Supplementary information

Supplementary methods

Whole genome shotgun sequencing

Castor bean inbred cultivar Hale[1] (NSL 4773) seeds were obtained from the National Center for Genetic Resources Preservation (NCGRP) at Ft. Collins, Colorado (Germplasm Resources Information Network). Nuclear DNA from etiolated castor bean seedlings grown in a growth chamber was purified as described[2] and was randomly sheared by nebulization, end-repaired with consecutive BAL31 nuclease and T4 DNA polymerase treatments, and size-selected using gel electrophoresis on 1% low-melting-point agarose. After ligation to *Bst*XI adapters, DNA was purified by three rounds of gel electrophoresis to remove excess adapters, and the fragments were ligated into the vector pHOS2 (a modified pBR322 vector) linearized with *Bst*XI. The pHOS2 plasmid contains two *Bst*XI cloning sites immediately flanked by sequencing-primer binding sites. Six libraries with small average insert size (3.5 to 9 kbp) were constructed by electroporation of the ligation reaction into *E. coli*. strain GC10. In addition, two fosmid libraries were constructed using 30 μg of DNA that was sheared by bead beating and end-repaired (as described above). Fragments between 39 and 40 kbp were isolated with a pulse field electrophoresis system and ligated to the blunt-end CopyControl pCC1FOS vector (Epicentre, Madison, WI). Lambda phage packaging and infection were performed following the manufacturer instructions. All clones were plated onto large format (16 × 16 cm) diffusion plates prepared by layering 150 ml of antibiotic-free LB-agar onto a previously set 50-ml layer of LB-agar containing ampicillin or chloramphenicol as required by the vector. Colonies were picked for template preparation using Qbot or QPix colony-picking robots (Genetix, http://www.genetix.com), inoculated into 384-well blocks containing liquid medium, and incubated overnight with shaking. High-purity plasmid DNA was prepared using the DNA purification robotic workstation custom-built by Thermo CRS (http://www.thermo.com) and based on the alkaline lysis miniprep[3] and isopropanol precipitation. The DNA precipitate was washed with 70% ethanol, dried, and re-suspended in 10 mM Tris.HCl buffer containing a trace of blue dextran. The typical yield of plasmid DNA from this method is approximately 600–800 ng per clone, providing sufficient DNA for at least four sequencing reactions per template. Sequencing was carried out using the di-deoxy sequencing method[4]. Two 384-well cycle-sequencing reaction plates were prepared from each plate of plasmid template DNA for opposite-end, paired-sequence reads. Sequencing reactions were completed using the Big Dye Terminator chemistry (Applied Biosystems, http://www.appliedbiosystems.com) and standard M13 forward and reverse primers. Reaction mixtures, thermal cycling profiles, and electrophoresis conditions were optimized to reduce the volume of the Big Dye Terminator mix and to extend read lengths on the AB3730xl sequencers (Applied Biosystems). Sequencing reactions were set up using a Biomek FX (Beckman Coulter, http://www.beckmancoulter.com) pipetting workstation. Robotics was used to aliquot and combine templates with reaction mixes consisting of deoxy- and fluorescently labeled di-deoxy-nucleotides, DNA polymerase, sequencing primers, and reaction buffer in a 5 µl

volume. Bar-coding and tracking systems promoted error-free template and reaction mix handling. After 30–40 consecutive cycles of amplification, reaction products were precipitated with isopropanol, dried at 25ºC, resuspended in water, and transferred to an AB3730xl DNA analyzer.

A total of 2,276,000 paired-end sequence reads were attempted yielding 2,079,000 high quality sequences, of which 12% correspond to fosmid clones (40 kbp insert size), 60% to 9 kbp insert size clones, 10% to 5kbp insert size clones and 18% to 3.5 kbp insert clones. The average read length was 839 bp. All reads were assembled into contigs using the Celera assembler[5] version 3.20 that utilizes an "overlay-layout-consensus" approach to produce consensus sequences or contigs. Celera also uses mate-pair read information to build scaffolds where contigs are ordered and oriented relative to each other. The Celera assembler was run using the default parameters for large genomes. In addition to the normal contigs, the assembler creates so-called "degenerate contigs" which have some kind of problem, such as excessive deviation from the expected level of coverage. We manually inpected the degenerate contigs and recovered approximately 12.4 Mbp of sequences that contained plant gene-like sequences as determined by BLAST analysis. The consensus sequences were entered in an in-house genome annotation relational database called RCA1.

As the genomic DNA used for sequencing was purified from non-axenic seedlings, plant-associated bacteria were likely to be present in our sequence. Therefore, conitgs smaller than 2 kbp that did not show high level of identity to plant organelle sequences (BLASTN E value cutoff < 10^{-50}), and showed sequence similarity to bacterial proteins from available bacterial genome sequences with BLASTX E values < 10^{-20} were removed.

Closure of sequence gaps

In order to increase the quality of the ricin gene family annotation, we performed finishing work on 8 scaffolds that contained members of this gene family to close sequence gaps or ambiguities within the corresponding gene models. Closure was conducted by editing the ends of sequence traces, primer walking on plasmid templates, sequencing genomic PCR products that spanned the gaps, or by transposon insertion and sequencing of selected fosmids clones[6].

