Published online 4 January 2012

Nucleic Acids Research, 2012, Vol. 40, No. 7 ***e49***
doi:10.1093/nar/gkr1293

MCScanX: a toolkit for detection and evolutionary analysis of gene synteny and collinearity

Yupeng Wang[1,2], Haibao Tang[1,3,4], Jeremy D. DeBarry[5], Xu Tan[1,3], Jingping Li[1,2], Xiyin Wang[1,6], Tae-ho Lee[1], Huizhe Jin[1,2], Barry Marler[1], Hui Guo[1,3], Jessica C. Kissinger[2,5,7] and Andrew H. Paterson[1,2,3,7,8,*]

[1]Plant Genome Mapping Laboratory, [2]Institute of Bioinformatics, [3]Department of Plant Biology, University of Georgia, Athens, GA 30602, [4]J. Craig Venter Institute, 9704 Medical Center Drive, Rockville, MD 20850, [5]Center for Tropical and Emerging Global Diseases, University of Georgia, Athens, GA 30602, USA, [6]Center for Genomics and Computational Biology, School of Life Sciences and School of Sciences, Hebei United University, Tangshan, Hebei 063009, China, [7]Department of Genetics and [8]Department of Crop and Soil Sciences, University of Georgia, Athens, GA 30602, USA

Received September 22, 2011; Revised December 11, 2011; Accepted December 15, 2011

ABSTRACT

MCScan is an algorithm able to scan multiple genomes or subgenomes in order to identify putative homologous chromosomal regions, and align these regions using genes as anchors. The *MCScanX* toolkit implements an adjusted MCScan algorithm for detection of synteny and collinearity that extends the original software by incorporating 14 utility programs for visualization of results and additional downstream analyses. Applications of *MCScanX* to several sequenced plant genomes and gene families are shown as examples. *MCScanX* can be used to effectively analyze chromosome structural changes, and reveal the history of gene family expansions that might contribute to the adaptation of lineages and taxa. An integrated view of various modes of gene duplication can supplement the traditional gene tree analysis in specific families. The source code and documentation of *MCScanX* are freely available at http://chibba.pgml.uga.edu/mcscan2/.

INTRODUCTION

Comparative genomic studies often rely on the accurate identification of homology (genes that share a common evolutionary origin) within or across genomes. Homologous genes are further classified as either orthologous, if they were separated by a speciation event, or paralogous, if they were separated by a gene duplication event. Recently, comparisons between related eukaryotic genomes reveal various degrees to which homologous genes remain on corresponding chromosomes (synteny) and in conserved orders (collinearity) during evolution (1). Over evolutionary time, genomes have been shaped and dynamically restructured by several forces such as whole-genome duplication (WGD), segmental duplication, inversions and translocations (2–5). These forces have acted in various combinations and to differing degrees to result in taxonomic groups with different modes of genome structure modification and gene family expansion. For example, angiosperm (flowering plants) genomes appear more volatile than mammalian genomes (1). Angiosperm genomes show remarkable fluctuations in size and organization, even among close relatives, and all examined angiosperms have undergone one or more ancient WGD (6). In contrast, karyotype evolution among major vertebrate lineages appears to have been slower, with a single whole-genome duplication event ~500 million years ago (4,7). However, hundreds of invertebrates are paleopolyploids (8) and their rates of chromosomal rearrangement have been suggested to be almost twice that of vertebrates (1,9,10). Further, there is also a remarkable lack of synteny and high rate of rearrangement in the parasitic and pathogenic protistan phylum Apicomplxa compared to what is seen in vertebrates (11).

Traditionally, synteny was identified via the clustering of neighboring matching gene pairs, as implemented in various programs including ADHoRe (12), TEAM (13), LineUp (14), the Max-gap Clusters by Multiple Sequence

*To whom correspondence should be addressed. Tel: +1 706 583 0162; Fax: +1 706 583 0160; Email: paterson@plantbio.uga.edu

The authors wish it to be known that, in their opinion, the first two authors should be regarded as joint First Author.

Comparison (MCMuSeC) (15) and OrthoCluster (16). However, detection of synteny is often complicated by gene loss, tandem duplications, gene transpositions and chromosomal rearrangements, any of which may produce artifacts. Collinearity, a more specific form of synteny, requires conserved gene order. More recent methods apply dynamic programming to chains of pair-wise collinear genes, and often specify a certain scoring scheme that rewards the adjacent collinear gene pairs (or 'anchor genes') and penalizes the distance between anchor genes. This class of methods has been implemented in software tools such as DAGchainer (17), ColinearScan (18), MCScan (19), SyMAP (20), FISH (21) and CYNTENATOR (22). In addition to algorithmic differences, synteny and collinearity detection tools often differ in application ranges, inputs, presentation of results and/or computational costs.

Although pair-wise collinear relationships among chromosomal regions have been widely studied, the multi-alignment (alignment of three or more regions) of collinear chromosomal regions (referred to as collinear blocks) is more important as it can reveal ancient WGD events (19,23) and complex chromosomal duplication/rearrangement relationships (24). Collinear blocks are comprised of anchor genes which are located at collinear positions and non-anchor genes which are assumed to have experienced gene gains, losses or transposition. Further, anchor genes are more likely to be homologs (25) and tend to be under stronger purifying selection than non-anchor genes (26). Patterns of synteny and collinearity can provide insight into the evolutionary history of a genome, and inform on potentially useful downstream analyses. However, although graphic interfaces for visualizing synteny and collinearity may be incorporated, many available software packages for synteny and collinearity detection do not directly provide downstream analysis tools. Further, genes may be duplicated by mechanisms other than whole-genome duplication, such as tandem, proximal and/or dispersed duplications, each of which may make different contributions to evolution (11,27). In addition, analysis of gene family evolution may require that it be placed in the context of genome evolution. To analyze the evolution of a genome, it may be helpful to correlate gene family analysis with different duplication modes for a more integrated view. To our knowledge, only the MicroSyn package (28) provides analysis of collinearity within gene families, but it cannot superimpose such analysis on a context of whole-genome collinearity.

MCScan is able to identify collinear blocks in genomes or subgenomes and then conduct multi-alignments of collinear blocks using collinear genes as anchors (19,23). MCScan is also customizable for genomes of different sizes and with different average intergenic distances. Using MCScan, a Plant Genome Duplication Database (PGDD) has been constructed and is publicly available at http://chibba.pgml.uga.edu/duplication/. The MCScan software package and PGDD database have been applied to a variety of research areas such as genome duplication and evolution (11,29–36), annotation of newly sequenced genomes (37) and the evolution of gene families (38–48).

Building on the MCScan algorithm, here we describe a software package named *MCScanX* for synteny and collinearity detection, visualization and diverse downstream analyses. Compared with MCScan, the usage of *MCScanX* has been greatly simplified. To more clearly show how frequently chromosomal regions are duplicated, multi-alignments of collinear blocks against reference chromosomes can be viewed through a web browser with various highlighted features (e.g. tandem arrays, coverage statistics). The overall pattern of synteny and collinearity between or among genomes can be visualized by up to four types of plots. Compared with existing synteny and collinearity detection tools, a distinct feature of *MCScanX* is that diverse tools for evolutionary analyses of synteny and collinearity are incorporated, aiding efforts to construct gene families using collinearity information, infer gene duplication modes and enrichments, characterize collinear genes with nucleotide substitution rates, detect collinear tandem arrays, perform statistical analyses of duplication depths and collinear orthologs, and analyze collinearity within gene families. *MCScanX* enables rapid and convenient conversion of synteny and collinearity information into evolutionary insights.

MATERIALS AND METHODS

Gene set and homology search

Whole-genome protein sequences and gene positions for *Arabidopsis thaliana*, *Populus trichocarpa*, *Vitis vinifera*, *Glycine max*, *Oryza sativa* and *Brachypodium distachyon* were retrieved from Phytozome v7.0 (http://www.phytozome.net/). Whole-genome protein sequences and gene positions for *Sorghum bicolor* and *Zea mays* were retrieved from EnsemblPlants (http://plants.ensembl.org/index.html) and MaizeSequence Release 5b.60 (http://www.maizesequence.org/index.html) respectively. If a gene had more than one transcript, only the first transcript in the annotation was used. To search for homology, the protein-coding genes from each genome was compared against itself and other genomes using BLASTP (49). For a protein sequence, the best five non-self hits in each target genome that met an *E*-value threshold of 10^{-5} were reported.

MCScanX algorithm

The *MCScanX* algorithm is a modified version of MCScan (19). Whole-genome BLASTP results are used to compute collinear blocks for all possible pairs of chromosomes and scaffolds. First, BLASTP matches are sorted according to gene positions. To avoid high numbers of local collinear gene pairs due to tandem arrays, if consecutive BLASTP matches have a common gene and its paired genes are separated by fewer than five genes, these matches are collapsed using a representative pair with the smallest BLASTP *E*-value. Then, dynamic programming is employed to find the highest scoring paths (i.e. chains of collinear gene pairs) using the following scoring schema, assuming that two gene pairs, u and v, are on the path where u precedes v,

$$ChainScore(v) = MatchScore(v) + \max\{ChainScore(u) + GapPenalty \cdot NumberofGaps(u, v), 0\}$$

where by default $MatchScore(v) = 50$ for one gene pair, $GapPenalty = -1$, and $NumberofGaps\ (u,v)$, the maximum number of intervening genes between u and v, should be fewer than 25. Non-overlapping chains with scores over 250 (i.e. involving at least 5 collinear gene pairs) are reported. In a pair of collinear blocks, there are two distinct genomic locations with aligned collinear genes as anchors.

