frontiers in
PLANT SCIENCE

ORIGINAL RESEARCH ARTICLE
published: 29 August 2012
doi: 10.3389/fpls.2012.00192

A phylogenetic analysis of the Brassicales clade based on an alignment-free sequence comparison method

Klas Hatje and Martin Kollmar*

Abteilung NMR-Basierte Strukturbiologie, Max-Planck-Institut für Biophysikalische Chemie, Göttingen, Germany

Edited by:
Xiaowu Wang, Chinese Academy of Agricultural Sciences, China

Reviewed by:
Ruiqiang Li, Peking University, China
Bernhard Haubold, Max-Planck-Society, Germany

****Correspondence:***
Martin Kollmar, Abteilung NMR-Basierte Strukturbiologie, Max-Planck-Institut für Biophysikalische Chemie, Am Fassberg 11, D-37077 Göttingen, Germany.
e-mail: mako@nmr.mpibpc.mpg.de

Phylogenetic analyses reveal the evolutionary derivation of species. A phylogenetic tree can be inferred from multiple sequence alignments of proteins or genes. The alignment of whole genome sequences of higher eukaryotes is a computational intensive and ambitious task as is the computation of phylogenetic trees based on these alignments. To overcome these limitations, we here used an alignment-free method to compare genomes of the Brassicales clade. For each nucleotide sequence a Chaos Game Representation (CGR) can be computed, which represents each nucleotide of the sequence as a point in a square defined by the four nucleotides as vertices. Each CGR is therefore a unique fingerprint of the underlying sequence. If the CGRs are divided by grid lines each grid square denotes the occurrence of oligonucleotides of a specific length in the sequence (Frequency Chaos Game Representation, FCGR). Here, we used distance measures between FCGRs to infer phylogenetic trees of Brassicales species. Three types of data were analyzed because of their different characteristics: (A) Whole genome assemblies as far as available for species belonging to the Malvidae taxon. (B) EST data of species of the Brassicales clade. (C) Mitochondrial genomes of the Rosids branch, a supergroup of the Malvidae. The trees reconstructed based on the Euclidean distance method are in general agreement with single gene trees. The Fitch–Margoliash and Neighbor joining algorithms resulted in similar to identical trees. Here, for the first time we have applied the bootstrap re-sampling concept to trees based on FCGRs to determine the support of the branchings. FCGRs have the advantage that they are fast to calculate, and can be used as additional information to alignment based data and morphological characteristics to improve the phylogenetic classification of species in ambiguous cases.

Keywords: Chaos game representation, Brassicales, *Brassica rapa*, phylogenetic tree, bootstrap re-sampling, frequency Chaos game representation

INTRODUCTION

Phylogenetic analyses reveal the evolutionary derivation of species. A phylogenetic tree can be inferred from multiple sequence alignments of proteins or genes, which assume the conservation and contiguity over the total sample length between homologous sequences (Blair and Murphy, 2011). The alignment of whole genome sequences of eukaryotes is a computational intensive and ambitious task as is the computation of phylogenetic trees based on these alignments (Dewey, 2012). In particular, genetic recombination and shuffling during species evolution complicate whole genome alignments limiting species genome versus single gene, multiple gene, or transcriptome comparisons. However, it would be beneficial for the significance of the species trees, if also whole genome assembly data were taken into account. In the past two decades several methods have been suggested for alignment-free sequence analyses that mainly group into word (oligomer) frequency methods and methods that do not resolve the fixed word-length distance measures and are thus absolutely independent from the assumption of conservation and contiguity (reviewed in Vinga and Almeida, 2003). The latter category includes the Chaos Theory (Jeffrey, 1990) and the theoretical concept of Kolmogorov complexity (Li et al., 2001). More recent methods include the alignment-free estimation of the number of substitutions per site (Domazet-Loso and Haubold, 2009) and feature frequency profiles (Sims et al., 2009).

The Chaos Game Representation (CGR) denotes an algorithm, which produces fractal pictures and can be adapted to reveal patterns in DNA (Li et al., 2001) and even protein sequences (Basu et al., 1997; Pleissner et al., 1997). These CGR pictures exhibit the fractal property that the overall pattern of the CGR picture is repeated in smaller parts of the picture. It has been shown that this self-similarity even holds for whole genome sequences and its sub-sequences, like single chromosomes, contigs, or genes (Deschavanne et al., 1999; Almeida et al., 2001; Joseph and Sasikumar, 2006). Commonly, the pictures of DNA sequences are generated as squares such that the lower $(A + T)$ and the upper $(C + G)$ halves indicate the base composition and the diagonals the purine/pyrimidine composition. CGRs are unique descriptions of each DNA sequence and, in the case of whole genome sequences, can therefore be regarded as genomic fingerprints. However, the CGRs are not directly comparable. If the CGR pictures are divided into smaller squares by grid lines, each grid square represents the frequencies of the respective oligonucleotides as found in the whole sequence (Deschavanne et al., 1999; Almeida et al., 2001). These frequencies can be represented in Frequency Chaos Game Representation (FCGR) pictures with a gray scale to express the number of points within each grid square and with pictures for each length k oligonucleotide (with $k = 1, 2, 3\ldots$). FCGRs are numerical matrices and can be used to infer phylogenetic trees based on distance methods (Wang et al., 2005). So far this approach has only been applied to reconstruct the phylogeny of 20 birds using nuclear genome data (Edwards et al., 2002), to analyze the mitochondrial genomes of 26 sample eukaryotes (Wang et al., 2005), and to sub-typing of HIV-I (Pandit and Sinha, 2010). One of the advantages of using FCGRs for phylogenetic reconstructions is that sequence, which cannot be aligned, can be used.

Here, we performed phylogenetic analyses based on three different types of data. Firstly we used the whole genomic sequence assemblies of all so far sequenced species in the taxon Malvidae, including that of *Brassica rapa.* Because a reference tree including all these species was not available we assembled and annotated all actin capping (CP) protein sequences (Cooper and Sept, 2008) and the sequences of the actin-related proteins Arp2 and Arp3 (Goley and Welch, 2006). These proteins are present in all eukaryotes and as single copies in the *Arabidopsis thaliana* genome. Thus they were not expected to exist in duplicates in the other analyzed species avoiding the ortholog-paralog problem. To infer the phylogeny of the different *Brassica* species, for which whole genome assemblies have not yet been produced, we used EST and mitochondrial genome DNA. The quality of the phylogenetic analyses depends on the resolution of the FCGRs (length of k) and thus on the length of the nucleotide sequences. Thus we only included those species for which a considerable number of EST clones were available. To estimate the support for the branchings, here, we apply the concept of bootstrap re-sampling to the comparison of FCGRs for the first time.

MATERIALS AND METHODS

DATA ACQUISITION

The genome files were retrieved from diArk[1] (Hammesfahr et al., 2011), and the mitochondrial genomes and EST reads from the NCBI database, each in FASTA format (**Table 1**). For the generation of the CGRs the contigs and reads of each dataset were concatenated. The whole genome assemblies as available from the sequencing centers contain both the nuclear and mitochondrial genomes, and potentially still some contaminations from other species' DNA. However, given the sizes of the whole genome datasets, the contributions of the mitochondrial genomes and contaminating DNA to the FCGRs are negligible. The FCGRs of the whole genome data can thus be regarded as identical to the FCGRs of the nuclear genomes.

IMPLEMENTATION OF THE ALGORITHM

The algorithm to calculate CGRs and FCGRs was implemented in C/C++. CGR positions were generated as lists in plain text and plotted for graphical presentations in the Scalable Vector Graphics (SVG) format[2]. Based on the CGR position values, FCGRs were calculated for each k in 1, …, 8. Distance calculations were implemented in Ruby[3].

GENERATING CHAOS GAME REPRESENTATIONS

Chaos game representations of the nucleotide sequences were generated by the following algorithm. A 1 × 1 square is drawn and each vertex labeled by a nucleotide. In agreement with other analyses we placed C in the upper left, G in the upper right, A in the lower left, and T in the lower right vertex. The starting point is defined as the geometric center of the square at position (0.5, 0.5). The respective nucleotide sequences are then plotted sequentially. For the first nucleotide a point is plotted on half the distance between the starting point (0.5, 0.5) and the vertex corresponding to this nucleotide. Subsequently for each following nucleotide a point is placed as mid-point between the previously plotted point and the vertex corresponding to the nucleotide (**Figure 1A**).

The algorithm can be expressed by the following equations:

$$\mathrm{CGR}_0 = (0.5, 0.5) \tag{1}$$

$$\mathrm{CGR}_i = \begin{cases} \mathrm{CGR}_{i-1} + 0.5 \cdot (\mathrm{CGR}_{i-1} + (0.0, 0.0)) & \text{if } \mathrm{seq}_i = \text{`C'} \\ \mathrm{CGR}_{i-1} + 0.5 \cdot (\mathrm{CGR}_{i-1} + (1.0, 0.0)) & \text{if } \mathrm{seq}_i = \text{`G'} \\ \mathrm{CGR}_{i-1} + 0.5 \cdot (\mathrm{CGR}_{i-1} + (0.0, 1.0)) & \text{if } \mathrm{seq}_i = \text{`A'} \\ \mathrm{CGR}_{i-1} + 0.5 \cdot (\mathrm{CGR}_{i-1} + (1.0, 1.0)) & \text{if } \mathrm{seq}_i = \text{`T'} \end{cases} \tag{2}$$

The resulting plot is unique for each sequence. The overall pattern of points is repeated in each sub-square of the plot (**Figure 1B**). In addition, each plot based on a sub-sequence of the whole sequence has a similar appearance. Thus similar sequences result in similar CGR plots. **Figure 1B** shows the CGR of the first 1,000,000 nt of the *B. rapa* genome sequence.

The calculation of the frequencies of points within each sub-square results in an FCGR. Thus each FCGR represents the occurrence of oligonucleotides in the whole sequence. For dinucleotides ($k = 2$) the binary square is divided into a 4 × 4 grid, for trinucleotides ($k = 3$) into an 8 × 8 grid, and in general into a $2^k \times 2^k$ grid. **Figure 1C** shows an FCGR ($k = 3$) of the whole *B. rapa* genome sequence.