Gene prediction and genome annotation

All *Ricinus communis* scaffolds were processed through the TIGR eukaryotic annotation pipeline. Prior to running the gene prediction software, RepeatMasker (http://repeatmasker.org) was used to mask the genomic sequence using a library of known plant repeats from an in-house plant repeat database and novel castor bean repeats identified by running RepeatScout, an algorithm that identifies sequences that are overrepresented in the assembly[7]. In order to prevent incorrect annotation of repeats as genes, we took a conservative approach and any sequence repeated at least 10 times in the genome was considered repetitive. Manual inspection of the list of repeats generated by RepeatScout was carried out to remove members of known gene families that were

wrongly reported as repeats. Further screening by manual review was carried out to remove putative gene families that were mistakenly identified as repeats, resulting in a final set of 1,517 consensus repeat sequences. With the so constructed repeat library, 50.33% of the castor bean genome was masked as repetitive sequences. Low complexity sequences and tandem repeats were identified but not masked because they are often part of protein coding sequences. The RepeatScout library masked 49.88% of the genome while the known plant repeat library masked 8.24% of the genome. Repeats were classified using 2994 Viridiplantae repeats from RepBase[8] and the consensus repetitive sequences identified by RepeatScout (Table 2).

Four gene finders were run on the masked genome: FgenesH gene prediction algorithm trained with a dicotyledonous matrix[9]; Augustus trained with *Arabidopsis*[10]; GlimmerHMM trained with *Arabidopsis*[11]; and SNAP trained with *Arabidopsis*[12].

We used the Program to Assemble Spliced Alignments[13] (PASA) to align 53,516 castor bean cDNA sequences to the castor bean genome. We used all available castor bean cDNA sequences from GenBank at the time, and 52,165 expressed sequence tags (EST) from 5 cDNA non-normalized libraries constructed from mRNAs from leaves, flowers, roots and two different seed developmental stages. cDNA clones were sequenced from the 5' end, except for the root cDNA clones, which were sequenced from both ends to increase the chances of obtaining full-length cDNA sequences. PASA also assembles the aligned cDNA sequences into so-called "PASA assemblies". Using the unmasked castor genome sequence, PASA aligned and assembled approximately 73% of the castor bean cDNA sequences. For a cDNA sequence to be aligned to the genome it should have at least 95% identity along 90% of its length, and consensus splice sites should be present at all inferred exon/intron boundaries. After alignment, PASA generated 8,132 non-redundant cDNA assemblies, of which 5,491 overlapped predicted gene models and 688 identified non-annotated regions. These PASA assemblies were used for identification new gene models as well as to validate or update existing ones. Other PASA assemblies were not incorporated into gene models due intron/exon structure conflicts or because the fragmentary nature of the genome assembly precluded the alignments to meet the stringency criteria.

Sequence homology to nucleotide and protein datasets was computed using the Analysis and Annotation Tool (AAT) package[14] on the unmasked castor bean genome. AAT utilizes a two-step approach consisting of a fast database homology search followed by a rigorous, splice-aware local alignment. The datasets used for AAT analyses included: i) *Oryza sativa* peptides (October 2006 release); ii) *Arabidopsis* proteins (TAIR 6, September 2006 release); iii) an in-house non-redundant amino acid database; iv) A database of transcript assemblies that contains clustered and assembled EST and other cDNA sequences from plant species[15] for which over 1,000 sequences are available in GB at the time (http://plantta.jcvi.org).

Proteins having the highest scoring amino acid alignment to our gene models were incorporated into the gene models using GeneWise[16] to increase protein prediction reliability.

All gene structures predicted by the methods described above as well as the alignments to protein and nucleotide databases were combined into consensus gene models using

Evidence Modeler[17] (EVM), a software package developed at The Institute for Genomic Research (TIGR, now the J.C. Venter Insittute or JCVI) that integrates data from multiple gene prediction programs as well as protein and cDNA similarity searches, in order to achieve the most accurate annotation possible with automated tools. It uses a non-stochastic weighted-evidence combining technique that accounts for both the type and abundance of evidence to compute weighted consensus gene structures. All potential gene structure components were scored based on manually set weights so that exon and intron structures supported by PASA alignments and high quality protein alignments had the highest relevance in determining a gene model's final structure, and the structure predicted by *ab initio* gene finding software were given lower weights according to their accuracy for castor bean. Evidence from transcript assemblies alignments, protein alignments, and gene prediction software were given a weight of 1, while GeneWise protein alignments received a weight of 5, and the weight of PASA assemblies was set at 20. Dynamic programming then was applied by EVM to find the highest scoring consensus gene structure, supported by all available evidence.

Gene models produced by EVM were then updated by new PASA assembly alignments. PASA extended untranslated regions (UTR) and added small missed exons. This resulted in a total of 31,237 gene models of which 19,768 have either EST or protein support (5,316 gene models have castor bean EST support determined by PASA, and 16,848 have protein evidence support determined by AAT searches). 3,150 models were labeled as "partial" because they missed either start and/or stop codons. 354 gene models contained an internal gap, which is represented by "Ns" in the nucleotide sequence and "Xs" in protein sequence, indicating the location and predicted size of the gap.

A dataset of 60 castor bean genes manually modeled based on highly conserved cDNA and protein alignments across multiple plant species were used as reference to evaluate the gene prediction algorithms' performance in comparison with EVM consensus predictions (Supplementary Table 1). Although this is a small set of genes, we used the exons to estimate the specificity and sensitivity of exon prediction by the different gene finder programs as described[17]. Future iterations of the annotation can be improved by using a larger set of genes for training and evaluation of the gene prediction software, as more castor bean cDNA sequences become available.