The expected number of occurrences (E-value) of a pair of collinear blocks is estimated using the formula introduced by Wang *et al.* (18),

$$E = 2P_N^m \prod_{i=1}^{m-1} \left(\frac{l_{1i}}{L_1} \cdot \frac{l_{2i}}{L_2}\right)$$

where N is the number of matching gene pairs between the two chromosomal regions defined by the pair of collinear blocks, m is the number of anchors in the pair of collinear blocks, L_1 and L_2 are respective lengths of the two chromosomal regions, and l_{1i} and l_{2i} are distances (in terms of nucleotide numbers) between two adjacent anchors in the pair of collinear blocks. The default E-value cutoff of *MCScanX* is 10^{-5}.

Multiple chromosomal regions threaded by consecutive ancestral loci are progressively aligned against reference chromosomes, where each genome being tested is used as a reference successively, according to the following procedure: (i) any reference chromosome is scanned from start to end, and empty tracks are placed alongside the reference chromosome to hold potential aligned collinear blocks; (ii) collinear blocks are progressively aligned against reference chromosomes pinpointed by anchors and assigned to the nearest empty tracks (once a track region is filled, it cannot be assigned collinear blocks again). In aligned collinear blocks, only symbols of anchor genes are shown while un-matched positions (gaps) between anchors (regardless of numbers of intervening genes) are denoted by '||'; (iii) at each locus of reference chromosomes, the number of tracks occupied by collinear blocks is recorded to reflect the duplication depth.

Classification of duplicate gene origins

Genes within a single genome can be classified as singletons, dispersed duplicates, proximal duplicates, tandem duplicates and segmental/WGD duplicates depending on their copy number and genomic distribution. The following procedure is used to assign gene classes: (i) All genes are initially classified as 'singletons' and assigned gene ranks according to their order of appearance along chromosomes; (ii) BLASTP results are evaluated and the genes with BLASTP hits to other genes are re-labeled as 'dispersed duplicates'; (iii) In any BLASTP hit, the two genes are re-labeled as 'proximal duplicates' if they have a difference of gene rank<20 (configurable); (iv) In any BLASTP hit, the two genes are re-labeled as 'tandem duplicates' if they have a difference of gene rank = 1; (v) *MCScanX* is executed. The anchor genes in collinear blocks are re-labeled as 'WGD/segmental'. So, if a gene appears in multiple BLASTP hits, it will be assigned a unique class according to the order of priority: WGD/segmental>tandem>proximal>dispersed.

Detection of orthologous gene pairs using OrthoMCL

Whole-genome protein sequences from *Arabidopsis*, *Populus*, *Vitis*, *Glycine*, *Oryza, Brachypodium*, *Sorghum* and *Zea* were merged and searched against themselves for homology using BLASTP with an E-value cutoff of 10^{-5}. Default parameters of OrthoMCL (50) were used. The combination of OrthoMCL intermediate files 'orthologs.txt' and 'coorthologs.txt' (generated by *orthomclDumpPairsFiles*) was used as the whole set of ortholog pairs.

Enrichment analysis

Enrichment analysis is performed using Fisher's exact test. The P-value was calculated for the null hypothesis that there is no association between the members of a gene family and a particular gene duplication mode and is corrected with the total number of duplication modes for multiple comparisons (i.e. Bonferroni correction). The P-value cutoff of 0.05 is used to suggest putative enrichment of certain gene duplication modes.

Computing *Ka* and *Ks*

Non-synonymous (*Ka*) and synonymous (*Ks*) substitution rates are estimated by Nei-Gojobori statistics (51), available through the 'Bio::Align::DNAStatistics' module of the BioPerl package (http://www.bioperl.org/wiki/Module:Bio::Align::DNAStatistics). Note that the 'Bio::Align::DNAStatistics' module may generate invalid *Ka* or *Ks* (i.e. non-digital output) for some homologous gene pairs due to mis-alignments.

Gene family examples

Lists of published *Arabidopsis* gene families were obtained from TAIR (http://www.arabidopsis.org/browse/genefamily/index.jsp). Only families with more than nine genes were considered in order to have enough statistical power to detect enrichment of duplication modes. *Arabidopsis* disease resistance gene homologs were downloaded from the NIBLRRS Project website (http://niblrrs.ucdavis.edu).

Execution of the *MCScanX* package

MCScanX is freely available at http://chibba.pgml.uga.edu/mcscan2. All programs in the *MCScanX* package should be executed using command line arguments on Mac OS or Linux systems. On Mac OS, Xcode (http://developer.apple.com/xcode/) should be installed prior to the installation of *MCScanX* package. On Linux systems, the Java SE Development Kit (JDK) and 'libpng' should be installed before the installation of *MCScanX* package. To list available command line options, the user can simply type the name of a program without any options.

RESULTS

Structure of the *MCScanX* package

The *MCScanX* package consists of two main components: (i) three core programs that implement an adjusted MCScan algorithm to generate pairwise and multiple alignments of collinear blocks and (ii) 12 downstream analysis programs for displaying and analyzing identified synteny and collinearity output by the core programs. The structure of the *MCScanX* package is shown in Figure 1. Compared with the previous version (0.8) of MCScan, there are numerous improvements in *MCScanX*. First, preprocessing of BLASTP input has been pipelined into the execution of core programs. Next, in MCScan, each gene was assigned a family ID to identify tandem genes, where the family ID has to be pre-computed using the Markov Clustering Algorithm (MCL) software (52). In *MCScanX*, tandem genes are assessed by gene rank according to chromosomal positions and thus, execution of MCL is no longer required. The aforementioned two improvements have made the installation and execution of *MCScanX* easier and more efficient. Furthermore, multi-alignments of collinear blocks, which are output as HTML files in *MCScanX*, can be easily and clearly viewed. In addition, numerous visualization and downstream analysis tools are incorporated into the *MCScanX* package, greatly enhancing the biological applications of the MCScan algorithm. In the following, we describe in detail each program in the *MCScanX* package.

The first core program, named *MCScanX*, can generate both pair-wise and multiple alignments of collinear blocks, similar to the previous MCScan version (0.8). However, *MCScanX* takes only a simplified GFF format file and a BLASTP tabular file as inputs. The simplified GFF file should contain the gene locations (which include chromosome, gene symbol, start and end) for the genomes to be compared. The BLASTP input file is one BLASTP output or combined multiple BLASTP outputs in tabular format (option '-m8' in BLAST and '−outfmt 6' in BLAST+) for all protein sequences in the species of interest. Note that when *MCScanX* is applied to multiple species, it may be useful to guard against over-enrichment of gene pairs from closely related species and we recommend that the BLASTP input file include the combined BLASTP outputs of pairwise genome comparisons and self-genome comparisons with a cutoff of best hits instead of a single BLASTP output of pooled protein sequences from different species. Alternatively, the BLASTP input can be replaced by a tab-delimited file containing pair-wise homologous relationships detected by third party software. In this case, the user needs to implement *MCScanX_h* (the second core program). In addition, *MCScanX_h* can generate statistics on numbers of collinear homolog pairs and their percentages (relative to the numbers of input homolog pairs).

We also adopted an adjusted MCScan algorithm. Matches among genes are first sorted according to chromosomal positions for all possible pairs of chromosomes and scaffolds, and in both transcriptional directions. Adjacent collinear genes are chained using dynamic programming (see 'Materials and Methods' section), outputting pairwise collinear blocks and tandem gene pairs to '.collinearity' and '.tandem' files respectively. Note that during the chaining of collinear genes, distances between genes are calculated in terms of differences in gene ranks. Use of differences in gene ranks provides relative gene distances, which can mitigate the effects of different gene densities (per unit physical DNA) among species on collinearity detection. Next, multiple chromosomal regions threaded by consecutive anchor loci are progressively aligned against 'reference' genomes. Because there could be many intervening/non-anchor genes between consecutive anchor genes, especially for divergent genomes, the alignment of non-anchor genes is highly flexible and could clutter the view of results. Thus, in *MCScanX*, the alignment among non-anchor genes is discarded in the output and non-anchor genes (mismatches) are simply denoted by '||' in the multi-alignment of gene orders. As a result, the layout of multiple alignments is less affected by alignment parameters and anchor genes and duplication depths can be easily discerned in the resulting multiple alignments.

The results of *MCScanX* multiple alignments are presented in HTML format with variously colored features that can be displayed using a web browser. An example is shown in Figure 2. In a reference chromosome, both anchor and non-anchor genes are shown, while in aligned collinear blocks only anchor genes are shown. Along the reference chromosome, duplication depth (i.e. number of aligned collinear blocks) is shown at each locus to indicate how frequently chromosomal regions are duplicated, and tandem genes are highlighted in red. In principle, all aligned collinear blocks can be also references. Note that in certain cases, in a specific alignment (e.g. A–B–C), an anchor locus is lost in the reference chromosome (A) and in turn cannot be shown in aligned collinear blocks (B and C) due to the non-reciprocity of the employed algorithm. To study differential gene loss, the user is suggested to analyze the results using the gene or genome of interest as the reference (i.e. the alignments B–A–C and C–A–B can show that the anchor locus exists between B and C but is lost in A) to ensure that complete chromosomal neighborhoods and matching segments are observed.