If the nucleotide sequences differ in length, the resulting FCGRs will also differ in there overall frequencies. To overcome this sequence length bias each FCGR was standardized (Wang et al., 2005). If the FCGR is represented as for example a $2^k \times 2^k$ matrix, the matrix $A = (a)_{2^k \times 2^k}$ is transformed to a standardized FCGR as follows:

$$\bar{A} = \frac{4^k}{\sum_{i=1}^{k} \sum_{j=1}^{k} a_{i,j}} A \tag{3}$$

The nucleotide sequences of each data file (whole genome, EST, or mitochondrial genome data) were concatenated and the reverse

[1] http://www.diark.org

[2] http://www.w3.org/Graphics/SVG

[3] http://ruby-lang.org

Table 1 | List of the species used in the analysis.

Species	Whole genome			EST		Mitochondrial genome		
	Contigs	Nucleotides	Accession numbers	Reads	Nucleotides	Contigs	Nucleotides	Accession numbers
Arabidopsis lyrata	695	206667935	GL348713–GL349407					
Arabidopsis thaliana	5	119145879	NC_003070–NC_003071, NC_003074–NC_003076	1529700	400512451	1	366924	NC_001284
Brassica rapa	51658	273071614	AENI01000001–AENI01051658	213605	122970377	1	219747	NC_016125
Capsella rubella	853	134834574						
Carica papaya	3207	331271729	DS981520–DS984726	77393	54789864			
Citrus clementina	1128	295550349						
Citrus sinensis	12574	319231331						
Eucalyptus camaldulensis	274001	654922307	DF097775–DF126446					
Eucalyptus grandis	4952	691297852						
Eutrema halophilum	639	243117811		38022	20080214			
Eutrema parvulum	7	114396853	CM001187–CM001193					
Gossypium raimondii	1448	763818933						
Theobroma cacao	1782	351351221	FR720657–FR725448					
Vitis vinifera	33	486265422	FN597015–FN597047	446643	284204927	1	773279	NC_012119
Brassica napus				643437	381399492	1	221853	NC_008285
Brassica oleracea				179150	125257248	1	360271	NC_016118
Limnanthes alba				15331	8582959			
Raphanus raphanistrum				164119	104536170			
Raphanus sativus				150680	97973638			
Tropaeolum majus				10507	6436290			
Brassica carinata						1	232241	NC_016120
Brassica juncea						1	219766	NC_016123
Lotus japonicus						1	380861	NC_016743
Millettia pinnata						1	425718	NC_016742
Ricinus communis						1	502773	NC_015141

The number of contigs/reads and the number of nucleotides for the whole genome, EST, and mitochondrial genome data files are given. In addition, for whole genome and mitochondrial genome data the NCBI accession numbers are given if available.

complement of the concatenated sequence was appended. Characters other than "C," "G," "A," or "T" were ignored. Some example FCGRs generated with $k = 8$ are shown in **Figures 1D–L**. Already by visual inspection it is obvious, that whole genome, EST, and mitochondrial genome FCGRs have distinct patterns (**Figures 1D–F**), while the FCGRs generated from the same data type of closely related species are very similar (**Figures 1G–L**). EST data disproportionately contain poly-A sequences, resulting in unusually high frequency values in the FCGRs. These subsequently dominate the distance matrix calculation for higher order FCGRs ($k > 5$) and misdirect the calculation of the phylogenetic trees (data not shown). Therefore, in the case of EST data, the two entries in each FCGR that contain poly-A and poly-T stretches were set to zero.

DISTANCES

In order to reveal the phylogenetic relation between the analyzed species we calculated pair-wise distances between the FCGRs. In general all distances that are applicable to matrices could be used. The following distances have already been described for comparing FCGRs: The Hamming distance (Campbell et al., 1999; Wang et al., 2005), the Euclidean distance (Edwards et al., 2002; Vinga and Almeida, 2003; Wang et al., 2005; Pandit and Sinha, 2010), the Image distance defined in Wang et al. (2005), and the Pearson distance (Almeida et al., 2001; Vinga and Almeida, 2003; Wang et al., 2005). Here, we chose the Pearson distance as a statistical distance and the Euclidean distance as a geometrical distance, which performed best in a comparison of difference distance methods (Wang et al., 2005). The Euclidean distance between two points in two-dimensional space is defined as the length of the line segment between these two points and can be calculated using the Pythagorean equation. This concept can be adapted to calculate the distance between two FCGRs. The Euclidean distance between two standardized FCGRs $A = (a)_{2^k \times 2^k}$ and $B = (b)_{2^k \times 2^k}$ is defined as follows:

$$d_{\text{Euclidean}}\left(\bar{A}, \bar{B}\right) = \frac{\sqrt{2^k}}{4^k}\sqrt{\sum_{i=1}^{2^k}\sum_{j=1}^{2^k}\left(a_{i,j} - b_{i,j}\right)^2} \tag{4}$$

The Pearson distance is based on a weighted Pearson correlation coefficient (Almeida et al., 2001; Wang et al., 2005). To calculate the Pearson distance, the FCGRs are represented as lists of the

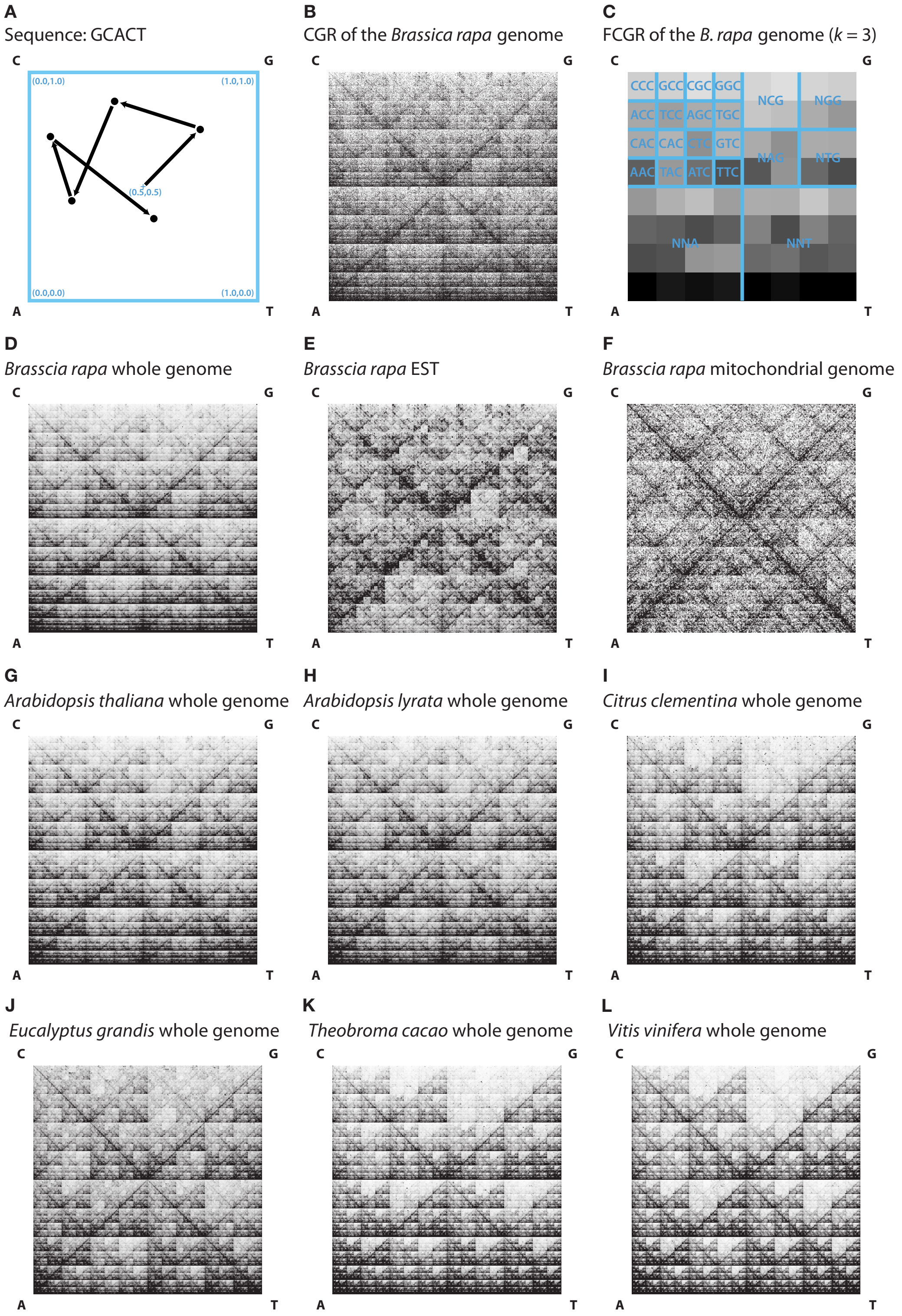
A
Sequence: GCACT
C
G
(0.0,1.0)
(1.0,1.0)
(0.5,0.5)
(0.0,0.0)
(1.0,0.0)
A
T
B
CGR of the Brassica rapa genome
C
G
A
T
C
FCGR of the B. rapa genome (k = 3)
C
G
CCC GCC CGC GGC
ACC TCC AGC TGC
CAC CAC CTC GTC
AAC TAC ATC TTC
NCG
NGG
NAG
NTG
NNA
NNT
A
T
D
Brasscia rapa whole genome
E
Brasscia rapa EST
F
Brasscia rapa mitochondrial genome
G
Arabidopsis thaliana whole genome
H
Arabidopsis lyrata whole genome
I
Citrus clementina whole genome
J
Eucalyptus grandis whole genome
K
Theobroma cacao whole genome
L
Vitis vinifera whole genome
C
G
A
T