Gene models were automatically named and their function was assigned by computationally extracting this information from BLASTP searches against the TAIR6 *Arabidopsis* peptides (http://www.arabidopsis.org), Uniprot-Swissprot (http://www.uniprot.org), and experimentally verified Panda (http://www.ebi.ac.uk/panda), Panther (http://www.pantherdb.org), and Interpro (http://www.ebi.ac.uk/interpro) databases. Gene models whose hits in those databases were defined as "unknown function" were labeled "conserved hypothetical protein" in our genome annotation. Gene models with no match in these databases above the selected threshold were labeled "hypothetical protein".

Automated Gene Ontology (http://www.geneontology.org) GO term assignments were done by extrapolating GO terms from matching *Arabidopsis* proteins using BLASTP with an E value threshold of 10^{-40}. Castor bean gene models with no match to *Arabidopsis* were screened against Pfam domains (http://pfam.sanger.ac.uk) and assigned the Pfam

associated GO term, if matches were above the selected cutoff. Altogether, this resulted in the assignment of 43,657 GO terms to 14,991 *R. communis* proteins.

Putative signal peptide sequences were identified using SignalP[18] and TargetP (http://www.cbs.dtu.dk/services/TargetP), and trans-membrane regions were predicted by tmHMM[19]. Castor bean protein domains were also compared against the Pfam database of conserved families[20]. Proteins were organized into putative paralogous families based on conserved domain composition, taking into account both previously identified domains from public databases and potential novel domains identified using independent methods[21,22].

Non-coding RNAs were identified by searching against various RNA libraries. tRNAscan-SE[23] was run on the assembled genomic sequence to identify tRNAs. All 20 tRNAs were found in the genome with a total of 717 copies. rRNA sequences were annotated based on homology to previously published rRNA sequences in plants. snRNA were searched by blasting against the NONCODE database (http://www.noncode.org).

We assigned Enzyme Commission (EC) classification developed by the Nomenclature Committee of the International Union of Biochemistry and Molecular Biology, to provide metabolic pathway annotation. Castor bean proteins were searched against PRIAM profiles[24] using PSI-BLAST, and EC numbers were assigned for hits with an E value lower than 10^{-10}.

Annotation data is displayed in the project website (http://castorbean.jcvi.org), which includes a generic genome browser (http://gmod.org/wiki/GBrowse), where gene models can be viewed in their sequence and genomic context. We used a gene model nomenclature that is composed by the scaffold ID number, followed by a period and the gene model number that consists of a letter "m" followed by the gene model number. This number can be used to locate genes in the castor bean genome browser. Gene models in the genome browser are linked to Manatee pages, which include additional annotation information (http://manatee.sourceforge.net).

The castor bean predicted proteome could be matched to over 3,000 protein domains from Pfam[20], several of which are not present in *Arabidopsis* or poplar, including secondary metabolism genes (Supplementary Figure 1). However, these results may have a substantial error due to inaccuracies of the automatic annotation both in poplar and castor bean.

We also searched for tandem gene duplications and found a total of 2,610 (8% of the total) genes forming part of tandem arrays.

Supplementary references

1. Brigham, R. Registration of castor variety Hale (Reg. No. 3). *Crop Sci* **10**, 457 (1970).
2. Rabinowicz, P. D. *et al*. Differential methylation of genes and retrotransposons facilitates shotgun sequencing of the maize genome. *Nat Genet* **23**, 305-308 (1999).
3. Sambrook, J. & Russell, D. W. *Molecular Cloning. A Laboratory Manual*. Third edn, (Cold Spring Harbor Laboratory Press, 2001).
4. Sanger, F., Nicklen, S. & Coulson, A. R. DNA sequencing with chain-terminating inhibitors. *Proc Natl Acad Sci U S A* **74**, 5463-5467 (1977).
5. Myers, E. W. *et al*. A whole-genome assembly of Drosophila. *Science* **287**, 2196-2204 (2000).
6. Birren, B., Green, E. D., Klapholz, S., Myers, R. M. & Roskams, J. *Genome Analysis. A Laboratory Manual. Analyzing DNA*. Vol. 1 (Cold Spring Harbor Laboratory Press, 1997).
7. Price, A. L., Jones, N. C. & Pevzner, P. A. De novo identification of repeat families in large genomes. *Bioinformatics* **21 Suppl 1**, i351-358 (2005).
8. Jurka, J. *et al*. Repbase Update, a database of eukaryotic repetitive elements. *Cytogenet Genome Res* **110**, 462-467 (2005).
9. Salamov, A. A. & Solovyev, V. V. Ab initio gene finding in Drosophila genomic DNA. *Genome Res* **10**, 516-522 (2000).
10. Stanke, M. & Waack, S. Gene prediction with a hidden Markov model and a new intron submodel. *Bioinformatics* **19 Suppl 2**, ii215-225 (2003).
11. Majoros, W. H., Pertea, M. & Salzberg, S. L. TigrScan and GlimmerHMM: two open source ab initio eukaryotic gene-finders. *Bioinformatics* **20**, 2878-2879 (2004).
12. Korf, I. Gene finding in novel genomes. *BMC Bioinformatics* **5**, 59 (2004).
13. Haas, B. J. *et al*. Improving the Arabidopsis genome annotation using maximal transcript alignment assemblies. *Nucleic Acids Res*. **31**, 5654-5666 (2003).
14. Huang, X., Adams, M. D., Zhou, H. & Kerlavage, A. R. A tool for analyzing and annotating genomic sequences. *Genomics* **46**, 37-45 (1997).
15. Childs, K. L. *et al*. The TIGR Plant Transcript Assemblies database. *Nucleic Acids Res* **35**, D846-851 (2007).
16. Birney, E., Clamp, M. & Durbin, R. GeneWise and Genomewise. *Genome Res* **14**, 988-995 (2004).
17. Haas, B. J. *et al*. Automated eukaryotic gene structure annotation using EVidenceModeler and the Program to Assemble Spliced Alignments. *Genome Biol* **9**, R7 (2008).
18. Bendtsen, J. D., Nielsen, H., von Heijne, G. & Brunak, S. Improved prediction of signal peptides: SignalP 3.0. *J Mol Biol* **340**, 783-795 (2004).
19. Krogh, A., Larsson, B., von Heijne, G. & Sonnhammer, E. L. Predicting transmembrane protein topology with a hidden Markov model: application to complete genomes. *J Mol Biol* **305**, 567-580 (2001).
20. Finn, R. D. *et al*. Pfam: clans, web tools and services. *Nucleic Acids Res* **34**, D247-251 (2006).