The third core program, named *duplicate gene classifier*, can classify the duplicate genes of a single species into WGD/segmental, tandem, proximal and dispersed duplicates. WGD/segmental duplicates are inferred by the anchor genes in collinear blocks. Tandem duplicates are defined as paralogs that are adjacent to each other on chromosomes, which are suggested to arise from illegitimate chromosomal recombination (27). Proximal duplicates are paralogs near each other, but interrupted by several other genes (e.g. separated by fewer than 20 genes, configurable). Proximal duplicates are inferred to result from localized transposon activities (53), or ancient tandem arrays interrupted by more recent gene insertions. Dispersed duplicates are paralogs that are neither near each other on chromosomes, nor do they show conserved synteny (54). Distant single gene translocations mediated

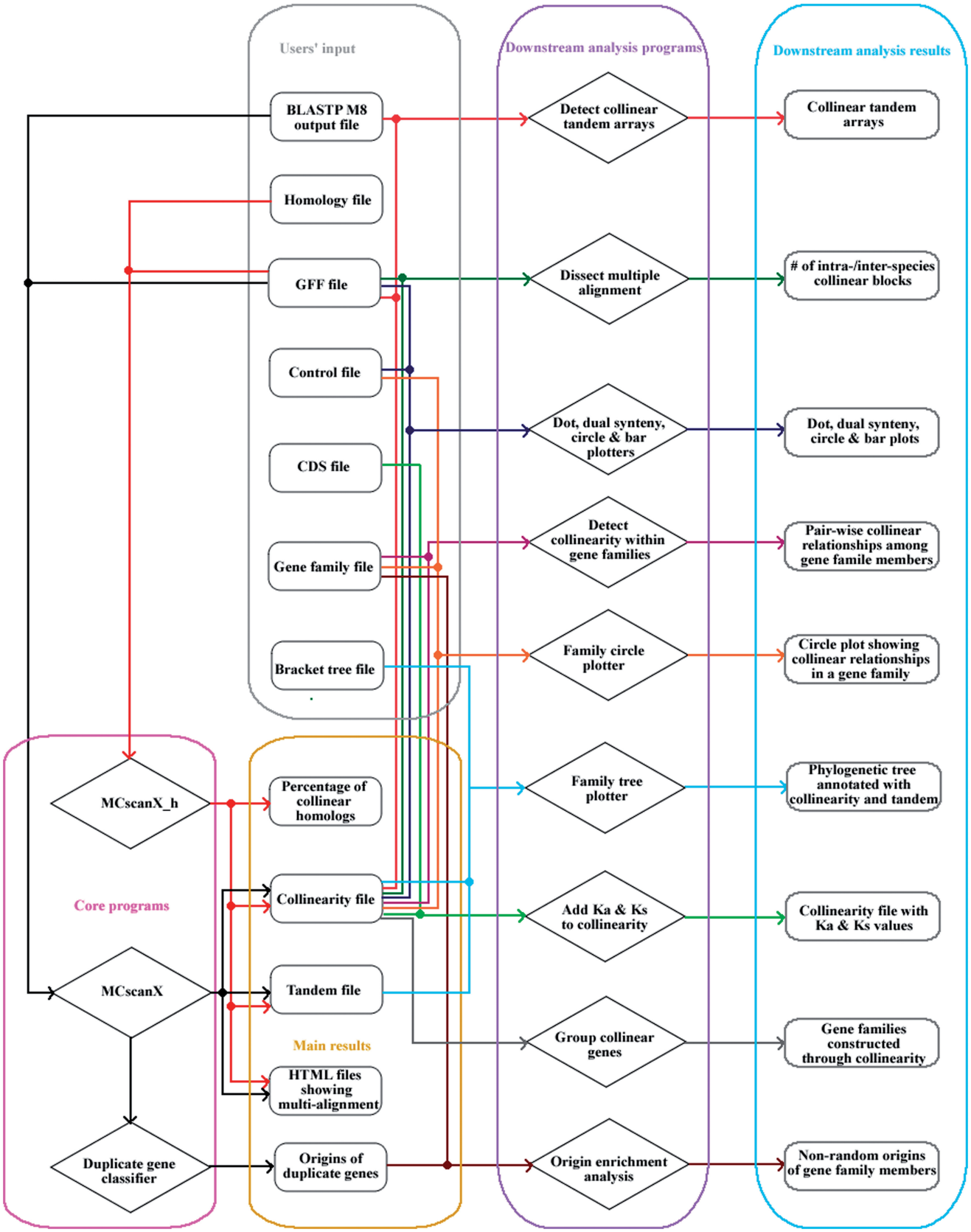

Figure 1. The structure of the *MCScanX* package illustrating major components and their dependencies.

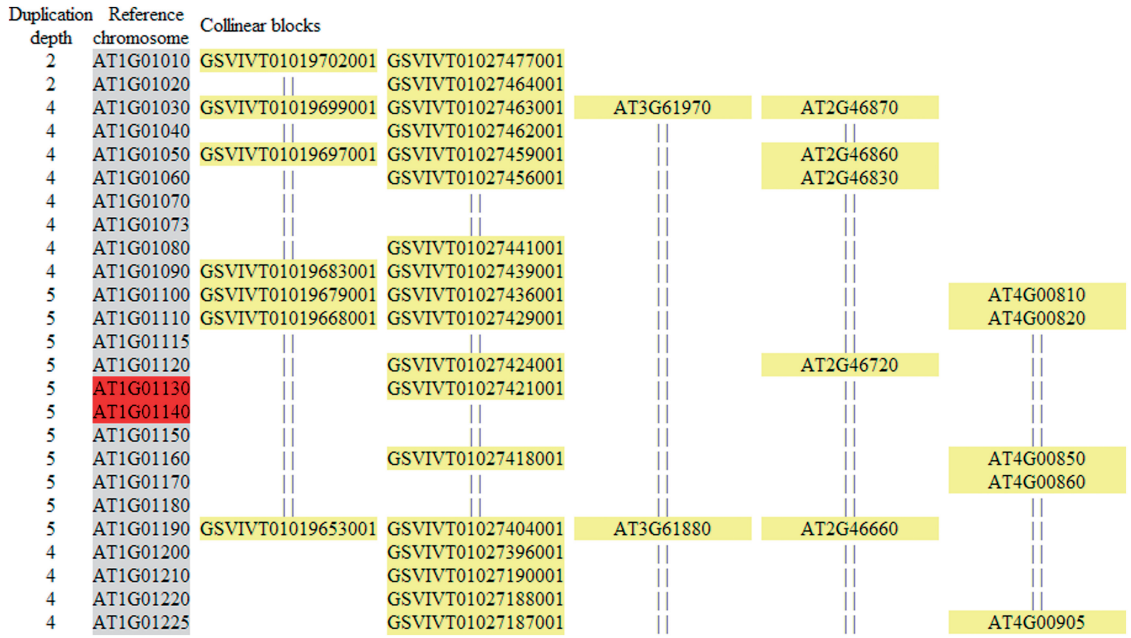

Figure 2. Sample HTML output displaying multiple alignments of collinear blocks by *MCScanX*. The first and second columns show duplication depth and gene symbol at each locus of reference chromosomes, where tandems are marked in red. The remaining columns show aligned collinear blocks, where only the symbols of anchor genes are shown.

by transposons may explain the wide spread of dispersed duplicates (27), often via pack-MULEs (55), helitrons (56), or CACTA elements (37) in plant genomes, or through 'retropositions' (57). Inferences about the mechanism(s) responsible for duplication of genes may reveal unusual evolutionary characteristics for particular lineages. *Duplicate gene classifier,* incorporating the *MCScanX* procedure, takes in the same input files as *MCScanX*, and returns statistics of duplicate gene origins and a file showing the likely origin of each gene.

Once the outputs of the core programs are generated, various visualization and downstream analysis tools can be applied. To display synteny and collinearity, four types of plots can be generated: dual synteny plot (Figure 3A), circle plot (Figure 3B), dot plot (Figure 3C) and bar plot (Figure 3D) using the Java programs: *dual synteny plotter, circle plotter, dot plotter* and *bar plotter*, respectively. The '.collinearity' file generated by *MCScanX* can be annotated with non-synonymous (*Ka*) and synonymous (*Ks*) substitution rates using the Perl program *add ka and ks to collinearity.pl*. Gene families constructed based on collinear relationships (instead of BLAST hits) can be generated based on the '.collinearity' file using the Perl program *group collinear genes*. It may be interesting to see how frequently chromosomal regions are duplicated within or across species for understanding species-specific or shared evolutionary events, and the program *dissect multiple alignment* can compute the number of intra- and inter-species collinear blocks at each locus of reference genomes and show statistics on gene numbers at different duplication depths. To avoid high numbers of local collinear gene pairs generated by *MCScanX* due to tandem arrays, tandem matches are collapsed using a representative pair with the smallest BLASTP *E*-value during *MCScanX* execution. However, a tandem array at an ancestral locus may imply positional gene family expansion (58). Thus, a tool named *detect collinear tandem arrays* is provided for detection of collinear tandem arrays.

The *MCScanX* package provides a variety of tools for analyzing gene family evolution based on the synteny and collinearity identified by *MCScanX*. *Origin enrichment analysis* can detect potential enrichment of duplicate gene origins for gene families, based on the classification of whole-genome duplicate genes (the output of *duplicate gene classifier*). *Detect collinearity within gene families* outputs all collinear gene pairs among gene family members. *Family circle plotter* can detect all collinear gene pairs within a gene family and plot them using a genomic circle *Family tree plotter*, with a Newick-format tree (direct results from most phylogenetic software) and '.collinearity' and '.tandem' files (generated by *MCScanX*) as inputs, can graphically annotate a phylogenetic tree with collinear and tandem relationships.