FIGURE 1 | (A) A Chaos game representation (CGR) image is generated by drawing a unit square and, starting at the center (0.5, 0.5), plotting for each nucleotide of the sequence a point on half the distance to the corresponding vertex. In this example the CGR for the sequence "GCACT" was drawn. **(B)** The image shows the CGR of the first 1,000,000 nt of the *Brassica rapa* genome. **(C)** The figure shows an FCGR ($k = 3$) of the whole *Brassica rapa* genome illustrating the frequencies of points in the CGR in an 8 × 8 grid. The squares of the grid represent the occurrence of specific trinucleotides, which are labeled in the figure. In **(D–L)** the FCGRs ($k = 8$) of the whole genome **(D)**, EST **(E)** and mitochondrial genome sequences **(F)** of *Brassica rapa* and the FCGRs ($k = 8$) of the whole genome sequences of some representatives of the different clades **(G–L)** are shown for visual comparison.

frequencies with $n = 4^k$ values. The Pearson distance between the non-standardized FCGRs $A = (x_1, \ldots, x_n)$ and $B = (y_1, \ldots, y_n)$ is defined as follows:

$$nw = \sum_{i=1}^{n} x_i \cdot y_i$$

$$\bar{x}w = \frac{\sum_{i=1}^{n} x_i^2 \cdot y_i}{nw}, \quad \bar{y}w = \frac{\sum_{i=1}^{n} y_i^2 \cdot x_i}{nw},$$

$$sx = \frac{\sum_{i=1}^{n} (x_i - \bar{x}w)^2 \cdot x_i \cdot y_i}{nw}, \quad sy = \frac{\sum_{i=1}^{n} \left(y_i - \bar{y}w\right)^2 \cdot x_i \cdot y_i}{nw}$$

$$d_{\text{Pearson}} = 1 - \frac{\sum_{i=1}^{n} \frac{x_i - \bar{x}w}{\sqrt{sx}} \cdot \frac{y_i - \bar{y}w}{\sqrt{sy}} \cdot x_i \cdot y_i}{nw} \qquad (5)$$

GENERATING PHYLOGENETIC TREES

To generate the phylogenetic trees, pair-wise distance matrices were calculated for each k in 1, . . ., 0.8 with the Euclidean distance method as defined in Eq. 4 and the Pearson distance as defined in Eq. 5. The distance matrices were subjected to the Neighbor joining (NJ) and Fitch–Margoliash algorithms as implemented in the Phylip package[4]. Statistical support for branchings was obtained by applying the bootstrap re-sampling method. For each FCGR, 500 datasets were generated by random sampling with replacement. Based on these re-sampled FCGRs 500 phylogenetic trees were reconstructed for each k in 1, . . ., 0.8. The trees of each dataset were summarized to consensus trees using the *consense* program of the Phylip package. The topologies of the consensus trees were fixed and the branch lengths calculated with the Fitch–Margoliash algorithm. In the case of the NJ trees, a bootstrapped tree was chosen that had the same topology as the consensus tree and the bootstrap values were plotted onto this tree. The bootstrap values represent the percentage each interior branch has the same partition as the consensus tree.

GENERATION OF THE REFERENCE TREE FOR THE WHOLE GENOME ANALYSIS

For the reference tree of those species for which whole genome assemblies are available we identified, assembled, and annotated the sequences of the heterodimeric actin capping protein (CAP), α- and β-CAP, and the sequences of the actin-related proteins Arp2 and Arp3. The *B. rapa* and *Gossypium raimondii* genomes contain duplicates of these genes due to species-specific duplications. Therefore, only one of the duplicates had been used for the phylogenetic tree reconstructions. The CAP and Arp sequences were aligned, concatenated, and phylogenetic trees reconstructed using the NJ and the Maximum likelihood (ML) method. The NJ tree was unrooted and generated using ClustalW (Chenna et al., 2003) with standard settings and the Bootstrap (1,000 replicates) method. The ML tree was calculated using the JTT (Jones et al., 1992) substitution model as suggested by ProtTest (Darriba et al., 2011) with estimated proportion of invariable sites and bootstrapping (1,000 replicates) using RAxML (Stamatakis et al., 2008).

RESULTS

Phylogenetic trees based on whole genome, mitochondrial genome, and EST data were generated using the Euclidean or Pearson distance methods in combination with the NJ or the Fitch–Margoliash tree reconstruction algorithms. In order to reveal the influence of the lengths of the oligonucleotides we report trees of FCGRs generated with $k = 3$ (trinucleotides, 64 data points) and $k = 8$ (octanucleotides, 65,536 data points).

INFLUENCE OF SEQUENCE LENGTHS ON THE PHYLOGENETIC TREES

First we tested whether different sequence lengths have an influence on the results (**Figure 2**). For the whole genome assemblies and the EST datasets, sub-sections of the sequences were generated with lengths of 10^6, 10^7, and 10^8 nt. For that purpose the contigs or EST entries of each organism were shuffled, concatenated, and subsequently the sub-sequences generated by cutting the sequences at the respective positions. In the case of the whole genome data (**Figure 2A**), the FCGRs of the whole genome assemblies and the sub-sequences of each organism grouped together forming clusters. The only exceptions were the shortest 10^6-nt sequences of *Citrus sinensis*, *Citrus clementina*, *Arabidopsis lyrata*, and *A. thaliana*, which group to different species. The FCGRs of the EST data group together for each species independently of the lengths of the sequences (**Figure 2B**). For the mitochondrial genomes datasets with shorter sequences of 10^4 and 10^5 nt were generated. Here the FCGRs of the 10^4 nt sequences do not cluster together with those of the longer sequences of the corresponding species. The FCGRs of the mitochondrial sequences have been calculated based on hexanucleotides ($k = 64{,}096$ data points). Here, $k = 6$ was chosen, because in the case of higher k values ($k = 7$ or $k = 8$), the sequence length of the shortest sequences (10^4 nt) would be less than the number of data points in the FCGRs. In the shortest sequences (10^4 nt) many of the hexanucleotides are not covered at all resulting in many zero values for frequency positions, which lead to the unusual grouping of these FCGRs.

WHOLE GENOME ANALYSIS

In order to analyze the phylogenetic grouping of *B. rapa* in a whole genome context we searched for closely related plant species, for which whole genome assemblies are available. According to diArk (Hammesfahr et al., 2011), that comprises the most reliable and complete compilation of eukaryotic genome projects for which genome assemblies are available, the genomes of 13 different species (excluding different *A. thaliana* strains) of the taxon Malvidae have been sequenced and assembled: *A. lyrata* (Hu et al., 2011), *A. thaliana* (thale cress; Arabidopsis Genome Initiative, 2000), *B. rapa* subsp. *pekinensis* (Chinese cabbage; Wang et al., 2011), *Capsella rubella*, *Carica papaya* (Ming et al., 2008), *C. clementina*, *C. sinensis* (sweet orange), *Eucalyptus camaldulensis* (Murray red gum), *Eucalyptus grandis* (Flooded gum), *Eutrema halophilum* (salt cress), *Eutrema parvulum* (Dassanayake et al., 2011), *G. raimondii*, and *Theobroma cacao* (cacao plant; Argout et al., 2011). In addition, the genome of *Vitis vinifera* (grape vine; Jaillon et al., 2007; Velasco et al., 2007) was chosen as outgroup