21. Haas, B. J. *et al*. Complete reannotation of the Arabidopsis genome: methods, tools, protocols and the final release. *BMC Biol* **3**, 7 (2005).
22. Wortman, J. R. *et al*. Annotation of the Arabidopsis genome. *Plant Physiol* **132**, 461-468 (2003).
23. Lowe, T. M. & Eddy, S. R. tRNAscan-SE: a program for improved detection of transfer RNA genes in genomic sequence. *Nucleic Acids Res* **25**, 955-964 (1997).
24. Claudel-Renard, C., Chevalet, C., Faraut, T. & Kahn, D. Enzyme-specific profiles for genome annotation: PRIAM. *Nucleic Acids Res* **31**, 6633-6639 (2003).

Supplementary tables

Supplementary Table 1. Sensitivity and specificity at exon level by different gene finders and EVM

	Sensitivity	Specificity
FgenesH	0.95	0.93
Augustus	0.84	0.95
GlimmerHMM	0.90	0.96
SNAP	0.86	0.93
EVM	0.99	0.98

Supplementary Table 2: Pfam domains found only in one or two of the three genomes analyzed (castor bean, poplar, and *Arabidopsis*). The numbers of castor bean- and poplar-specific domains may be overestimatied as they are based in purely automatic annotation. Once annotation is improved by manual inspection, some of the gene models with domains found only in castor bean may be deemed misannotations.

Domains specific of *Arabidopsis*:

Pfam accession	Description
PF05617	Arabidopsis thaliana protein of unknown function
PF05325	Protein of unknown function
PF06975	Protein of unknown function
PF07620	SLEI
PF02721	Domain of unknown function DUF223
PF01097	Arthropod defensin
PF07303	Occludin and RNA polymerase II elongation factor ELL
PF06746	Protein of unknown function
PF07727	Reverse transcriptase
PF08226	Domain of unknown function
PF06881	RNA polymerase II transcription factor SIII
PF00321	Plant thionin
PF07557	Shugoshin C terminus
PF04776	Protein of unknown function
PF04642	Protein of unknown function, DUF601
PF05391	Lsm interaction motif
PF00339	Arrestin
PF04845	PurA ssDNA and RNA-binding protein
PF06683	Protein of unknown function
PF02606	tetraacyldisaccharide-1-P 4'-kinase
PF00537	Scorpion toxin-like domain
PF07265	Tapetum specific protein TAP35/TAP44
PF04510	Family of unknown function
PF04325	Protein of unknown function
PF04863	Alliinase EGF-like domain
PF06521	PAR1 protein
PF07918	CAP160 repeat
PF06721	Protein of unknown function
PF03024	Folate receptor family
PF03778	Protein of unknown function
PF07794	Protein of unknown function
PF03299	Transcription factor AP-2
PF03384	Drosophila protein of unknown function, DUF287
PF05278	Arabidopsis phospholipase-like protein
PF06651	Protein of unknown function

Domains specific of poplar:

Pfam accession	Description
PF02603	HPr
PF03626	Prokaryotic Cytochrome C oxidase subunit IV
PF00961	LAGLIDADG endonuclease
PF06434	aconitate hydratase 2, N-terminal domain
PF00325	transcriptional regulator, Crp/Fnr family
PF04257	Exodeoxyribonuclease V, gamma subunit
PF05758	Ycf1
PF01127	Succinate dehydrogenase cytochrome b subunit