Application examples

Estimation of the number of WGD events. MCScan version 0.8 was implemented to estimate the number of WGD events of *Arabidopsis, Carica, Populus* and *Vitis*, through analysis of the duplication depths of their collinear blocks using *Vitis* as the reference genome (19,23). To facilitate this analysis using the output of *MCScanX*, the tool *dissect multiple alignment* is provided. When the user applies the *MCScanX* package, the BLASTP and GFF inputs should be restricted to a single genome for self-genome comparison or between two genomes for cross-genome comparison. Alternatively, a BLASTP of self-genome comparison and cross-genome comparison

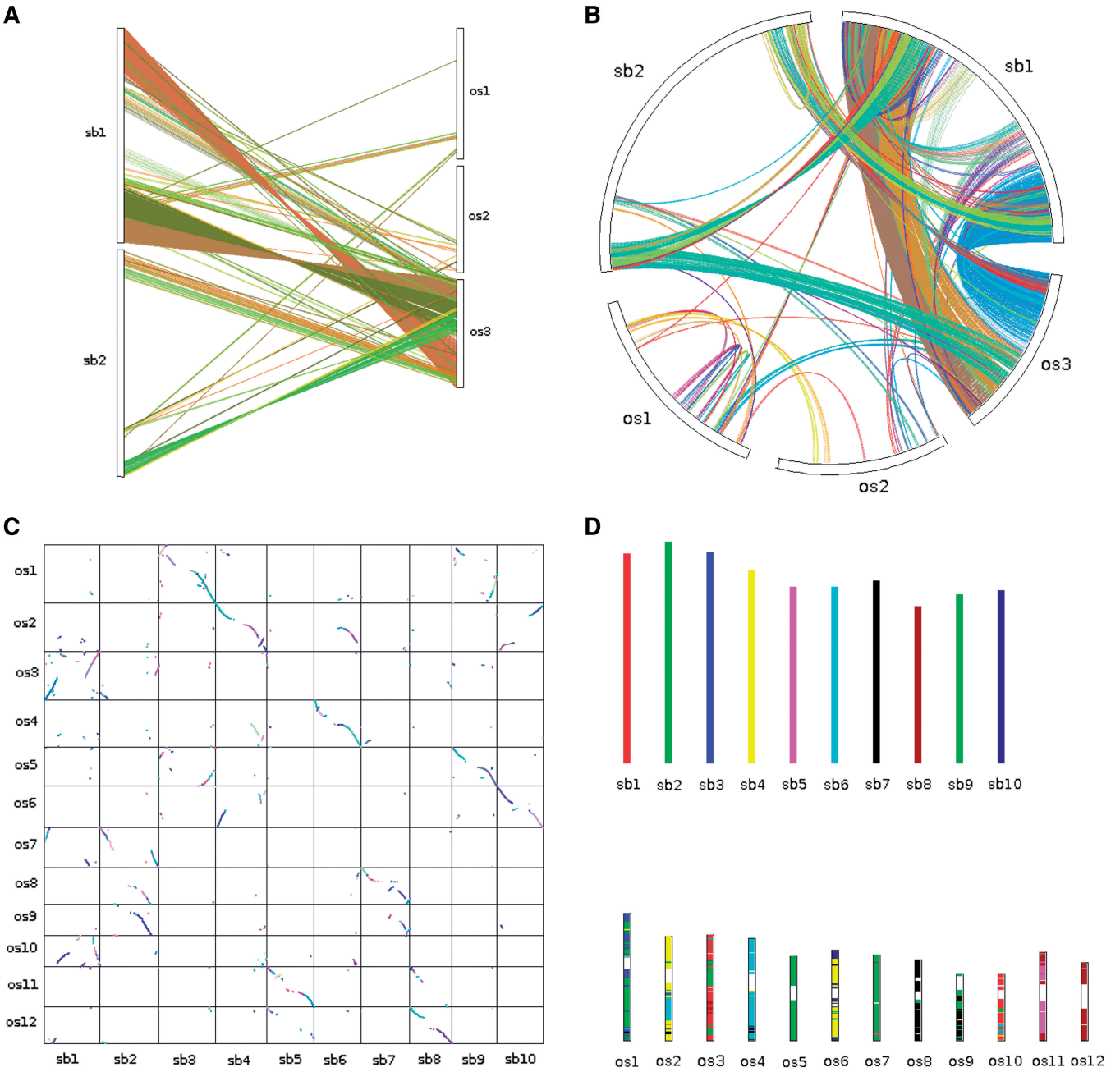

Figure 3. Different types of plots showing patterns of synteny and collinearity: (**A**) dual synteny plot, (**B**) circle plot, (**C**) dot plot and (**D**) bar plot, generated by '*dual synteny plotter, circle plotter, dot plotter and bar plotter*', respectively. Chromosomes are labeled in the format 'species abbreviation' + 'chromosome ID'. os, *Oryza sativa*; sb, *Sorghum bicolor*.

may be merged for both comparisons. However, self-genome comparison may not be as sensitive as cross-genome comparison due to the differential loss of functionally redundant genes, sometimes in a complementary fashion (19). Although the determination of an exact number of WGD events may be heuristic, the output of '*dissect multiple alignment*' can give a reasonable estimate. Note that a duplication depth x indicates that there are x and x+1 aligned collinear blocks in the target genome using cross-genome and self-genome comparisons respectively. For example, '*dissect multiple alignment*' was applied to both self-genome and cross-genome comparisons between *Arabidopsis* and *Vitis*. Using *Arabidopsis* and *Vitis* as references, the maximum duplication depths of *Arabidopsis* collinear blocks are 7 (self-genome comparison, so the maximum number of aligned *Arabidopsis* collinear blocks is 8) and 11 (cross-genome comparison, so the maximum number of aligned *Arabidopsis* collinear blocks is 11), respectively, suggesting that the lineage experienced at least three WGD events to achieve this duplication depth, i.e. a triplication WGD event $\gamma \times$ two duplication WGD events α and β (6,19,23). By applying

dissect multiple alignment to self-genome comparison of *Vitis*, the maximum duplication depth of *Vitis* collinear blocks is 4. However, the gene numbers at levels 3 and 4 (297 and 6, respectively) are much smaller than at level 2 (6993). A whole-genome triplication (WGT) plus small scale chromosomal duplications is the simplest explanation for this duplication pattern (19,23). Note that analysis of duplication depths of collinear blocks can generate good estimates on relatively recent WGD events. Very ancient WGD events often do not result in discernable collinear blocks in extant species due to extensive chromosome rearrangement, loss or gain of chromosomal segments, loss or transposition of duplicate genes, horizontal gene transfers, etc. A recent study, through analyzing the phylogenetic trees of cross-species gene families, reported two ancestral WGD events for seed plants and angiosperms respectively (59).

Detection of collinear orthologs. Detection of collinear orthologs is important for understanding gene evolution. The comparison between collinear orthologs and all orthologs can reveal how gene orders are conserved (or inversely, how frequently chromosomes are rearranged) between species. Limited only by the state of a genome's annotation and the assumption that sufficient sequence similarity is present for detection, a complete set of orthologs for a set of species can be generated by third-party software such as OrthoMCL (50). We implemented OrthoMCL to find ortholog pairs among *Arabidopsis*, *Populus*, *Vitis*, *Glycine*, *Rice, Brachypodium*, *Sorghum* and *Zea*. The ortholog pairs identified by OrthoMCL were regarded as the whole set of orthologs, and were then used as the input of *MCScanX_h*. Besides standard *MCScanX* output, *MCScanX_h* generated statistics on the numbers of collinear ortholog pairs and all ortholog pairs, and percentages of collinear ortholog pairs between any two of the selected angiosperm genomes (Table 1). As expected, gene order is better conserved within monocots and within eudicots than between monocots and eudicots. Within eudicots, *Vitis* shows the highest level of collinearity with the other 3 species, suggesting that *Vitis* most closely resemble the gene order of the eudicot ancestral genome, due in part to the lack of recent WGDs (60).

Differences in duplicate gene origins among angiosperms. Using self-genome BLASTP outputs and the tool *duplicate gene classifier*, we classified the origins of duplicate genes for *Arabidopsis*, *Populus*, *Vitis*, *Glycine*, *Oryza, Brachypodium*, *Sorghum* and *Zea* respectively. The results are shown in Table 2. The collinear blocks in the self-genome comparisons result from segmental or whole-genome duplications. Most collinear blocks within these flowering plant genomes were derived from WGDs because of their high coverage throughout the genome as well as supporting *Ks* evidence (19).