[4] http://evolution.genetics.washington.edu/phylip.html

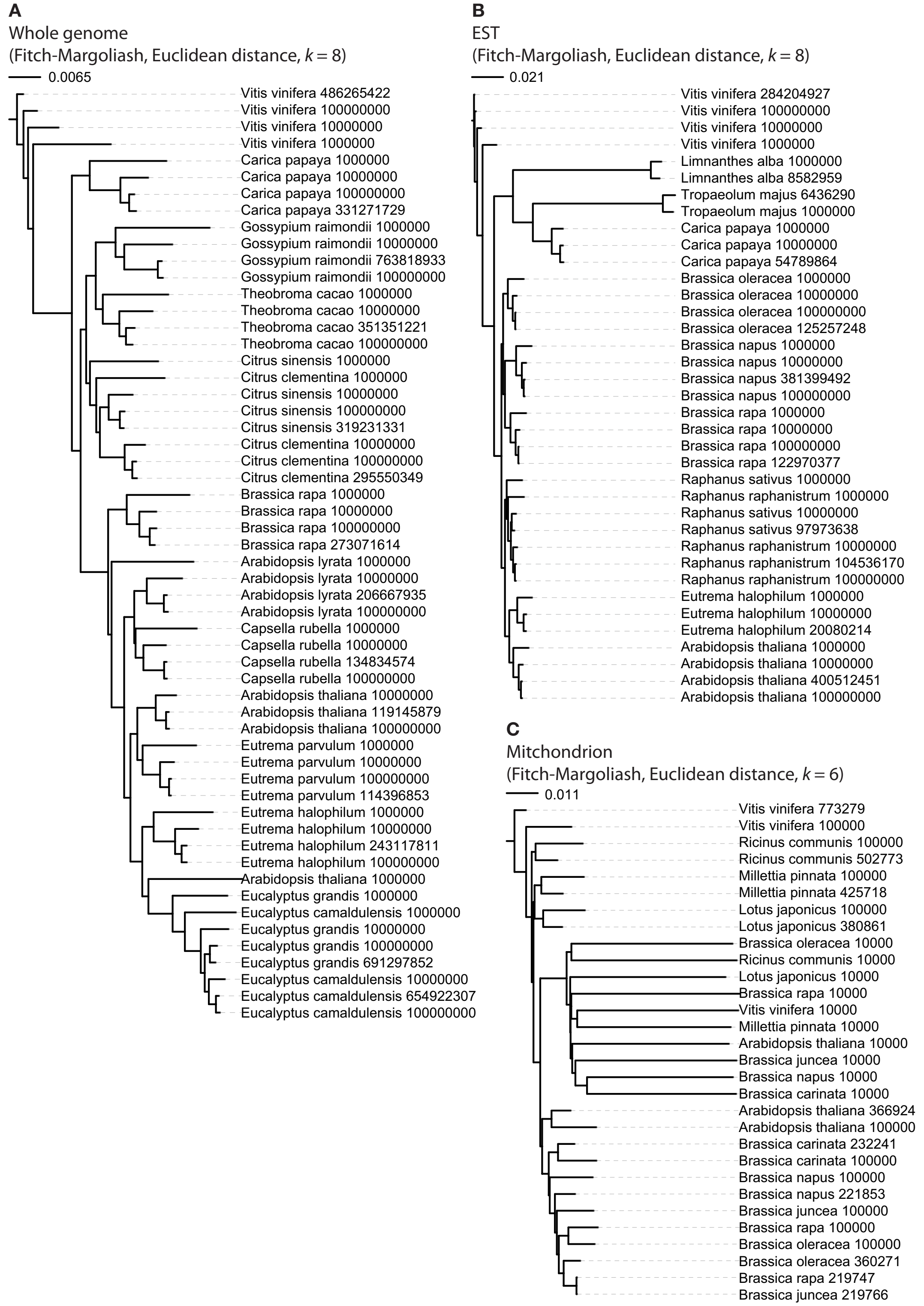
A
Whole genome
(Fitch-Margoliash, Euclidean distance, k = 8)
0.0065
Vitis vinifera 486265422
Vitis vinifera 100000000
Vitis vinifera 10000000
Vitis vinifera 1000000
Carica papaya 1000000
Carica papaya 10000000
Carica papaya 100000000
Carica papaya 331271729
Gossypium raimondii 1000000
Gossypium raimondii 10000000
Gossypium raimondii 763818933
Gossypium raimondii 100000000
Theobroma cacao 1000000
Theobroma cacao 10000000
Theobroma cacao 351351221
Theobroma cacao 100000000
Citrus sinensis 1000000
Citrus clementina 1000000
Citrus sinensis 10000000
Citrus sinensis 100000000
Citrus sinensis 319231331
Citrus clementina 10000000
Citrus clementina 100000000
Citrus clementina 295550349
Brassica rapa 1000000
Brassica rapa 10000000
Brassica rapa 100000000
Brassica rapa 273071614
Arabidopsis lyrata 1000000
Arabidopsis lyrata 10000000
Arabidopsis lyrata 206667935
Arabidopsis lyrata 100000000
Capsella rubella 1000000
Capsella rubella 10000000
Capsella rubella 134834574
Capsella rubella 100000000
Arabidopsis thaliana 10000000
Arabidopsis thaliana 119145879
Arabidopsis thaliana 100000000
Eutrema parvulum 1000000
Eutrema parvulum 10000000
Eutrema parvulum 100000000
Eutrema parvulum 114396853
Eutrema halophilum 1000000
Eutrema halophilum 10000000
Eutrema halophilum 243117811
Eutrema halophilum 100000000
Arabidopsis thaliana 1000000
Eucalyptus grandis 1000000
Eucalyptus camaldulensis 1000000
Eucalyptus grandis 10000000
Eucalyptus grandis 100000000
Eucalyptus grandis 691297852
Eucalyptus camaldulensis 10000000
Eucalyptus camaldulensis 654922307
Eucalyptus camaldulensis 100000000
B
EST
(Fitch-Margoliash, Euclidean distance, k = 8)
0.021
Vitis vinifera 284204927
Vitis vinifera 100000000
Vitis vinifera 10000000
Vitis vinifera 1000000
Limnanthes alba 1000000
Limnanthes alba 8582959
Tropaeolum majus 6436290
Tropaeolum majus 1000000
Carica papaya 1000000
Carica papaya 10000000
Carica papaya 54789864
Brassica oleracea 1000000
Brassica oleracea 10000000
Brassica oleracea 100000000
Brassica oleracea 125257248
Brassica napus 1000000
Brassica napus 10000000
Brassica napus 381399492
Brassica napus 100000000
Brassica rapa 1000000
Brassica rapa 10000000
Brassica rapa 100000000
Brassica rapa 122970377
Raphanus sativus 1000000
Raphanus raphanistrum 1000000
Raphanus sativus 10000000
Raphanus sativus 97973638
Raphanus raphanistrum 10000000
Raphanus raphanistrum 104536170
Raphanus raphanistrum 100000000
Eutrema halophilum 1000000
Eutrema halophilum 10000000
Eutrema halophilum 20080214
Arabidopsis thaliana 1000000
Arabidopsis thaliana 10000000
Arabidopsis thaliana 400512451
Arabidopsis thaliana 100000000
C
Mitchondrion
(Fitch-Margoliash, Euclidean distance, k = 6)
0.011
Vitis vinifera 773279
Vitis vinifera 100000
Ricinus communis 100000
Ricinus communis 502773
Millettia pinnata 100000
Millettia pinnata 425718
Lotus japonicus 100000
Lotus japonicus 380861
Brassica oleracea 10000
Ricinus communis 10000
Lotus japonicus 10000
Brassica rapa 10000
Vitis vinifera 10000
Millettia pinnata 10000
Arabidopsis thaliana 10000
Brassica juncea 10000
Brassica napus 10000
Brassica carinata 10000
Arabidopsis thaliana 366924
Arabidopsis thaliana 100000
Brassica carinata 232241
Brassica carinata 100000
Brassica napus 100000
Brassica napus 221853
Brassica juncea 100000
Brassica rapa 100000
Brassica oleracea 100000
Brassica oleracea 360271
Brassica rapa 219747
Brassica juncea 219766

FIGURE 2 | Phylogenetic trees to reveal the potential influence of sequence length. For each dataset sub-sequences with defined lengths were generated and FCGRs calculated. The lengths of the sequences were supposed to be sufficient for reliable tree reconstructions if the datasets generated from the same species grouped together. For whole genome **(A)** and EST **(B)** data 10,000,000 nt should be sufficient while the full-length mitochondrial genomes **(C)** are needed for reliable tree reconstructions.

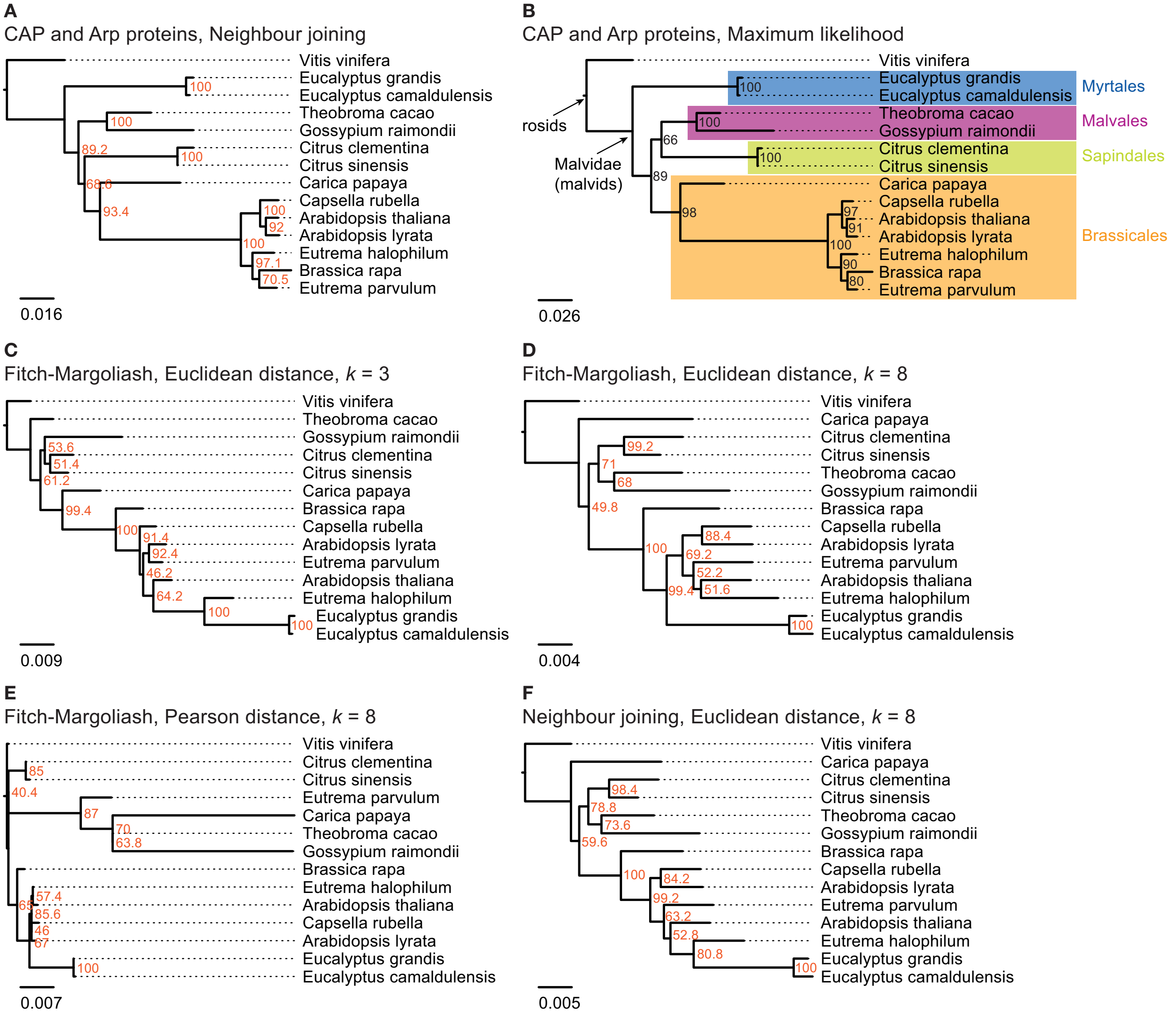

FIGURE 3 | The trees in (A,B) are based on a multiple sequence alignment of manually assembled CAP and Arp2/3 protein sequences. The trees were calculated using the Neighbor joining and the Maximum likelihood method, respectively, with 1,000 bootstraps for each tree. In **(C–F)** phylogenetic trees were generated applying different methods on FCGRs of whole genome sequence data of species of the taxon Malvidae. In **(C–E)** the Fitch–Margoliash algorithm was used to calculate trees for 500 re-sampled datasets. Subsequently, a consensus tree was built and branch lengths were calculated based on the fixed consensus tree. The method used for the distance calculation and the resolution of the FCGRs are given on top of the trees. In **(F)** the Neighbor joining algorithm was used to calculate the tree.