PF05013	N-formylglutamate amidohydrolase
PF02233	NAD
PF07793	Protein of unknown function
PF01497	Periplasmic binding protein
PF02501	Bacterial type II secretion system protein I/J
PF03524	Conjugal transfer protein
PF00345	gram-negative pili assembly chaperone, N-terminal domain
PF03576	Peptidase family T4
PF00772	replicative DNA helicase
PF04341	Protein of unknown function, DUF485
PF04262	glutamate-cysteine ligase
PF03605	Anaerobic c4-dicarboxylate membrane transporter
PF06750	Bacterial Peptidase A24 N-terminal domain
PF00871	Acetokinase family
PF06983	3-demethylubiquinone-9 3-methyltransferase domain protein
PF04546	Sigma-70, non-essential region
PF06863	Protein of unknown function
PF01790	prolipoprotein diacylglyceryl transferase
PF04945	YHS domain
PF02435	Levansucrase/Invertase
PF01107	viral movement protein
PF03279	Bacterial lipid A biosynthesis acyltransferase
PF06243	Phenylacetic acid degradation B
PF01693	caulimovirus viroplasmin
PF02607	B12 binding domain
PF05138	Phenylacetic acid catabolic protein
PF00908	dTDP-4-dehydrorhamnose 3,5-epimerase
PF01654	Bacterial Cytochrome Ubiquinol Oxidase
PF07485	Domain of Unknown Function
PF02633	creatininase
PF06995	Phage P2 GpU
PF01943	Polysaccharide biosynthesis protein
PF06481	COX Aromatic Rich Motif
PF01833	IPT/TIG domain
PF07840	FadR C-terminal domain
PF00401	ATP synthase, Delta/Epsilon chain, long alpha-helix domain
PF02600	disulfide bond formation protein DsbB
PF04166	pyridoxal phosphate biosynthetic protein PdxA
PF06032	Protein of unknown function
PF03814	Potassium-transporting ATPase A subunit
PF07947	YhhN-like protein
PF02533	Photosystem II 4 kDa reaction center component
PF00419	Fimbrial protein
PF04993	TfoX N-terminal domain
PF06455	NADH dehydrogenase subunit 5 C-terminus
PF03988	Repeat of Unknown Function
PF05119	phage terminase, small subunit
PF02754	Cysteine-rich domain
PF04965	GPW / gp25 family
PF00662	NADH-Ubiquinone oxidoreductase
PF06158	phage tail protein E
PF02558	Ketopantoate reductase PanE/ApbA
PF02086	D12 class N6 adenine-specific DNA methyltransferase
PF02743	Cache domain
PF01010	NADH-Ubiquinone oxidoreductase
PF04717	phage-related baseplate assembly protein

PF06029	AlkA N-terminal domain
PF04954	Siderophore-interacting protein
PF08021	Siderophore-interacting FAD-binding domain
PF04261	Dyp-type peroxidase family
PF04809	HupH hydrogenase expression protein, C-terminal conserved region
PF02160	cauliflower mosaic virus peptidase
PF01824	MatK/TrnK amino terminal region
PF03459	TOBE domain
PF04002	DNA repair protein RadC
PF02682	allophanate hydrolase subunit 1
PF03379	CcmB protein
PF02899	Phage integrase, N-terminal SAM-like domain
PF01223	DNA/RNA non-specific endonuclease
PF03137	Organic Anion Transporter Polypeptide
PF03480	Bacterial extracellular solute-binding protein, family 7
PF07862	Nitrogen fixation protein of unknown function
PF02642	Uncharacterized ACR, COG2107
PF02468	Photosystem II reaction centre N protein
PF01791	Deoxyribose-phosphate aldolase
PF05227	CHASE3 domain
PF02411	MerT mercuric transport protein
PF06742	Protein of unknown function
PF04461	Protein of unknown function
PF04081	DNA polymerase delta, subunit 4
PF03118	Bacterial RNA polymerase, alpha chain C terminal domain
PF02253	Phospholipase A1
PF01175	urocanate hydratase
PF00821	phosphoenolpyruvate carboxykinase
PF03916	Polysulphide reductase, NrfD
PF00359	phosphoenolpyruvate-dependent sugar phosphotransferase system, EIIA 2
PF04303	Protein of unknown function
PF03934	general secretion pathway protein K
PF01345	Domain of unknown function
PF02378	Phosphotransferase system, EIIC
PF00912	Transglycosylase
PF02419	PsbL protein
PF01311	Bacterial export proteins, family 1
PF02420	Insect antifreeze protein
PF04060	Putative Fe-S cluster
PF00284	Lumenal portion of Cytochrome b559, alpha
PF01051	Initiator RepB protein
PF05159	Capsule polysaccharide biosynthesis protein
PF00796	photosystem I reaction center subunit VIII
PF08132	S-adenosyl-l-methionine decarboxylase leader peptide
PF07804	HipA-like C-terminal domain
PF03681	Uncharacterised protein family
PF01475	transcriptional regulator, Fur family
PF03976	Polyphosphate kinase 2
PF07478	D-ala D-ala ligase C-terminus
PF01333	Apocytochrome F, C-terminal
PF02674	CvpA family protein
PF00712	DNA polymerase III beta subunit, N-terminal domain
PF04977	Septum formation initiator
PF03748	flagellar basal body-associated protein FliL
PF02355	Protein export membrane protein
PF01701	photosystem I reaction center subunit IX

