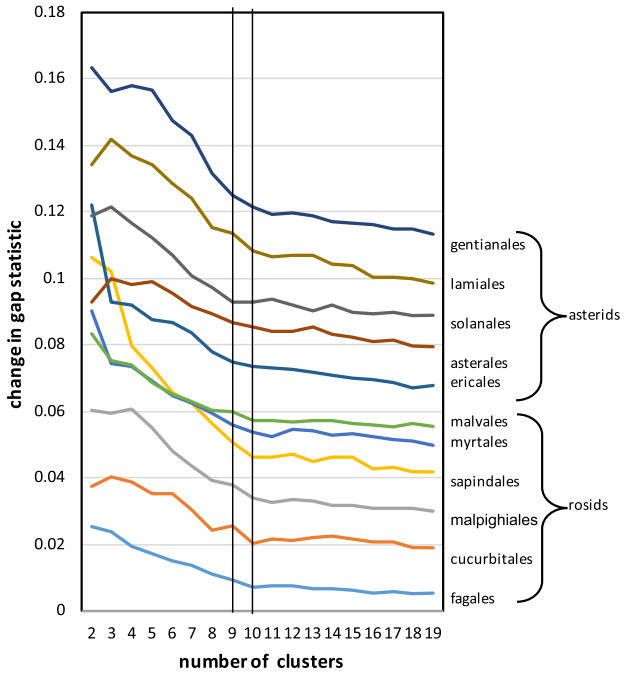

Figure 11. Transition between meaningful and noisy phases of increase in gap statistic for 11 core eudicot orders. The y-axis values are displaced + 0.01 for each order in the list above the previous item. Vertical lines at $k = 9$ and $k = 10$ highlight that for most orders, the increment at $k = 10$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

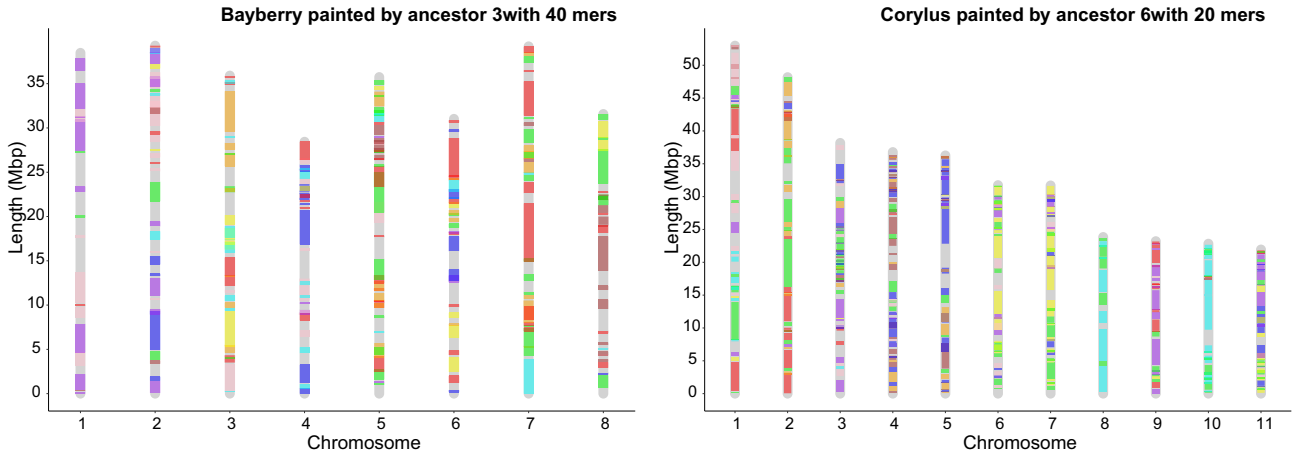

Figure 12. Painting the chromosomes of the extant genomes using colors corresponding to the ancestral chromosomes. Example of ancestral Fagales colors on the oak and the bayberry genomes.

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

Malpighiales chromosome \ Fagales 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.

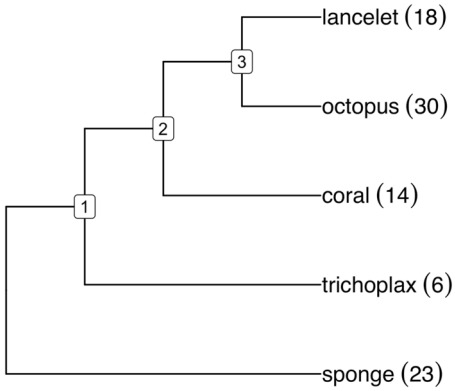

Figure 15. Metazoan phylogeny with haploid numbers of chromosomes. We can consider ancestral genome 1 to represent either the eumatazoan ancestor, a sister group to the sponges (porifera), or the more recent parahoxoan ancestor, giving rise to the placazoans like trichoplax, the cnidarians like coral 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.

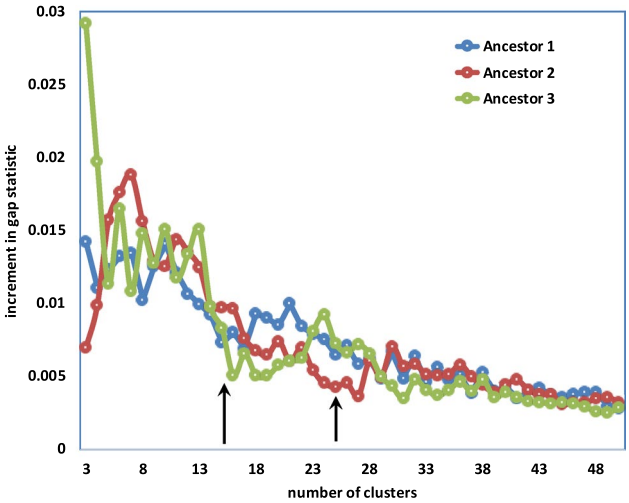

Figure 16. Significance increment graph for cluster-based metazoan karyotypes.

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.