to root the trees. A species tree including all these organisms is not available. For comparison we therefore reconstructed trees of these species based on the alignment of the concatenated protein sequences of the actin CAP, Arp2, and Arp3 proteins (**Figures 3A,B**). The trees based on the NJ and ML methods are almost identical and differ only in the grouping of the two *Citrus* species (Sapindales clade) as independent clade (NJ, **Figure 3A**) or as sister clade of the Malvales (ML, **Figure 3B**). While the bootstrap support for all branchings is high, the support for the grouping of the *Citrus* clade is low in both trees (68.6% in the NJ and 66% in the ML tree, respectively). Both trees are in general agreement with phylogenetic analyses of the mitochondrial matR proteins (Zhu et al., 2007) and 61 chloroplast protein-coding genes (Bausher et al., 2006), and the combined analysis of 10 plastid and 2 nuclear (18S and 26S rDNA) genes (Cantino et al., 2007) that also show different groupings of the Sapindales clade. All trees agree with the grouping of the Malvales, Sapindales, and Brassicales into one clade and the grouping of the Myrtales as a sister clade, *C. papaya* being the most divergent of the analyzed Brassicales species and *C. rubella* being the closest relative of the *Arabidopsis* species. Except for the grouping of the two *Citrus* species the topology of the tree based on the ubiquitous cytoskeletal proteins CAP and Arp2/3 can thus be regarded as reference.

The resulting phylogenetic trees of the FCGRs differ as a function of data and methods used (**Figures 3C–F**). We reconstructed two trees based on the Euclidean distance and the Fitch–Margoliash algorithm but based on FCGRs with different resolution ($k = 3$ and $k = 8$ in **Figures 3C,D**, respectively), a tree using a different method for the distance calculation, the Pearson distance (**Figure 3E**), and a tree by applying a different method for the tree reconstruction, the NJ method (**Figure 3F**). In general, the trees agree with the reference tree except for the *Eucalyptus* species, which are either placed as sister group to *E. halophilum* (**Figures 3C,F**) or at the base of the Brassicales (**Figures 3D,E**) and thus far from their position according to the reference tree. In addition, *T. cacao* in **Figure 2C**, *C. papaya* in **Figures 3D–F**, and *E. parvulum* in **Figure 2E** are in wrong positions. None of the combinations of methods and data resulted in a correct resolution of the very closely related *Arabidopsis*, *Eutrema*, and *Capsella* species.

The tree based on the Pearson distance method (**Figure 3E**) contains the most deviations from the reference tree and this method therefore seems to be the least appropriate for reconstructing phylogenetic trees of whole genome sequences. This observation is in accordance with Wang et al. (2005). In addition, the bootstrap values do not provide reasonable support for most of the branchings except for the monophyly of the *Citrus* and the *Eucalyptus* clades. The trees based on high-resolution FCGRs ($k = 8$) using the Euclidean distance method (**Figures 3D,F**) have identical topologies except for the *Eucalyptus* outliers. In both trees *C. papaya* is placed as closest species to *V. vinifera* and not at the base of the Brassicales, *A. thaliana* grouped to the *Eutrema* species instead to its closest relative *A. lyrata*, and *B. rapa* is found at the base of the Brassicales instead of grouping to the *Eutrema* species. However, the misplacement of *Carica* and *A. thaliana* is not well supported (bootstrap values of 50–60%). Thus, the considerably faster NJ algorithm is a good alternative to the Fitch–Margoliash algorithm if run time is important. In contrast, the phylogenetic tree based on the low-resolution FCGRs ($k = 3$) contains more differences compared to the reference tree (**Figure 3C**).

EST DATA ANALYSIS

For this analysis related species of *B. rapa* were chosen, for which more than 1,000 EST entries are available in the EST database of NCBI. There are ten species that belong to the Brassicales taxon and match this criteria: *A. thaliana*, *Brassica napus*, *Brassica oleracea*, *B. rapa*, *C. papaya*, *E. halophilum*, *Limnanthes alba*, *Raphanus raphanistrum*, *Raphanus sativus*, and *Tropaeolum majus* (**Table 1**). Again, *V. vinifera* was included as outgroup. The trees reconstructed from the FCGRs of the EST datasets are shown in **Figure 4**. The tree based on the Pearson distance and calculated with the Fitch–Margoliash algorithm (**Figure 4C**) shows many deviations from the known relationships of the species but also low support for the branchings. Like for the whole genome analysis, the Pearson distance concept is not appropriate for the reconstruction of reliable phylogenetic trees based on FCGRs. The trees based on the Euclidean distance (**Figures 4A–D**) have almost identical (low-resolution $k = 3$ compared to high-resolution data $k = 8$) to identical topologies (Fitch–Margoliash compared to NJ algorithm). Especially the species of the Brassicaceae clade are well resolved and their topology is highly supported in all trees. The Limnanthaceae, Tropaeolaceae, and Caricaceae are sister groups of the Brassicaceae. To our knowledge there is no highly resolved tree of these groups available that we could use as reference. Based on our experience with the whole genome data we suppose that the trees based on high-resolution data represent the more reliable topologies.

MITOCHONDRIAL GENOME ANALYSIS

For this analysis close relatives of *B. rapa* were chosen, for which sequenced mitochondria are available from NCBI. There were nine species in the Rosids taxon, whose mitochondrial genome sequences were available: *A. thaliana*, *Brassica carinata*, *Brassica juncea*, *B. napus*, *B. oleracea*, *B. rapa*, *Lotus japonicus*, *Millettia pinnata*, and *Ricinus communis* (**Table 1**). The mitochondrial genome of *V. vinifera* was used as outgroup. In contrast to the analyses of the other datasets, the trees based on the FCGRs of the mitochondrial genomes were very similar for the four different methods (**Figure 5**). Especially the sub-branches containing the five closely related *Brassica* species show exactly the same topology supported by high bootstrap values. While the topology of the Brassicales subfamily tree is well resolved the grouping of the Fabales *L. japonicus* and *M. pinnata* and the Malpighiales *R. communis*, which all belong to the fabids, is different in the four trees. Here, the trees based on the Euclidean distance with high-resolution FCGRs ($k = 8$) have the same well supported topology grouping the Fabales together (**Figures 5B,D**) independently which method has been used for the tree reconstruction. This is in agreement with the results from the whole genome and EST analysis that the use of FCGRs with high-resolution results in more reasonable trees, and that the Euclidean method for the calculations of the distances is more appropriate than the Pearson method.

COMPUTATIONAL RESOURCE COMPARISON

The algorithm to calculate the CGRs and FCGRs has linear time complexity $O(L)$ and space constant complexity $O(1)$, where L is the length of the nucleotide sequence. In the case of whole genomes, the calculation of the CGRs and FCGRs took about 7,600 s for each genome, for EST data 2,800 s for each species, and 140 s for each mitochondrial genome. The time the algorithm needs to calculate the phylogenetic trees mainly depends on the distance matrix calculated for each species against each other species. This calculation has time complexity $O(4^k s^2)$ and space complexity $O(s^2)$, where s is the number of species and k is the length of the oligonucleotide. The reconstructions of the phylogenetic trees took 98 s for $k = 8$ and the whole genome datasets ($k = 7$: 41 s, $k = 6$: 10 s, $k = 3$: 4 s), 86 s with $k = 8$ for the EST datasets ($k = 7$: 22 s, $k = 3$: 2 s) and 58 s for $k = 8$ and the mitochondrial genome datasets ($k = 7$: 13 s, $k = 3$: 1 s). These values refer to one round of bootstrapping. For comparison, one of the fastest whole genome alignment tools, called Mugsy, needs 45,000 s (ca. 12 h) to align the human and the mouse genomes (Angiuoli and Salzberg, 2011). However, whole genomes can only be aligned if they are from closely related species and, to our knowledge, phylogenies of multiple sequence alignments of the whole genomes from different eukaryotes have not been reconstructed yet.

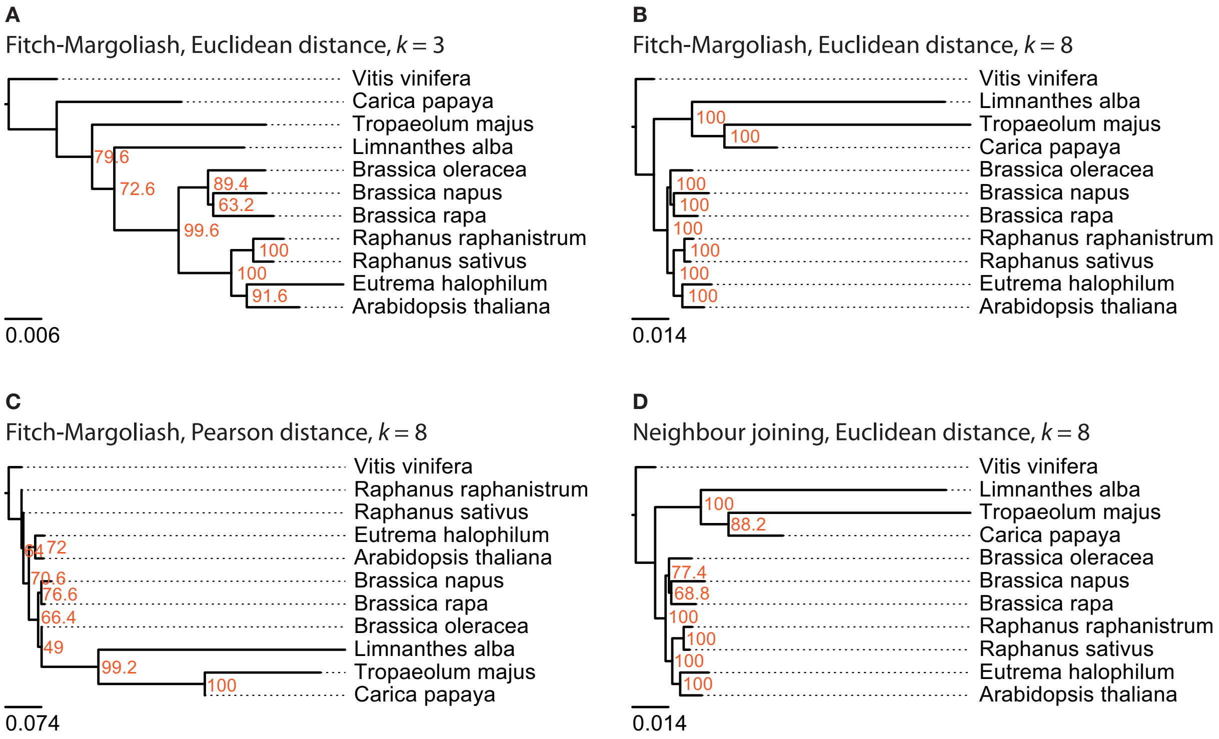

FIGURE 4 | The phylogenetic trees were generated applying different methods on FCGRs of public available EST data of the Brassicales taxon. In **(A–C)** the Fitch–Margoliash algorithm was used to calculate trees for 500 re-sampled datasets. Subsequently, a consensus tree was built and branch lengths were calculated based on the fixed consensus tree. The methods used for the distance calculation and the resolution of the FCGRs are given on top of the trees. In **(D)** the Neighbor joining algorithm was used to calculate the tree.