PF03775	Septum formation inhibitor MinC, C-terminal domain
PF04205	FMN-binding domain
PF02965	Vitamin B12 dependent methionine synthase, activation domain
PF03979	Sigma-70 factor, region 11
PF05816	Toxic anion resistance protein
PF02306	major spike protein
PF04403	Paraquat-inducible protein A
PF01910	Protein of unknown function
PF02554	Carbon starvation protein CstA
PF01520	N-acetylmuramoyl-L-alanine amidase
PF06169	Protein of unknown function
PF01252	signal peptidase
PF00737	photosystem II 10 kDa phosphoprotein
PF03796	DnaB-like helicase C terminal domain
PF02694	Uncharacterized BCR, YnfA/UPF0060 family
PF01515	Phosphate acetyl/butyryl transferase
PF05954	Phage late control gene D protein
PF02950	Conotoxin
PF06035	Bacterial protein of unknown function
PF03972	MmgE/PrpD family
PF04378	Protein of unknown function
PF04860	phage portal protein
PF05838	Protein of unknown function
PF01730	UreF
PF02572	ATP:corrinoid adenosyltransferase BtuR/CobO/CobP
PF01144	Coenzyme A transferase
PF02669	K+-transporting ATPase, c chain
PF04216	Protein involved in formate dehydrogenase formation
PF03745	Domain of unknown function
PF07063	Protein of unknown function
PF03922	OmpW family
PF00455	transcriptional regulator, DeoR family
PF00499	NADH-ubiquinone/plastoquinone oxidoreductase chain 6
PF04198	Putative sugar-binding domain
PF00417	Ribosomal protein S3, N-terminal domain
PF07549	SecD/SecF GG Motif
PF02962	5-carboxymethyl-2-hydroxymuconate delta isomerase
PF00873	RND transporter, Hydrophobe/Amphiphile Efflux-1
PF04052	TolB amino-terminal domain
PF04352	ProQ activator of osmoprotectant transporter ProP
PF03004	Plant transposase
PF06891	P2 phage tail completion protein R
PF01498	transposase
PF06865	Protein of unknown function
PF03960	ArsC family
PF01689	Hydratase/decarboxylase
PF02310	B12 binding domain
PF05272	Virulence-associated protein E
PF03449	Prokaryotic transcription elongation factor, GreA/GreB, N-terminal domain
PF04381	Putative exonuclease, RdgC
PF04314	Protein of unknown function
PF08187	Myoactive tetradecapeptides family
PF06805	bacteriophage lambda tail assembly protein I
PF02643	uncharacterized ACR, COG1430
PF03702	Uncharacterised protein family
PF04264	YceI like family

PF02615	Malate/L-lactate dehydrogenase
PF00283	Cytochrome b559, alpha
PF01580	FtsK/SpoIIIE family
PF00978	RNA dependent RNA polymerase
PF05976	Bacterial membrane protein of unknown function
PF00872	transposase, Mutator family
PF04655	Aminoglycoside/hydroxyurea antibiotic resistance kinase
PF02635	DsrE/DsrF-like family
PF06791	prophage tail length tape measure protein
PF07086	Protein of unknown function
PF00375	transporter, dicarboxylate/amino acid:cation
PF04183	siderophore biosynthesis protein, IucA/IucC family
PF03797	Autotransporter beta-domain
PF04279	intracellular septation protein A
PF06821	Alpha/Beta hydrolase family of unknown function
PF06264	Protein of unknown function
PF04985	phage tail tube protein FII
PF06527	TniQ
PF06751	ethanolamine ammonia-lyase, large subunit
PF07536	HWE histidine kinase
PF01418	transcriptional regulator, RpiR family
PF02381	Domain of unknown function UPF0040 family
PF07475	HPr Serine kinase C-terminus
PF08007	Cupin superfamily protein
PF07733	DNA polymerase III, alpha subunit
PF04226	Transglycosylase associated protein

Domains specific of castor bean:

Pfam accession	Description
PF08472	Sucrose-6-phosphate phosphohydrolase C-terminal
PF08311	Mad3/BUB1 homology region 1
PF08342	Phosphopentomutase N-terminal
PF05443	ROS/MUCR transcriptional regulator protein
PF04271	DnaD-like domain
PF08282	haloacid dehalogenase-like hydrolase
PF08668	HDOD domain
PF08645	Polynucleotide kinase 3 phosphatase
PF08699	Domain of unknown function
PF09298	Domain of unknown function
PF09339	IclR helix-turn-helix domain
PF04313	Type I restriction enzyme R protein N terminus
PF03808	Glycosyl transferase WecB/TagA/CpsF family
PF08718	Glycolipid transfer protein
PF01904	Protein of unknown function
PF01653	NAD-dependent DNA ligase adenylation domain
PF08498	Sterol methyltransferase C-terminal
PF03872	Anti sigma-E protein RseA, N-terminal domain
PF08659	KR domain
PF02108	flagellar assembly protein FliH
PF08424	Protein of unknown function
PF04351	pilus assembly protein PilP
PF03888	MucB/RseB family
PF01882	Protein of unknown function
PF07023	Protein of unknown function
PF08458	Plant pleckstrin homology-like region
PF02120	flagellar hook-length control protein