WGDs have had different impacts on the gene repertoires of the investigated taxa. Strikingly, ~76.0% of *Glycine* genes were duplicated and retained from WGD events, versus only 14.5% of *Oryza* genes. The proportions of genes involved in WGD events may reflect the relative timing of the most recent WGD event, as well as the level of gene retention following the WGD. For example, *Vitis*, with only 15.0% of genes created by WGD (actually WGT), was inferred to have undergone the γ WGT event, which likely predated the divergence of most eudicots >100 million years ago (19,23). Other eudicot lineages have experienced lineage-specific WGDs in addition to the shared γ event. Twenty-seven percent of *Arabidopsis* appear to have been created through WGD, having experienced α and β WGD events since its divergence from other members of the Brassicales clade (6,23). *Populus*, with 51.6% of genes created by WGD, was inferred to have undergone an additional WGD event in the Salicoid lineage (23). *Glycine*, with the highest proportion of WGD genes, was reported to have experienced two additional WGD events, with the most recent occurring 13 million years ago (61). A total of 29.2% of *Zea* genes were created through WGD, which experienced a lineage-specific WGD after its divergence from *Sorghum*

Table 1. Numbers of collinear ortholog pairs and total ortholog pairs and percentage of collinear ortholog pairs in selected angiosperm genomes

Species	No. of collinear ortholog pairs, No. of total ortholog pairs and percentage of collinear ortholog pairs						
	Pt	Gm	Vv	Os	Bd	Sb	Zm
At	14 278, 46 944, 30.4%	17 498, 58 038, 30.1%	7378, 24 086, 30.6%	319, 24 992, 1.3%	202, 22 719, 0.9%	350, 24 120, 1.5%	142, 24 689, 0.6%
Pt	–	34 545, 92 901, 37.2%	15 734, 38 727, 40.6%	2121, 37 575, 5.6%	1632, 32 790, 5.0%	1523, 36 059, 4.2%	687, 35 596, 1.9%
Gm	–	–	18 310, 47 652, 38.4%	1437, 46 916, 3.1%	1308, 43 130, 3.0%	1263, 46 631, 2.7%	501, 47 326, 1.1%
Vv	–	–	–	1315, 19 678, 6.7%	981, 18 080, 5.4%	1194, 19 137, 6.2%	293, 19 501, 1.5%
Os	–	–	–	–	15 492, 34 413, 45.0%	15 664, 39 695, 39.5%	14 112, 35 206, 40.1%
Bd	–	–	–	–	–	14 070, 32 701, 43.0%	13 111, 30 841, 42.5%
Sb	–	–	–	–	–	–	18084, 36826, 49.1%

At, *Arabidopsis thaliana*; Pt, *Populus trichocarpa*; Gm, *Glycine max*; Vv, *Vitis vinifera*; Os, *Oryza sativa*; Bd, *Brachypodium distachyon*; Sb, *Sorghum bicolor*; Zm, *Zea mays*.

Table 2. Numbers of genes from different origins as classified by *duplicate gene classifier* in eight angiosperm genomes

Species	No. of genes	No. of genes from different origins (percentage)				
		Singletons	WGD	Tandem	Proximal	Dispersed
Arabidopsis	27 105	5272 (19.5)	7321 (27.0)	769 (2.8)	892 (3.3)	12 851 (47.4)
Populus	40 650	5014 (12.3)	20 989 (51.6)	713 (1.8)	999 (2.5)	12 935 (31.8)
Glycine	46 360	1459 (3.1)	35 233 (76.0)	582 (1.3)	670 (1.4)	8416 (18.2)
Vitis	23 647	6275 (26.5)	3539 (15.0)	688 (2.9)	1590 (6.7)	11 555 (48.9)
Oryza	40 634	12 720 (31.3)	5896 (14.5)	960 (2.4)	2184 (5.4)	18 874 (46.4)
Brachypodium	25 524	4842 (19.0)	4575 (17.9)	697 (2.7)	827 (3.2)	14 583 (57.1)
Sorghum	34 564	5839 (16.9)	5260 (15.2)	895 (2.6)	1283 (3.7)	21 287 (61.6)
Zea	39 365	8212 (20.9)	11 506 (29.2)	774 (2.0)	1175 (3.0)	17 698 (45.0)

(15.2% genes created by WGD) (62,63). Although tandem genes are volatile after gene duplication, those retained may indicate functional significance. We find that tandem genes account for about 1–3% of genes in each genome, smaller than ~10% reported by Rizzon *et al.* (64). This difference is due to the algorithm of *duplicate gene classifier*, which treats the tandem duplicates located at ancestral loci as WGD duplicates. Proximal duplicates account for larger proportions of genes in the genomes with fewer WGD duplicates, e.g. there are 5.4% of *Oryza* genes and 6.7% of *Vitis* genes created by proximal duplications, while in other genomes, the numbers of proximal duplicates are comparable to those of tandem duplicates.

Detection of collinear tandem arrays. In the *MCScanX* package, tandem arrays are defined as clusters of consecutive tandem duplicates. Via '*detect collinear tandem arrays*', tandem arrays are first determined according to successive gene ranks in all chromosomes. Collinear gene pairs are then searched against these tandem arrays. If any gene of a collinear pair is located within a tandem array, the gene is replaced by the tandem array and then reported. If a tandem array is located at an anchor locus of a collinear block, it is termed a collinear tandem array. Collinear tandem arrays can indicate positional gene family expansions (58), which could be important for forming large gene families, or adopted as an alternative path to increasing gene copy number in the genomes that experienced fewer WGD events. For example, we applied the tool '*detect collinear tandem arrays*' to a comparison of the *Arabidopsis* and *Vitis* genomes. A total of 1160 pairs of collinear tandem arrays were detected between *Arabidopsis* and *Vitis*, of which only 68 (5.9%) pairs have equal numbers of tandem duplicates in each species, while 54.3% of pairs have more tandem duplicates in *Vitis* than *Arabidopsis*. In conjunction with the finding above that *Vitis* has more proximal duplicates than other species, we suggest that tandem and proximal duplications contribute relatively more to the expansion of the *Vitis* genome than to other eudicots that experienced more WGDs in their evolutionary histories.

Analysis of gene family evolution. While *MCScanX* can detect synteny and collinearity using whole-genome homology and gene positional information, it is also of interest to analyze collinearity within a gene family, toward clarifying gene family evolution (65). We used the *Arabidopsis* MADS-box gene family as an example to illustrate the usefulness of *MCScanX* for analyzing the history of gene family expansion. Using the tool '*detect collinearity with gene families*', we detected 14 collinear gene pairs from the members of the MADS box gene family. The inferred collinear relationships of the MADS box gene family members can be displayed and placed within the context of whole-genome collinearity using a genomic circle generated by '*family circle plotter*' (Figure 4). Next, a phylogenetic tree was constructed for the MADS box gene family using PhyML package (66). The Newick tree was then used as the input of '*family tree plotter*'. A plot that showed the phylogenetic tree, collinear and tandem relationships for the MADS box gene family was generated (Figure 5). The overlay of positional

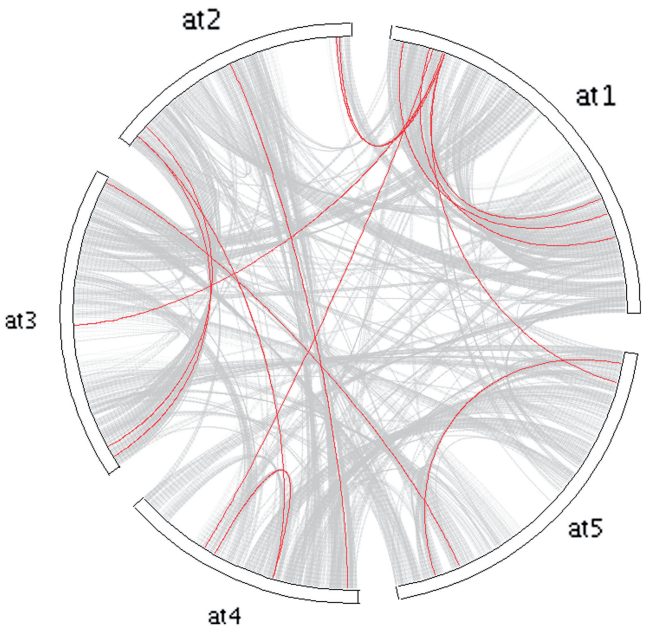

Figure 4. Circle plot showing collinearity in the MADS box gene family over the gray background of collinearity in *Arabidopsis* (the collinear blocks in *Arabidopsis*). The circle plot can be generated by '*family tree plotter*'. Chromosomes are labeled in the format 'species abbreviation' + 'chromosome ID'. at, *Arabidopsis thaliana*.

history over the gene clades reveals interesting characteristics of the MADS-box gene family. We note that the clade with many collinear relationships (WGD or segmentally duplicated) appears to be the MIKC^c-type (67). In contrast, the remaining clades of MADS-box genes appear to favor dispersed duplications (27,68).

The tool '*origin enrichment analysis*', which is able to detect potential enrichments of duplicate gene origins, was applied to 126 published *Arabidopsis* gene families of 10 or more genes, available at TAIR (http://www.arabidopsis.org/). We found that 46 (36.5%) gene families were enriched for at least one of the four types of origins at $\alpha = 0.05$. For example, disease resistance gene homologs and the cytochrome P450 gene family are enriched for dispersed and proximal duplicates, while the cytoplasmic ribosomal protein gene family and C2H2 zinc finger proteins are enriched for WGD duplicates, as previously noted (68).

Comparison with other synteny and collinearity tools

Existing tools for synteny and collinearity detection mainly include ADHoRe (12), TEAM (13), LineUp (14), MCMuSeC (15), OrthoCluster (16), DiagHunter (69), DAGChainer (17), ColinearScan (18), MCScan (19), SyMAP (20), FISH (21), Cyntenator (22), MicroSyn (28) and Cinteny (70), of which OrthoCluster, ADHoRe and SyMAP are currently upgraded to OrthoClusterDB (71), i-ADHoRe 3 (72) and SyMAP 3.4 (73), respectively. We summarized the functions of synteny and collinearity detection tools regarding five elements: graphic visualization, operation on multiple (>2) genomes, multi-alignments, evolutionary analyses of synteny and collinearity (e.g. estimating WGD events, gene-order conservation and duplicate gene origins, constructing collinear gene groups/families, etc.) and analyses of gene families. Functional comparison of different synteny and collinearity detection tools is shown in Table 3. If there were multiple versions for a tool, we used the latest one for comparison. Seven tools output synteny or collinearity information as plain texts, while the other tools provide graphic visualization options, though types and numbers of plots vary among different tools. As for the data scale, most tools published in the past 4 years can operate on multiple genomes. Five tools can perform multi-alignments of collinear blocks. MicroSyn is focused on collinearity analysis within gene families. i-ADHoRe 3 has provided several post-processing programs for dissecting multi-alignments of collinear blocks, in addition to detecting and visualizing synteny and collinearity. Among these synteny and collinearity detection tools, 11 tools cover no more than two functions, and OrthoclusterDB, MicroSyn and i-ADHoRe 3 cover three functions. *MCScanX*, with all five functions, can perform more biological analyses than any other synteny or collinearity detection tool.