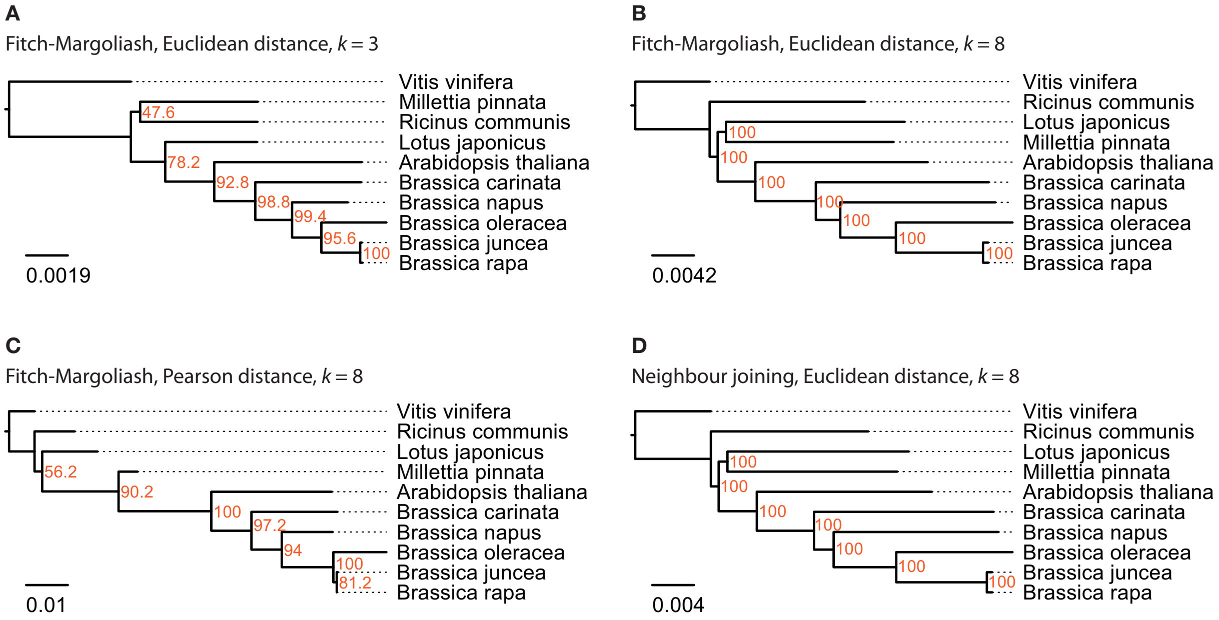

FIGURE 5 | The phylogenetic trees were generated applying different methods on FCGRs of available mitochondrial genome sequence data of the Rosids taxon. In **(A–C)** the Fitch–Margoliash algorithm was used to calculate trees for 500 re-sampled datasets. Subsequently, a consensus tree was built and branch lengths were calculated based on the fixed consensus tree. The method used for the distance calculation and the resolution of the FCGRs are given on top of the trees. In **(D)** the Neighbor joining algorithm was used to calculate the tree.

DISCUSSION

In general, phylogenetic trees of species are reconstructed from amino acid or nucleotide sequence data, by comparing morphological characteristics, or by combining these data. While most of the sequence-based analyses are built on single genes, concatenated sequences are increasingly used, which could consist of even whole transcriptomes (phylogenomics). Here, we wanted to reconstruct the phylogeny of selected Brassicales species based on alignment-free sequence data. As approach we chose CGRs, which are scale-independent representations for genomic sequences (Jeffrey, 1990). Because CGRs are unique fingerprints of the corresponding sequences they cannot be compared directly. To reconstruct phylogenetic trees we therefore generated FCGRs at different resolutions. For the calculation of the distances between FCGRs we used the Euclidean (a geometric distance) and the Pearson (a statistical distance) method, and trees were reconstructed with the Fitch–Margoliash and the NJ algorithm.

Because of their different characteristics we compared three types of nucleotide sequences, nuclear genome sequences, mitochondrial genome sequences, and EST reads. Nuclear and mitochondrial genomes have been shown to have different GC contents and codon usage patterns (Zhang et al., 2007). EST data just comprise the exons and thus only part of the nuclear genome sequences. In addition EST data are potentially biased toward highly abundant genes and 5′- and 3′-terminal sequences. In order to reduce this bias we decided to include only those species for which at least 1,000 EST clones were available. Unfortunately, appropriate species from the Brassicales clade are not available for which all three types of nucleotide data have been sequenced. Therefore, we compared different sets of species for the three data types. Also, it is not known whether the mitochondrial genome data have been extracted from the whole genome datasets. As most of these are denoted as "draft assembly" we assume that the whole genome datasets still contain mitochondrial data. However, because of the very small size of the mitochondrial genomes compared to the nuclear genomes the results should be identical to those obtained from pure nuclear genome data. We would have liked to compare the results of each type of nucleotide sequence with the results of combined datasets but appropriate sequence data is not available. However, the EST and mitochondrial data do not comprise 1% of the whole genome data (**Table 1**) and a combined analysis should therefore be dominated by and be identical to the whole genome data.

The mitochondrial and whole genomes of the analyzed Brassicales species are of considerably different size, and different amounts of EST data are available. FCGRs naturally depend on the presence and frequency of the respective oligonucleotides and thus on the length of the analyzed sequence. For a reasonable result it is therefore essential to find the best balance between sequence length and FCGR resolution (oligonucleotide length), which represents the number of data available for the tree calculations and is also the main determinant for computing time. To exclude that the lengths of the concatenated sequences have an influence on the phylogenetic tree reconstructions of the Brassicales species at high FCGR resolution we calculated trees including the full-lengths sequences and specific defined subsets (**Figure 2**). At the resolution of octanucleotides, all partial sequences of whole genome assemblies containing more than 10 million nucleotides of each species group together while sets with 1 million nucleotides result in the ambiguous grouping of some species. In contrast, one million nucleotides of EST data, which correspond to the exon sequences, already result in consistent monophyly of all datasets of each species. Remarkably, this holds even true for the closely related *Brassica* species. The mitochondrial genomes of the analyzed species have sizes of 220–780 kbp. Thus, at the resolution of hexanucleotides it is not surprising that many oligonucleotides do not exist in sub-sections of 10 kbp leading to the artificial attraction of all these datasets in the reconstructed tree. Also, datasets of 100 kbp of the different *Brassica* species do not consistently group to the full-length mitochondrial genomes. Therefore, for mitochondrial data the resolution has to be reduced or full-length data to be used. As outgroup we choose *V. vinifera* in all analyses.

According to the diArk database, whole genome assemblies are available for 34 species belonging to the Malvidae/malvids (Hammesfahr et al., 2011). Twenty-two of them are *A. thaliana* strains of which we only included the reference strain into the analysis. A species tree including all these sequenced Malvidae is not available. Therefore, we assembled and annotated the CAPs α- and β-CAP, and the actin-related proteins Arp2 and Arp3 to generate a reference tree. The CAP and Arp proteins have been chosen for the reference tree because they are ubiquitous and well conserved in all eukaryotes (Goley and Welch, 2006; Cooper and Sept, 2008), and duplicates were most probably removed after the many whole genome duplication events that happened in plant evolution (Van de Peer, 2011). For example, the *A. thaliana* genome has experienced two duplications since its divergence from *Carica* (Tang et al., 2008), but has retained single copies of the CAP and Arp genes (Hammesfahr and Kollmar, 2012). Nevertheless, duplicated CAP and Arp2/3 genes have been identified in the *B. rapa* and *G. raimondii* genomes that are, however, the result of species-specific duplications. Only one of each duplicate has been used in this analysis. The phylogenetic tree of the concatenated CAP and Arp proteins is in agreement with other recent analyses containing part of the species (Bausher et al., 2006; Zhu et al., 2007; Wang et al., 2009) and can thus be regarded as reference tree. Compared to this reference tree, the FCGR tree based on the Pearson distance displays the most discrepancies followed by the tree based on low-resolution data ($k=3$, trinucleotides). In addition, most of the branchings have low bootstrap values. The trees based on high-resolution data ($k=8$,65,536 data points) and the Euclidean distance method show overall agreement with the reference trees independent of the method used for the tree reconstruction. Notably, *C. papaya* and *B. rapa* group wrongly, although both are only shifted by one branching event. Most surprisingly, the *Eucalyptus* species are completely wrongly grouped in all FCGR trees. Their exclusion from the tree calculation did not change the grouping of the other species (data not shown). However, the grouping of the Myrtales branch, which contains the *Eucalyptus* species, is different in all published trees (Bausher et al., 2006; Zhu et al., 2007; Wang et al., 2009) and their wrong placement in the FCGR trees might be due to some unknown characteristics of the genomes. Probably, they would group better, if species from other branches like the Crossosomatales, Geraniales, and Fabidae branches were included in the analysis. The phylogenetic trees of the FCGRs of the mitochondrial genomes are very similar independently of the resolution, distance measure, and tree reconstruction method. Therefore either the species selection was fortunate or mitochondrial genome data is less sensitive with respect to these parameters.