PF08780	Nucleotidyltransferase substrate binding protein like
PF09261	Alpha mannosidase, middle domain
PF02806	Alpha amylase, C-terminal all-beta domain
PF03417	Acyl-coenzyme A:6-aminopenicillanic acid acyl-transferase
PF08238	Sel1 repeat
PF08670	MEKHLA domain
PF07508	Recombinase
PF08448	PAS fold
PF06689	ClpX C4-type zinc finger
PF09072	Translation machinery associated TMA7
PF08510	PIG-P
PF08546	Ketopantoate reductase PanE/ApbA C terminal
PF05853	Prokaryotic protein of unknown function
PF05681	fumarate hydratase I, N-terminal region or alpha subunit
PF01098	cell cycle protein, FtsW/RodA/SpoVE family
PF08620	RPAP1-like, C-terminal
PF08818	Domain of unknown function
PF08170	POPLD
PF04032	RNAse P Rpr2/Rpp21 subunit domain
PF08787	Alginate lyase
PF08571	Yos1-like
PF08491	Squalene epoxidase
PF08246	Cathepsin propeptide inhibitor domain
PF06041	Bacterial protein of unknown function
PF05235	CHAD domain
PF01232	Mannitol dehydrogenase rossman domain
PF04999	cell division protein FtsL
PF01613	Flavin reductase like domain
PF02277	Phosphoribosyltransferase
PF08610	Peroxisomal membrane protein
PF08771	Rapamycin binding domain
PF08767	CRM1 C terminal
PF08502	LeuA allosteric
PF04331	Family of unknown function
PF04445	Protein of unknown function
PF08327	Activator of Hsp90 ATPase homolog 1-like protein
PF08392	FAE1/Type III polyketide synthase-like protein
PF08551	Eukaryotic integral membrane protein
PF09179	Domain of unkown function
PF09243	Mitochondrial small ribosomal subunit Rsm22
PF07468	Agglutinin
PF08285	Dolichol-phosphate mannosyltransferase subunit 3
PF05662	Haemagglutinin
PF08541	3-Oxoacyl-[acyl-carrier-protein
PF08615	Ribonuclease H1 small subunit
PF08241	Methyltransferase domain
PF02933	Cell division protein 48
PF04967	HTH DNA binding domain
PF09328	Domain of unknown function
PF08324	PUL domain
PF02561	flagellar protein FliS
PF06202	Amylo-alpha-1,6-glucosidase
PF08597	Translation initiation factor eIF3 subunit
PF01297	Periplasmic solute binding protein family
PF09341	Transcription factor Pcc1
PF02698	Uncharacterized ACR, COG1434

PF05012	Prophage maintenance system killer protein
PF09336	Vps4 C terminal oligomerisation domain
PF08623	TATA-binding protein interacting
PF00358	phosphoenolpyruvate-dependent sugar phosphotransferase system, EIIA 1
PF08755	Hemimethylated DNA-binding protein YccV like
PF08514	STAG domain
PF08534	Redoxin
PF08291	Peptidase M15
PF05164	Family of unknown function
PF08298	PrkA AAA domain
PF02660	Domain of unknown function DUF
PF08603	DE Adenylate cyclase associated
PF05649	peptidase family M13
PF03235	Protein of unknown function DUF262
PF02686	glutamyl-tRNA
PF08996	DNA Polymerase alpha zinc finger
PF08372	Plant phosphoribosyltransferase C-terminal
PF01430	chaperonin HslO
PF08625	Utp13 specific WD40 associated domain
PF03739	putative permease, YjgP/YjgQ family
PF03313	Serine dehydratase alpha chain
PF08799	pre-mRNA processing factor 4
PF08389	Exportin 1-like protein
PF04329	Family of unknown function
PF08652	RAI1 like
PF06684	Protein of unknown function
PF08125	Mannitol dehydrogenase C-terminal domain
PF00532	Periplasmic binding proteins and sugar binding domain of the LacI family
PF08737	Rgp1
PF08781	Transcription factor DP
PF05717	IS66 family element, Orf2 protein
PF05036	sporulation and cell division repeat protein
PF08704	tRNA methyltransferase complex GCD14 subunit
PF08250	Sperm-activating peptides
PF04632	Fusaric acid resistance protein conserved region
PF08295	Histone deacetylase
PF02121	Phosphatidylinositol transfer protein
PF03050	IS66 family element, transposase
PF08569	Mo25-like
PF06719	AraC-type transcriptional regulator N-terminus
PF08536	Whirly transcription factor
PF01087	Galactose-1-phosphate uridyl transferase, N-terminal domain
PF00353	type I secretion target GGXGXDXXX repeat
PF07110	EthD protein
PF05067	Manganese containing catalase
PF07470	glycosyl hydrolase, family 88
PF05772	NinB protein
PF02979	Nitrile hydratase, alpha chain
PF01568	Molydopterin dinucleotide binding domain
PF08239	Bacterial SH3 domain
PF08019	Domain of unknown function
PF08501	Shikimate dehydrogenase substrate binding domain
PF05594	Haemagluttinin repeat
PF09324	Domain of unknown function
PF09118	Domain of unknown function
PF08646	Replication factor-A C terminal domain

PF07167	Poly-beta-hydroxybutyrate polymerase
PF04402	Protein of unknown function
PF08545	3-Oxoacyl-[acyl-carrier-protein
TIGR03168	1-phosphofructokinase
PF08920	Splicing factor 3B subunit 1
PF08310	LGFP repeat
PF02055	O-Glycosyl hydrolase family 30
PF08927	Domain of unknown function
PF08697	Tuftelin interacting protein 11
PF08698	Fcf2 pre-rRNA processing
PF06187	Protein of unknown function
PF00700	Bacterial flagellin C-terminus
PF06580	Histidine kinase
PF08752	Coatomer gamma subunit appendage domain
PF02238	Cytochrome c oxidase subunit VIIa
PF07772	TP901-1 ORF40-like protein
PF08490	Domain of unknown function
PF08590	Domain of unknown function
PF08628	Sorting nexin C terminal
PF09333	ATG C terminal domain
PF08338	Domain of unknown function
PF08991	Domain of unknown function
PF08557	Sphingolipid Delta4-desaturase
PF01292	Nickel-dependent hydrogenases b-type cytochrome subunit
PF02625	xanthine dehydrogenase accessory factor, putative
PF08784	Replication protein A C terminal
PF05299	peptidase, M61
PF04330	Family of unknown function
PF02433	Cytochrome C oxidase, mono-heme subunit/FixO
PF01242	6-pyruvoyl tetrahydropterin synthase
PF01339	protein-glutamate methylesterase CheB
PF01815	rop protein
PF08417	Pheophorbide a oxygenase
PF03548	Outer membrane lipoprotein carrier protein LolA
PF08676	MutL C terminal dimerisation domain
PF09262	Peroxisome biogenesis factor 1, N-terminal
PF08449	UAA transporter family
PF04962	5-keto 4-deoxyuronate isomerase
PF06965	Na+/H+ antiporter 1
PF02706	Chain length determinant protein
PF09258	EXTL2, alpha-1,4-N-acetylhexosaminyltransferase
PF08642	Histone deacetylation protein Rxt3
PF07484	Phage Tail Collar Domain
PF08325	WLM domain
PF08513	LisH
PF03747	ADP-ribosylglycohydrolase
PF06629	MltA-interacting protein MipA
PF07001	BAT2 N-terminus
PF09229	Activator of Hsp90 ATPase, N-terminal
PF03819	MazG nucleotide pyrophosphohydrolase domain
PF08506	Cse1
PF09329	Primase zinc finger
PF07494	Two component regulator propeller
PF03472	Autoinducer binding domain
PF04946	DGPF domain
PF05065	phage capsid family