MCScanX is unique in providing multiple programs for evolutionary analysis of synteny and collinearity, which are a necessary step towards biological discovery. Further, *MCScanX* has connected collinearity analyses between whole-genome and gene family scales. To our

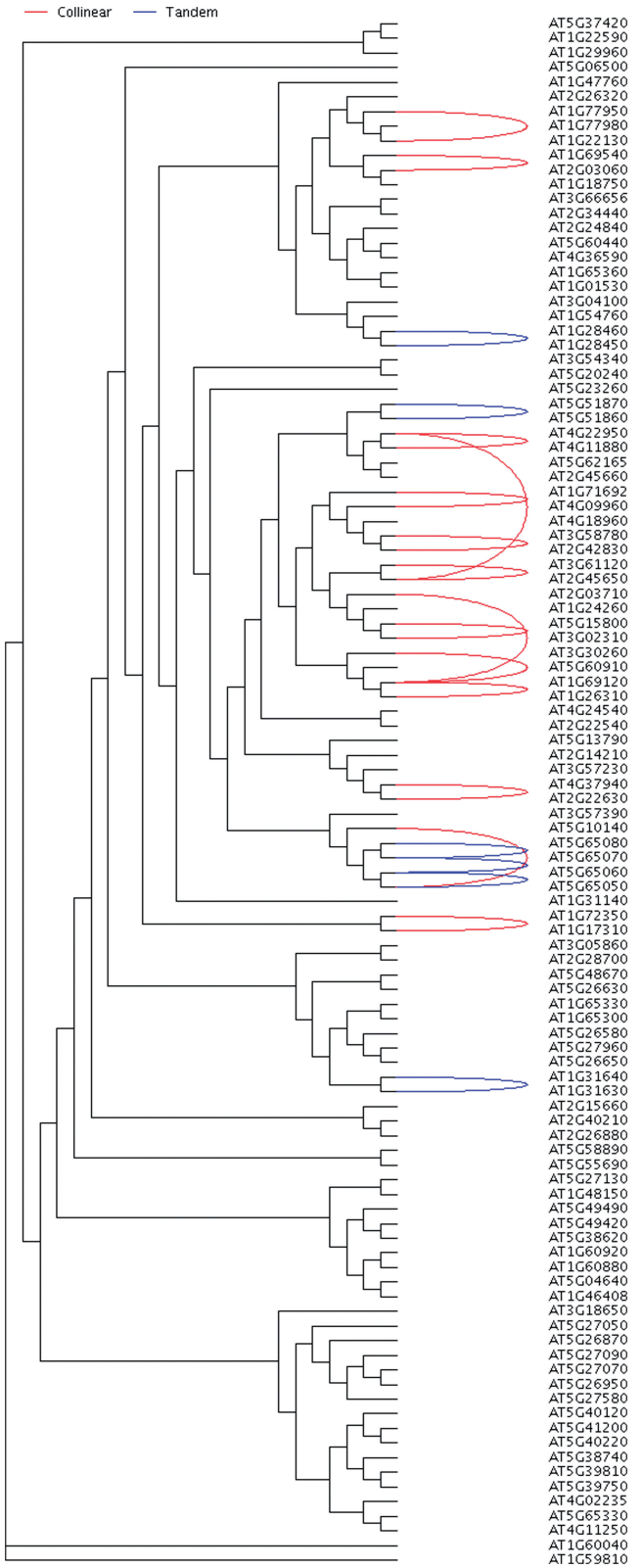

Figure 5. Phylogenetic tree of the MADS box gene family in *Arabidopsis* annotated with collinear and tandem relationships. Curves connecting pairs of gene names suggest either the collinear relationship (red) or tandem relationship (blue). This annotated tree is output from '*family tree plotter*'.

Table 3. Functional comparison of different synteny and collinearity detection tools ('+' and '−' represent 'yes' and 'no', respectively)

Tool	Year published	Graphic visualization	Multiple genomes	Multi-alignments	Evolutionary analyses of synteny and collinearity	Analyses of gene families
i-ADHoRe 3	2011	+	+	+	−	−
LineUp	2003	−	−	−	−	−
TEAM	2003	−	+	−	−	−
MCMuSeC	2009	−	+	−	−	−
OrthoClusterDB	2009	+	+	+	−	−
DiagHunter	2003	+	−	−	−	−
DAGChainer	2004	+	−	−	−	−
ColinearScan	2006	−	−	−	−	−
MCScan	2008	−	+	+	−	−
SyMAP 3.4	2011	+	+	−	−	−
FISH	2003	−	−	−	−	−
Cyntenator	2010	−	+	+	−	−
MicroSyn	2011	+	+	−	−	+
Cinteny	2007	+	+	−	−	−
MCScanX		+	+	+	+	+

knowledge, the following biological analyses implemented in *MCScanX* are not yet available in other synteny and collinearity detection tools: constructing gene families using collinearity information, inferring gene duplication modes and enrichments, detecting collinear tandem arrays, performing statistical analyses of duplication depths and collinear orthologs and annotating phylogenetic trees with collinearity and tandems.

For synteny and collinearity detection tools, effective identification of collinear gene pairs is the basis for collinear block construction and downstream analyses. It is informative to perform a quantitative evaluation of *MCScanX* on the identification of collinear gene pairs. Two widely implemented tools, MCScan and i-ADHoRe 3 were chosen as competitors. Since a benchmark for assessing synteny and collinearity tools has not been established (72), we compared their performances by applying them to the *Arabidopsis thaliana* genome. Note that a higher number of detected collinear gene pairs does not simply indicate better performance, as true and false positives must be simultaneously considered and well balanced (69). A total of 5794 collinear gene pairs (i.e. WGD duplicate gene pairs) in the *Arabidopsis* genome including 3822 α, 1451 β and 521 γ pairs profiled using an integrated phylogenomic approach in the study from Bowers *et al*. (6), were regarded as the whole set of collinear gene pairs. The performances of MCScan, *MCScanX* and i-ADHoRe 3 were evaluated by power (i.e. sensitivity), defined as the ratio between numbers of true positives and all collinear gene pairs; and precision, defined as the ratio between numbers of true positives and all positives (i.e. true positives + false positives). When MCScan and *MCScanX* were compared, the same parameters were used. Based on the default parameters of *MCScanX* (match size = 5, max gaps = 25), MCScan and *MCScanX* identified 4134 and 4225 collinear gene pairs, of which 3375 and 3407 were true positives, respectively. Power was 0.58 and 0.59, and precision was 0.82 and 0.81 for MCScan and *MCScanX*, respectively. The above statistics suggest that MCScan and *MCScanX* are generally comparable in detecting collinear gene pairs, while *MCScanX* has a slightly higher power and a slightly lower precision. Based on its default parameters, i-ADHoRe 3 identified 6233 non-overlapping collinear gene pairs, of which 3459 were true positives. Its power and precision was 0.60 and 0.55. However, direct comparison between *MCScanX* and i-ADHoRe 3 using their respective default parameters was not reasonable because i-ADHoRe 3 output many more positives. To this end, we executed MCScan and *MCScanX* using a more relaxed set of parameters (match size = 3, max gaps = 50), which output 5554 and 6110 positives, respectively. Based on the new parameters, power was 0.65 and 0.67, and precision was 0.68 and 0.64 for MCScan and *MCScanX,* respectively. The new statistics suggest that in terms of identification of collinear gene pairs, MCScan and *MCScanX* each perform better than i-ADHoRe 3 and remain comparable to one another, with MCScan having higher precision and *MCScanX* having higher power. The small difference between MCScan and *MCScanX* is because in order to make *MCScanX* more easily and efficiently implemented, pre-processing of BLASTP input was pipelined into the execution of the main programs and the dependency of MCL was dropped. In MCScan, cross-family BLASTP hits are removed based on MCL output, while in *MCScanX*, all non-self BLASTP hits are considered, leading to an enlarged pool of BLASTP hits. MCL may generate 5–20% incorrect families and its performance is affected by inflation value (a parameter of the MCL algorithm used to control the granularity/tightness of protein clusters) (52). So the cross-family BLASTP hits based on MCL gene families indeed contain some collinear gene pairs, though the proportion of collinear gene pairs is smaller in cross-family BLASTP hits than in within-family BLASTP hits. This results in marginally higher power and lower precision for *MCScanX* than MCScan, though their performances on identifying collinear gene pairs are very similar. Since MCScan was successfully applied to the distantly related apicomplexans (11), we believe that

MCScanX is also applicable over a wide range of organisms besides angiosperms.

DISCUSSION

Synteny and collinearity information is important for elucidating the evolutionary histories of both genomes and gene families. Although many synteny and collinearity tools are available, their output files are often difficult to read and downstream evolutionary analysis programs are rarely provided. For this reason, users often have to write additional programs or reformat the synteny and collinearity output files in order to use third-party evolutionary analysis tools. This incompleteness of functionality has reduced the usefulness of existing synteny and collinearity detection tools. A distinguishing feature of *MCScanX* is that diverse tools for evolutionary analyses of synteny and collinearity are incorporated, which enables rapid and convenient conversion of synteny and collinearity information into evolutionary insights. In addition, many biological analyses implemented in *MCScanX* are unique. *MCScanX* can be used to effectively analyze chromosome structural changes and evolution, annotate new genomes and reveal the history of gene family expansions.