When working with the EST data we observed disproportionate high frequencies for poly-A and poly-T oligonucleotides in the FCGRs. Probably, the poly-A tails were not consistently removed during the cDNA library construction. For low-resolution data (up to $k = 5$) the differences of the frequencies of these oligonucleotides to the next-highest values were not large enough to considerably bias the phylogenetic tree reconstructions. However, the topologies of trees based on high-resolution data ($k > 5$) are strongly disturbed. Therefore, we set the values for the frequencies of the poly-A and poly-T oligonucleotides to zero before we started the tree calculations. The artificial oligonucleotides generated at the boundaries of the concatenated EST reads apparently do not influence the resulting trees. The phylogeny of the *Brassica* species is slightly different compared to that obtained from the mitochondrial genome data. The genus *Brassica* includes 41 species (Velasco and Fernández-Martínez, 2010) the six with the highest economic importance being *B. rapa* (A), *Brassica nigra* (B), *Brassia olerácea* (C), *B. napus* (AC), *B. juncea* (AB), and *B. carinata* (BC). The first three comprise the three elementary species while the other three are amphidiploids that originated from natural hybridizations between two of the elementary species (Velasco and Fernández-Martínez, 2010). Thus the amphidiploid EST data contain mixtures of the hybridized species and dependent on which part is overrepresented in the data they will look closer related to one of their parent species. Although the distance in the phylogenetic tree is very small, *B. napus* seems to be closer to *B. rapa* based on the mitochondrial data. Based on the EST data, the hybrids *B. juncea* and *B. carinata* are more divergent than the parent species *B. rapa* and *B. oleracea.* Probably the part of the more divergent parent species *B. nigra* is dominating in this case.

In general we could show that FCGRs are well suited to phylogenetically group plant genomes and exonomes from even closely related species. We assume that FCGRs could also be used to group all eukaryotes provided that a balanced set of species from all lineages is taken. This has in part already been demonstrated on the phylogeny of 26 mitochondrial genomes of which only three were placed completely wrong when using the Euclidean distance method (Wang et al., 2005). However, this analysis was solely based on data from mitochondrions and biased against fish and mammalian species. Our analysis of the Brassicales clade has shown that high-resolution data (octanucleotides and longer sequences) result in better tree topologies and higher support for branchings. Trees based on the Pearson distance, which is a statistical distance measure, are less reliable than those based on Euclidean distances. The Fitch–Margoliash and NJ algorithms result in similar to identical trees. We have shown for the first time that the bootstrap concept to determine the support of the branchings in the tree, which is well established for trees based on sequence alignments since decades ("taxon-by-character" data matrix; Felsenstein, 1985), can also be applied to trees based on FCGRs. In another study it has been shown that although longer word lengths could reveal the correct clustering of the HIV-I subtypes in contrast to shorter word lengths (Pandit and Sinha, 2010) the grouping within the subtypes was always different. Also in this case a bootstrap analysis could have helped in the interpretation of the various branchings and we would recommend applying the bootstrap concept to all phylogenies based on FCGRs. FCGRs are fast to calculate and could be used in combination with alignment based data and morphological characteristics to improve the phylogenetic classification in ambiguous cases.

ACKNOWLEDGMENTS

We would like to thank Björn Hammesfahr for helpful suggestions and discussion, and Prof. Christian Griesinger for continuous support.

REFERENCES

Almeida, J. S., Carriço, J. A., Maretzek, A., Noble, P. A., and Fletcher, M. (2001). Analysis of genomic sequences by chaos game representation. *Bioinformatics* 17, 429–437.

Angiuoli, S. V., and Salzberg, S. L. (2011). Mugsy: fast multiple alignment of closely related whole genomes. *Bioinformatics* 27, 334–342.

Arabidopsis Genome Initiative. (2000). Analysis of the genome sequence of the flowering plant Arabidopsis thaliana. *Nature* 408, 796–815.

Argout, X., Salse, J., Aury, J.-M., Guiltinan, M. J., Droc, G., Gouzy, J., Allegre, M., Chaparro, C., Legavre, T., Maximova, S. N., Abrouk, M., Murat, F., Fouet, O., Poulain, J., Ruiz, M., Roguet, Y., Rodier-Goud, M., Barbosa-Neto, J. F., Sabot, F., Kudrna, D., Ammiraju, J. S., Schuster, S. C., Carlson, J. E., Sallet, E., Schiex, T., Dievart, A., Kramer, M., Gelley, L., Shi, Z., Bérard, A., Viot, C., Boccara, M., Risterucci, A. M., Guignon, V., Sabau, X., Axtell, M. J., Ma, Z., Zhang, Y., Brown, S., Bourge, M., Golser, W., Song, X., Clement, D., Rivallan, R., Tahi, M., Akaza, J. M., Pitollat, B., Gramacho, K., D'Hont, A., Brunel, D., Infante, D., Kebe, I., Costet, P., Wing, R., McCombie, W. R., Guiderdoni, E., Quetier, F., Panaud, O., Wincker, P., Bocs, S., and Lanaud, C. (2011). The genome of Theobroma cacao. *Nat. Genet.* 43, 101–108.

Basu, S., Pan, A., Chitra., and Das, J. (1997). Chaos game representation of proteins. *J. Mol. Graph. Model.* 15, 279–289.

Bausher, M. G., Singh, N. D., Lee, S.-B., Jansen, R. K., and Daniell, H. (2006). The complete chloroplast genome sequence of Citrus sinensis (L.) Osbeck var "Ridge Pineapple": organization and phylogenetic relationships to other angiosperms. *BMC Plant Biol.* 6, 21. doi:10.1186/1471-2229-6-21

Blair, C., and Murphy, R. W. (2011). Recent trends in molecular phylogenetic analysis: where to next? *J. Hered.* 102, 130–138.

Campbell, A., Mrázek, J., and Karlin, S. (1999). Genome signature comparisons among prokaryote, plasmid, and mitochondrial DNA. *Proc. Natl. Acad. Sci. U.S.A.* 96, 9184–9189.

Cantino, P. D., Doyle, J. A., Graham, S. W., Judd, W. S., Olmstead, R. G., Soltis, D. E., Soltis, P. S., and Donoghue, M. J. (2007). Towards a phylogenetic nomenclature of Tracheophyta. *Taxon* 56, 1E–44E.

Chenna, R., Sugawara, H., Koike, T., Lopez, R., Gibson, T. J., Higgins, D. G., and Thompson, J. D. (2003). Multiple sequence alignment with the Clustal series of programs. *Nucleic Acids Res.* 31, 3497–3500.

Cooper, J. A., and Sept, D. (2008). New insights into mechanism and regulation of actin capping protein. *Int. Rev. Cell Mol. Biol.* 267, 183–206.

Darriba, D., Taboada, G. L., Doallo, R., and Posada, D. (2011). ProtTest 3: fast selection of best-fit models of protein evolution. *Bioinformatics* 27, 1164–1165.

Dassanayake, M., Oh, D.-H., Haas, J. S., Hernandez, A., Hong, H., Ali, S., Yun, D.-J., Bressan, R. A., Zhu, J.-K., Bohnert, H. J., and Cheeseman, J. M. (2011). The genome of the extremophile crucifer Thellungiella parvula. *Nat. Genet.* 43, 913–918.

Deschavanne, P. J., Giron, A., Vilain, J., Fagot, G., and Fertil, B. (1999). Genomic signature: characterization and classification of species assessed by chaos game representation of sequences. *Mol. Biol. Evol.* 16, 1391–1399.

Dewey, C. N. (2012). Whole-genome alignment. *Methods Mol. Biol.* 855, 237–257.

Domazet-Loso, M., and Haubold, B. (2009). Efficient estimation of pairwise distances between genomes. *Bioinformatics* 25, 3221–3227.

Edwards, S. V., Fertil, B., Giron, A., and Deschavanne, P. J. (2002). A genomic schism in birds revealed by phylogenetic analysis of DNA strings. *Syst. Biol.* 51, 599–613.

Felsenstein, J. (1985). Confidence limits on phylogenies: an approach using the bootstrap. *Evolution* 39, 783–791.

Goley, E. D., and Welch, M. D. (2006). The ARP2/3 complex: an actin nucleator comes of age. *Nat. Rev. Mol. Cell Biol.* 7, 713–726.

Hammesfahr, B., and Kollmar, M. (2012). Evolution of the eukaryotic dynactin complex, the activator of cytoplasmic dynein. *BMC Evol. Biol.* 12, 95. doi:10.1186/1471-2148-12-95

Hammesfahr, B., Odronitz, F., Hellkamp, M., and Kollmar, M. (2011). diArk 2.0 provides detailed analyses of the ever increasing eukaryotic genome sequencing data. *BMC Res. Notes* 4, 338. doi:10.1186/1756-0500-4-338

Hu, T. T., Pattyn, P., Bakker, E. G., Cao, J., Cheng, J.-F., Clark, R. M., Fahlgren, N., Fawcett, J. A., Grimwood, J., Gundlach, H., Haberer, G., Hollister, J. D., Ossowski, S., Ottilar, R. P., Salamov, A. A., Schneeberger, K., Spannagl, M., Wang, X., Yang, L., Nasrallah, M. E., Bergelson, J., Carrington, J. C., Gaut, B. S., Schmutz, J., Mayer, K. F., van de Peer, Y., Grigoriev, I. V., Nordborg, M., Weigel, D., and Guo, Y. L. (2011). The Arabidopsis lyrata genome sequence and the basis of rapid genome size change. *Nat. Genet.* 43, 476–481.