In conclusion, *MCScanX* is a toolkit that implements an adjusted MCScan algorithm for detection of synteny and collinearity and incorporates 14 computer programs for visualizing and analyzing identified synteny and collinearity. The usefulness of the *MCScanX* toolkit has been demonstrated through a series of real data applications and comparison with other synteny and collinearity detection tools. *MCScanX* is freely available at http://chibba.pgml.uga.edu/mcscan2/.

FUNDING

National Science Foundation (NSF: DBI 0849896, MCB 0821096, MCB 1021718 to A.H.P.); National Institutes of Health (R01 AI068908to J.C.K.) in part; resources and technical expertise from the University of Georgia Georgia Advanced Computing Resource Center in part. Funding for open access charge: NSF (DBI 0849896).

Conflict of interest statement. None declared.

REFERENCES

1. Coghlan,A., Eichler,E.E., Oliver,S.G., Paterson,A.H. and Stein,L. (2005) Chromosome evolution in eukaryotes: a multi-kingdom perspective. *Trends Genet.*, **21**, 673–682.
2. Dietrich,F.S., Voegeli,S., Brachat,S., Lerch,A., Gates,K., Steiner,S., Mohr,C., Pohlmann,R., Luedi,P., Choi,S.D. *et al.* (2004) The *Ashbya gossypii* genome as a tool for mapping the ancient *Saccharomyces cerevisiae* genome. *Science*, **304**, 304–307.
3. Dujon,B., Sherman,D., Fischer,G., Durrens,P., Casaregola,S., Lafontaine,I., De Montigny,J., Marck,C., Neuveglise,C., Talla,E. *et al.* (2004) Genome evolution in yeasts. *Nature*, **430**, 35–44.
4. Nakatani,Y., Takeda,H., Kohara,Y. and Morishita,S. (2007) Reconstruction of the vertebrate ancestral genome reveals dynamic genome reorganization in early vertebrates. *Genome Res.*, **17**, 1254–1265.
5. Salse,J., Abrouk,M., Bolot,S., Guilhot,N., Courcelle,E., Faraut,T., Waugh,R., Close,T.J., Messing,J. and Feuillet,C. (2009) Reconstruction of monocotelydoneous proto-chromosomes reveals faster evolution in plants than in animals. *Proc. Natl Acad. Sci. USA*, **106**, 14908–14913.
6. Bowers,J.E., Chapman,B.A., Rong,J. and Paterson,A.H. (2003) Unravelling angiosperm genome evolution by phylogenetic analysis of chromosomal duplication events. *Nature*, **422**, 433–438.
7. Hufton,A.L., Groth,D., Vingron,M., Lehrach,H., Poustka,A.J. and Panopoulou,G. (2008) Early vertebrate whole genome duplications were predated by a period of intense genome rearrangement. *Genome Res.*, **18**, 1582–1591.
8. Otto,S.P. and Whitton,J. (2000) Polyploid incidence and evolution. *Annu. Rev. Genet.*, **34**, 401–437.
9. Ranz,J.M., Casals,F. and Ruiz,A. (2001) How malleable is the eukaryotic genome? Extreme rate of chromosomal rearrangement in the genus *Drosophila*. *Genome Res.*, **11**, 230–239.
10. Bourque,G., Pevzner,P.A. and Tesler,G. (2004) Reconstructing the genomic architecture of ancestral mammals: lessons from human, mouse, and rat genomes. *Genome Res.*, **14**, 507–516.
11. Debarry,J.D. and Kissinger,J.C. (2011) Jumbled genomes: missing Apicomplexan synteny. *Mol. Biol. Evol.*, **28**, 2855–2871.
12. Vandepoele,K., Saeys,Y., Simillion,C., Raes,J. and Van De Peer,Y. (2002) The automatic detection of homologous regions (ADHoRe) and its application to microcolinearity between *Arabidopsis* and rice. *Genome Res.*, **12**, 1792–1801.
13. Luc,N., Risler,J.L., Bergeron,A. and Raffinot,M. (2003) Gene teams: a new formalization of gene clusters for comparative genomics. *Comput. Biol. Chem.*, **27**, 59–67.
14. Hampson,S., McLysaght,A., Gaut,B. and Baldi,P. (2003) LineUp: statistical detection of chromosomal homology with application to plant comparative genomics. *Genome Res.*, **13**, 999–1010.
15. Ling,X., He,X. and Xin,D. (2009) Detecting gene clusters under evolutionary constraint in a large number of genomes. *Bioinformatics*, **25**, 571–577.
16. Vergara,I.A. and Chen,N. (2009) Using OrthoCluster for the detection of synteny blocks among multiple genomes. *Curr. Protoc. Bioinform.*, 27, 6.10.1–6.10.18.
17. Haas,B.J., Delcher,A.L., Wortman,J.R. and Salzberg,S.L. (2004) DAGchainer: a tool for mining segmental genome duplications and synteny. *Bioinformatics*, **20**, 3643–3646.
18. Wang,X., Shi,X., Li,Z., Zhu,Q., Kong,L., Tang,W., Ge,S. and Luo,J. (2006) Statistical inference of chromosomal homology based on gene colinearity and applications to *Arabidopsis* and rice. *BMC Bioinformatics*, **7**, 447.
19. 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.
20. Soderlund,C., Nelson,W., Shoemaker,A. and Paterson,A. (2006) SyMAP: a system for discovering and viewing syntenic regions of FPC maps. *Genome Res.*, **16**, 1159–1168.
21. Calabrese,P.P., Chakravarty,S. and Vision,T.J. (2003) Fast identification and statistical evaluation of segmental homologies in comparative maps. *Bioinformatics*, **19(Suppl. 1)**, i74–i80.
22. Rodelsperger,C. and Dieterich,C. (2010) CYNTENATOR: progressive gene order alignment of 17 vertebrate genomes. *PLoS One*, **5**, e8861.
23. Tang,H., Bowers,J.E., Wang,X., Ming,R., Alam,M. and Paterson,A.H. (2008) Synteny and collinearity in plant genomes. *Science*, **320**, 486–488.
24. Abrouk,M., Murat,F., Pont,C., Messing,J., Jackson,S., Faraut,T., Tannier,E., Plomion,C., Cooke,R., Feuillet,C. *et al.* (2010) Palaeogenomics of plants: synteny-based modelling of extinct ancestors. *Trends Plant Sci.*, **15**, 479–487.
25. Jun,J., Mandoiu,I.I. and Nelson,C.E. (2009) Identification of mammalian orthologs using local synteny. *BMC Genomics*, **10**, 630.
26. Casneuf,T., De Bodt,S., Raes,J., Maere,S. and Van de Peer,Y. (2006) Nonrandom divergence of gene expression following gene and genome duplications in the flowering plant *Arabidopsis thaliana*. *Genome Biol.*, **7**, R13.

27. Freeling,M. (2009) Bias in plant gene content following different sorts of duplication: tandem, whole-genome, segmental, or by transposition. *Annu. Rev. Plant Biol.*, **60**, 433–453.
28. Cai,B., Yang,X., Tuskan,G.A. and Cheng,Z.M. (2011) MicroSyn: a user friendly tool for detection of microsynteny in a gene family. *BMC Bioinformatics*, **12**, 79.
29. Wang,X., Tang,H. and Paterson,A.H. (2011) Seventy million years of concerted evolution of a homoeologous chromosome pair, in parallel, in major Poaceae lineages. *Plant Cell*, **23**, 27–37.
30. Wang,X., Tang,H., Bowers,J.E. and Paterson,A.H. (2009) Comparative inference of illegitimate recombination between rice and sorghum duplicated genes produced by polyploidization. *Genome Res.*, **19**, 1026–1032.
31. Lyons,E., Pedersen,B., Kane,J., Alam,M., Ming,R., Tang,H., Wang,X., Bowers,J., Paterson,A., Lisch,D. *et al.* (2008) Finding and comparing syntenic regions among *Arabidopsis* and the outgroups papaya, poplar, and grape: CoGe with rosids. *Plant Physiol.*, **148**, 1772–1781.
32. Charles,M., Tang,H.B., Belcram,H., Paterson,A., Gornicki,P. and Chalhoub,B. (2009) Sixty million years in evolution of soft grain trait in grasses: emergence of the softness locus in the common ancestor of *Pooideae* and *Ehrhartoideae*, after their divergence from *Panicoideae*. *Mol. Biol. Evol.*, **26**, 1651–1661.
33. Lin,L., Pierce,G.J., Bowers,J.E., Estill,J.C., Compton,R.O., Rainville,L.K., Kim,C., Lemke,C., Rong,J., Tang,H. *et al.* (2010) A draft physical map of a D-genome cotton species (*Gossypium raimondii*). *BMC Genomics*, **11**, 395.
34. Lin,L., Tang,H., Compton,R.O., Lemke,C., Rainville,L.K., Wang,X., Rong,J., Rana,M.K. and Paterson,A.H. (2011) Comparative analysis of *Gossypium* and *Vitis* genomes indicates genome duplication specific to the *Gossypium* lineage. *Genomics*, **97**, 313–320.
35. Tang,H., Bowers,J.E., Wang,X. and Paterson,A.H. (2010) Angiosperm genome comparisons reveal early polyploidy in the monocot lineage. *Proc. Natl Acad. Sci. USA*, **107**, 472–477.
36. Wang,Y., Wang,X., Tang,H., Tan,X., Ficklin,S.P., Feltus,F.A. and Paterson,A.H. (2011) Modes of gene duplication contribute differently to genetic novelty and redundancy, but show parallels across divergent angiosperms. *PLoS One*, **6**, e28150.
37. Paterson,A.H., Bowers,J.E., Bruggmann,R., Dubchak,I., Grimwood,J., Gundlach,H., Haberer,G., Hellsten,U., Mitros,T., Poliakov,A. *et al.* (2009) The *Sorghum bicolor* genome and the diversification of grasses. *Nature*, **457**, 551–556.
38. Causier,B., Castillo,R., Xue,Y.B., Schwarz-Sommer,Z. and Davies,B. (2010) Tracing the evolution of the floral homeotic B- and C-function genes through genome synteny. *Mol. Biol. Evol.*, **27**, 2651–2664.
39. Knoller,A.S., Blakeslee,J.J., Richards,E.L., Peer,W.A. and Murphy,A.S. (2010) Brachytic2/ZmABCB1 functions in IAA export from intercalary meristems. *J. Exp. Bot.*, **61**, 3689–3696.
40. Watanabe,M., Mochida,K., Kato,T., Tabata,S., Yoshimoto,N., Noji,M. and Saito,K. (2008) Comparative genomics and reverse genetics analysis reveal indispensable functions of the serine acetyltransferase gene family in *Arabidopsis*. *Plant Cell*, **20**, 2484–2496.
41. Okazaki,Y., Shimojima,M., Sawada,Y., Toyooka,K., Narisawa,T., Mochida,K., Tanaka,H., Matsuda,F., Hirai,A., Hirai,M.Y. *et al.* (2009) A chloroplastic UDP-glucose pyrophosphorylase from *Arabidopsis* is the committed enzyme for the first step of sulfolipid biosynthesis. *Plant Cell*, **21**, 892–909.
42. Hyun,T.K., Kim,J.S., Kwon,S.Y. and Kim,S.H. (2010) Comparative genomic analysis of mitogen activated protein kinase gene family in grapevine. *Genes Genom.*, **32**, 275–281.
43. Li,C. and Zhang,Y.M. (2011) Molecular evolution of glycinin and beta-conglycinin gene families in soybean (*Glycine max* L. Merr.). *Heredity*, **106**, 633–641.
44. Li,W., Liu,B., Yu,L., Feng,D., Wang,H. and Wang,J. (2009) Phylogenetic analysis, structural evolution and functional divergence of the 12-oxo-phytodienoate acid reductase gene family in plants. *BMC Evol. Biol.*, **9**, 90.
45. Kopriva,S., Mugford,S.G., Matthewman,C. and Koprivova,A. (2009) Plant sulfate assimilation genes: redundancy versus specialization. *Plant Cell Rep.*, **28**, 1769–1780.
46. Palmieri,F., Pierri,C.L., De Grassi,A., Nunes-Nesi,A. and Fernie,A.R. (2011) Evolution, structure and function of mitochondrial carriers: a review with new insights. *Plant J.*, **66**, 161–181.
47. Higgins,J.A., Bailey,P.C. and Laurie,D.A. (2010) Comparative genomics of flowering time pathways using *Brachypodium distachyon* as a model for the temperate grasses. *PLoS One*, **5**, e10065.
48. Wang,X., Gowik,U., Tang,H., Bowers,J.E., Westhoff,P. and Paterson,A.H. (2009) Comparative genomic analysis of C4 photosynthetic pathway evolution in grasses. *Genome Biol.*, **10**, R68.
49. Altschul,S.F., Gish,W., Miller,W., Myers,E.W. and Lipman,D.J. (1990) Basic local alignment search tool. *J. Mol. Biol.*, **215**, 403–410.
50. Li,L., Stoeckert,C.J. Jr and Roos,D.S. (2003) OrthoMCL: identification of ortholog groups for eukaryotic genomes. *Genome Res.*, **13**, 2178–2189.
51. Nei,M. and Gojobori,T. (1986) Simple methods for estimating the numbers of synonymous and nonsynonymous nucleotide substitutions. *Mol. Biol. Evol.*, **3**, 418–426.
52. Enright,A.J., Van Dongen,S. and Ouzounis,C.A. (2002) An efficient algorithm for large-scale detection of protein families. *Nucleic Acids Res.*, **30**, 1575–1584.
53. Zhao,X.P., Si,Y., Hanson,R.E., Crane,C.F., Price,H.J., Stelly,D.M., Wendel,J.F. and Paterson,A.H. (1998) Dispersed repetitive DNA has spread to new genomes since polyploid formation in cotton. *Genome Res.*, **8**, 479–492.
54. Ganko,E.W., Meyers,B.C. and Vision,T.J. (2007) Divergence in expression between duplicated genes in *Arabidopsis*. *Mol. Biol. Evol.*, **24**, 2298–2309.
55. Jiang,N., Bao,Z., Zhang,X., Eddy,S.R. and Wessler,S.R. (2004) Pack-MULE transposable elements mediate gene evolution in plants. *Nature*, **431**, 569–573.
56. Yang,L. and Bennetzen,J.L. (2009) Distribution, diversity, evolution, and survival of Helitrons in the maize genome. *Proc. Natl Acad. Sci. USA*, **106**, 19922–19927.
57. Wang,W., Zheng,H., Fan,C., Li,J., Shi,J., Cai,Z., Zhang,G., Liu,D., Zhang,J., Vang,S. *et al.* (2006) High rate of chimeric gene origination by retroposition in plant genomes. *Plant Cell*, **18**, 1791–1802.
58. Vergara,I.A. and Chen,N. (2010) Large synteny blocks revealed between *Caenorhabditis elegans* and *Caenorhabditis briggsae* genomes using OrthoCluster. *BMC Genomics*, **11**, 516.
59. Jiao,Y., Wickett,N.J., Ayyampalayam,S., Chanderbali,A.S., Landherr,L., Ralph,P.E., Tomsho,L.P., Hu,Y., Liang,H., Soltis,P.S. *et al.* (2011) Ancestral polyploidy in seed plants and angiosperms. *Nature*, **473**, 97–100.
60. Jaillon,O., Aury,J.M., Noel,B., Policriti,A., Clepet,C., Casagrande,A., Choisne,N., Aubourg,S., Vitulo,N., Jubin,C. *et al.* (2007) The grapevine genome sequence suggests ancestral hexaploidization in major angiosperm phyla. *Nature*, **449**, 463–467.
61. Schmutz,J., Cannon,S.B., Schlueter,J., Ma,J., Mitros,T., Nelson,W., Hyten,D.L., Song,Q., Thelen,J.J., Cheng,J. *et al.* (2010) Genome sequence of the palaeopolyploid soybean. *Nature*, **463**, 178–183.
62. Wei,F., Coe,E., Nelson,W., Bharti,A.K., Engler,F., Butler,E., Kim,H., Goicoechea,J.L., Chen,M., Lee,S. *et al.* (2007) Physical and genetic structure of the maize genome reflects its complex evolutionary history. *PLoS Genet.*, **3**, e123.
63. Salse,J., Bolot,S., Throude,M., Jouffe,V., Piegu,B., Quraishi,U.M., Calcagno,T., Cooke,R., Delseny,M. and Feuillet,C. (2008) Identification and characterization of shared duplications between rice and wheat provide new insight into grass genome evolution. *Plant Cell*, **20**, 11–24.
64. Rizzon,C., Ponger,L. and Gaut,B.S. (2006) Striking similarities in the genomic distribution of tandemly arrayed genes in *Arabidopsis* and rice. *PLoS Comput. Biol.*, **2**, e115.
65. Sampedro,J., Lee,Y., Carey,R.E., dePamphilis,C. and Cosgrove,D.J. (2005) Use of genomic history to improve

phylogeny and understanding of births and deaths in a gene family. *Plant J.*, **44**, 409–419.
66. Guindon,S., Dufayard,J.F., Lefort,V., Anisimova,M., Hordijk,W. and Gascuel,O. (2010) New algorithms and methods to estimate maximum-likelihood phylogenies: assessing the performance of PhyML 3.0. *Syst. Biol.*, **59**, 307–321.
67. Becker,A. and Theissen,G. (2003) The major clades of MADS-box genes and their role in the development and evolution of flowering plants. *Mol. Phylogenet. Evol.*, **29**, 464–489.
68. Freeling,M., Lyons,E., Pedersen,B., Alam,M., Ming,R. and Lisch,D. (2008) Many or most genes in *Arabidopsis* transposed after the origin of the order Brassicales. *Genome Res.*, **18**, 1924–1937.
69. Cannon,S.B., Kozik,A., Chan,B., Michelmore,R. and Young,N.D. (2003) DiagHunter and GenoPix2D: programs for genomic comparisons, large-scale homology discovery and visualization. *Genome Biol.*, **4**, R68.
70. Sinha,A.U. and Meller,J. (2007) Cinteny: flexible analysis and visualization of synteny and genome rearrangements in multiple organisms. *BMC Bioinformatics*, **8**, 82.
71. Ng,M.P., Vergara,I.A., Frech,C., Chen,Q., Zeng,X., Pei,J. and Chen,N. (2009) OrthoClusterDB: an online platform for synteny blocks. *BMC Bioinformatics*, **10**, 192.
72. Fostier,J., Proost,S., Dhoedt,B., Saeys,Y., Demeester,P., Van de Peer,Y. and Vandepoele,K. (2011) A greedy, graph-based algorithm for the alignment of multiple homologous gene lists. *Bioinformatics*, **27**, 749–756.
73. Soderlund,C., Bomhoff,M. and Nelson,W.M. (2011) SyMAP v3.4: a turnkey synteny system with application to plant genomes. *Nucleic Acids Res.*, **39**, e68.