Jaillon, O., Aury, J.-M., Noel, B., Policriti, A., Clepet, C., Casagrande, A., Choisne, N., Aubourg, S., Vitulo, N., Jubin, C., Vezzi, A., Legeai, F., Hugueney, P., Dasilva, C., Horner, D., Mica, E., Jublot, D., Poulain, J., Bruyère, C., Billault, A., Segurens, B., Gouyvenoux, M., Ugarte, E., Cattonaro, F., Anthouard, V., Vico, V., Del Fabbro, C., Alaux, M., Di Gaspero, G., Dumas, V., Felice, N., Paillard, S., Juman, I., Moroldo, M., Scalabrin, S., Canaguier, A., Le Clainche, I., Malacrida, G., Durand, E., Pesole, G., Laucou, V., Chatelet, P., Merdinoglu, D., Delledonne, M., Pezzotti, M., Lecharny, A., Scarpelli, C., Artiguenave, F., Pè, M. E., Valle, G., Morgante, M., Caboche, M., Adam-Blondon, A. F., Weissenbach, J., Quétier, F., Wincker, P., and French-Italian Public Consortium for Grapevine Genome Characterization. (2007). The grapevine genome sequence suggests ancestral hexaploidization in major angiosperm phyla. *Nature* 449, 463–467.

Jeffrey, H. J. (1990). Chaos game representation of gene structure. *Nucleic Acids Res.* 18, 2163–2170.

Jones, D. T., Taylor, W. R., and Thornton, J. M. (1992). The rapid generation of mutation data matrices from protein sequences. *Comput. Appl. Biosci.* 8, 275–282.

Joseph, J., and Sasikumar, R. (2006). Chaos game representation for comparison of whole genomes. *BMC Bioinformatics* 7, 243. doi:10.1186/1471-2105-7-243

Li, M., Badger, J. H., Chen, X., Kwong, S., Kearney, P., and Zhang, H. (2001). An information-based sequence distance and its application to whole mitochondrial genome phylogeny. *Bioinformatics* 17, 149–154.

Ming, R., Hou, S., Feng, Y., Yu, Q., Dionne-Laporte, A., Saw, J. H., Senin, P., Wang, W., Ly, B. V., Lewis, K. L. T., Salzberg, S. L., Feng, L., Jones, M. R., Skelton, R. L., Murray, J. E., Chen, C., Qian, W., Shen, J., Du, P., Eustice, M., Tong, E., Tang, H., Lyons, E., Paull, R. E., Michael, T. P., Wall, K., Rice, D. W., Albert, H., Wang, M. L., Zhu, Y. J., Schatz, M., Nagarajan, N., Acob, R. A., Guan, P., Blas, A., Wai, C. M., Ackerman, C. M., Ren, Y., Liu, C., Wang, J., Wang, J., Na, J. K., Shakirov, E. V., Haas, B., Thimmapuram, J., Nelson, D., Wang, X., Bowers, J. E., Gschwend, A. R., Delcher, A. L., Singh, R., Suzuki, J. Y., Tripathi, S., Neupane, K., Wei, H., Irikura, B., Paidi, M., Jiang, N., Zhang, W., Presting, G., Windsor, A., Navajas-Pérez, R., Torres, M. J., Feltus, F. A., Porter, B., Li, Y., Burroughs, A. M., Luo, M. C., Liu, L., Christopher, D. A., Mount, S. M., Moore, P. H., Sugimura, T., Jiang, J., Schuler, M. A., Friedman, V., Mitchell-Olds, T., Shippen, D. E., dePamphilis, C. W., Palmer, J. D., Freeling, M., Paterson, A. H., Gonsalves, D., Wang, L., and Alam, M. (2008). The draft genome of the transgenic tropical fruit tree papaya (Carica papaya Linnaeus). *Nature* 452, 991–996.

Pandit, A., and Sinha, S. (2010). Using genomic signatures for HIV-1 sub-typing. *BMC Bioinformatics* 11(Suppl. 1), S26.

Pleissner, K. P., Wernisch, L., Oswald, H., and Fleck, E. (1997). Representation of amino acid sequences as two-dimensional point patterns. *Electrophoresis* 18, 2709–2713.

Sims, G. E., Jun, S.-R., Wu, G. A., and Kim, S.-H. (2009). Alignment-free genome comparison with feature frequency profiles (FFP) and optimal resolutions. *Proc. Natl. Acad. Sci. U.S.A.* 106, 2677–2682.

Stamatakis, A., Hoover, P., and Rougemont, J. (2008). A rapid bootstrap algorithm for the RAxML Web servers. *Syst. Biol.* 57, 758–771.

Tang, H., Wang, X., Bowers, J. E., Ming, R., Alam, M., and Paterson, A. H. (2008). Unraveling ancient hexaploidy through multiply-aligned angiosperm gene maps. *Genome Res.* 18, 1944–1954.

Van de Peer, Y. (2011). A mystery unveiled. *Genome Biol.* 12, 113.

Velasco, L., and Fernández-Martínez, J. M. (2010). “Other Brassicas,” in *Oil Crops Handbook of Plant Breeding*, eds J. Vollmann and I. Rajcan (Springer New York), 127–153.

Velasco, R., Zharkikh, A., Troggio, M., Cartwright, D. A., Cestaro, A., Pruss, D., Pindo, M., Fitzgerald, L. M., Vezzulli, S., Reid, J., Malacarne, G., Iliev, D., Coppola, G., Wardell, B., Micheletti, D., Macalma, T., Facci, M., Mitchell, J. T., Perazzolli, M., Eldredge, G., Gatto, P., Oyzerski, R., Moretto, M., Gutin, N., Stefanini, M., Chen, Y., Segala, C., Davenport, C., Demattè, L., Mraz, A., Battilana, J., Stormo, K., Costa, F., Tao, Q., Si-Ammour, A., Harkins, T., Lackey, A., Perbost, C., Taillon, B., Stella, A., Solovyev, V., Fawcett, J. A., Sterck, L., Vandepoele, K., Grando, S. M., Toppo, S., Moser, C., Lanchbury, J., Bogden, R., Skolnick, M., Sgaramella, V., Bhatnagar, S. K., Fontana, P., Gutin, A., Van de Peer, Y., Salamini, F., and Viola, R. (2007). A high quality draft consensus sequence of the genome of a heterozygous grapevine variety. *PLoS ONE* 2, e1326. doi:10.1371/journal.pone.0001326

Vinga, S., and Almeida, J. (2003). Alignment-free sequence comparison-a review. *Bioinformatics* 19, 513–523.

Wang, H., Moore, M. J., Soltis, P. S., Bell, C. D., Brockington, S. F., Alexandre, R., Davis, C. C., Latvis, M., Manchester, S. R., and Soltis, D. E. (2009). Rosid radiation and the rapid rise of angiosperm-dominated forests. *Proc. Natl. Acad. Sci. U.S.A.* 106, 3853–3858.

Wang, X., Wang, H., Wang, J., Sun, R., Wu, J., Liu, S., Bai, Y., Mun, J.-H., Bancroft, I., Cheng, F., Huang, S., Li, X., Hua, W., Wang, J., Wang, X., Freeling, M., Pires, J. C., Paterson, A. H., Chalhoub, B., Wang, B., Hayward, A., Sharpe, A. G., Park, B. S., Weisshaar, B., Liu, B., Li, B., Liu, B., Tong, C., Song, C., Duran, C., Peng, C., Geng, C., Koh, C., Lin, C., Edwards, D., Mu, D., Shen, D., Soumpourou, E., Li, F., Fraser, F., Conant, G., Lassalle, G., King, G. J., Bonnema, G., Tang, H., Wang, H., Belcram, H., Zhou, H., Hirakawa, H., Abe, H., Guo, H., Wang, H., Jin, H., Parkin, I. A., Batley, J., Kim, J. S., Just, J., Li, J., Xu, J., Deng, J., Kim, J. A., Li, J., Yu, J., Meng, J., Wang, J., Min, J., Poulain, J., Wang, J., Hatakeyama, K., Wu, K., Wang, L., Fang, L., Trick, M., Links, M. G., Zhao, M., Jin, M., Ramchiary, N., Drou, N., Berkman, P. J., Cai, Q., Huang, Q., Li, R., Tabata, S., Cheng, S., Zhang, S., Zhang, S., Huang, S., Sato, S., Sun, S., Kwon, S. J., Choi, S. R., Lee, T. H., Fan, W., Zhao, X., Tan, X., Xu, X., Wang, Y., Qiu, Y., Yin, Y., Li, Y., Du, Y., Liao, Y., Lim, Y., Narusaka, Y., Wang, Y., Wang, Z., Li, Z., Wang, Z., Xiong, Z., Zhang, Z., and Brassica rapa Genome Sequencing Project Consortium. (2011). The genome of the mesopolyploid crop species Brassica rapa. *Nat. Genet.* 43, 1035–1039.

Wang, Y., Hill, K., Singh, S., and Kari, L. (2005). The spectrum of genomic signatures: from dinucleotides to chaos game representation. *Gene* 346, 173–185.

Zhang, W., Zhou, J., Li, Z., Wang, L., Gu, X., and Zhong, Y. (2007). Comparative analysis of codon usage patterns among mitochondrion, chloroplast and nuclear genes in Triticum aestivum L. *J. Integr. Plant Biol.* 49, 246–254.

Zhu, X.-Y., Chase, M. W., Qiu, Y.-L., Kong, H.-Z., Dilcher, D. L., Li, J.-H., and Chen, Z.-D. (2007). Mitochondrial matR sequences help to resolve deep phylogenetic relationships in Rosids. *BMC Evol. Biol.* 7, 217.

Conflict of Interest Statement: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Received: 25 May 2012; accepted: 06 August 2012; published online: 29 August 2012.

*Citation: Hatje K and Kollmar M (2012) A phylogenetic analysis of the Brassicales clade based on an alignment-free sequence comparison method. Front. Plant Sci. **3**:192. doi: 10.3389/fpls.2012.00192*

This article was submitted to Frontiers in Plant Genetics and Genomics, a specialty of Frontiers in Plant Science.