PF08433	Chromatin associated protein KTI12
PF09273	Rubisco LSMT substrate-binding
PF09070	PFU
PF08542	Replication factor C
PF01878	Protein of unknown function
PF05658	Hep_Hag
PF08694	Protein of unknown function
PF02753	gram-negative pili assembly chaperone, C-terminal domain
PF05860	haemagglutination activity domain
PF08638	Mediator complex subunit MED14
PF08766	DEK C terminal domain
PF07648	Kazal-type serine protease inhibitor domain
PF08722	TnsA endonuclease N terminal
PF08880	QLQ
PF09265	Cytokinin dehydrogenase 1, FAD and cytokinin binding
PF08381	Disease resistance/zinc finger/chromosome condensation-like region
PF08267	Cobalamin-independent synthase, N-terminal domain
PF00496	bacterial extracellular solute-binding proteins, family 5
PF09282	Mago binding
PF09313	Domain of unknown function
PF08321	PPP5
PF02344	Myc leucine zipper domain
PF08334	Bacterial type II secretion system protein G
PF05145	Putative ammonia monooxygenase
PF08279	HTH domain
PF07691	PA14 domain
PF05960	Bacterial protein of unknown function
PF01041	DegT/DnrJ/EryC1/StrS aminotransferase family
PF00896	phosphorylases family 2
PF08450	SMP-30/Gluconolaconase/LRE-like region
PF00768	D-alanyl-D-alanine carboxypeptidase
PF09320	Domain of unknown function
PF08825	E2 binding domain
PF05736	OmpF membrane domain
PF03845	Spore germination protein
PF03963	flagellar hook capping protein
PF07228	Stage II sporulation protein E
PF04336	Protein of unknown function, DUF479
PF03648	Glycosyl hydrolase family 67
PF08330	Protein of unknown function
PF07394	Protein of unknown function
PF06073	Bacterial protein of unknown function
PF02397	Bacterial sugar transferase
PF08672	Anaphase promoting complex
PF08292	RNA polymerase III subunit Rpc25
PF08613	Cyclin
PF05336	Protein of unknown function
PF08606	Prp19/Pso4-like
PF08242	Methyltransferase domain
PF08388	Group II intron, maturase-specific domain
PF09325	Vps5 C terminal like
PF08785	Ku C terminal domain like
PF03349	Outer membrane protein transport protein
PF02342	Bacterial stress protein
PF08234	Kinetochore protein Spc25
PF01548	Transposase

PF07497	Rho termination factor, RNA-binding domain
PF06776	Invasion associated locus B
PF04820	Tryptophan halogenase
PF08561	Mitochondrial ribosomal protein L37
PF06483	Chitinase C
PF01757	putative acyltransferase
PF08806	Sep15/SelM redox domain
PF09066	Beta2-adaptin appendage, C-terminal sub-domain
PF08512	Histone chaperone Rttp106-like
PF08523	Multiprotein bridging factor 1
PF07681	DoxX
PF08318	COG4 transport protein
PF08269	Cache domain
PF08244	Glycosyl hydrolases family 32 C terminal
PF08370	Plant PDR ABC transporter associated
PF08617	Kinase binding protein CGI-121
PF03965	transcriptional regulator, BlaI/MecI/CopY family
PF08446	PAS fold
PF09084	NMT1/THI5 like
PF07811	TadE-like protein
PF05947	Bacterial protein of unknown function
PF09110	HAND
PF09296	NADH pyrophosphatase-like rudimentary NUDIX domain
PF08637	ATP synthase regulation protein NCA2
PF07277	SapC
PF09268	Clathrin, heavy-chain linker
PF03713	Domain of unknown function
PF03119	NAD-dependent DNA ligase C4 zinc finger domain
PF04011	LemA family
PF08696	DNA replication factor Dna2
PF08777	RNA binding motif
PF02547	Queuosine biosynthesis protein
PF09139	Mitochondrial matrix Mmp37
PF05992	SbmA/BacA-like family
PF06055	Exopolysaccharide synthesis, ExoD
PF08662	Eukaryotic translation initiation factor eIF2A
PF08492	SRP72 RNA-binding domain
PF03668	Uncharacterised P-loop ATPase protein family
PF05433	Rickettsia 17 kDa surface antigen
PF05524	phosphoenolpyruvate-protein phosphotransferase, N-terminal
PF09177	Syntaxin 6, N-terminal
PF05683	fumarate hydratase I, C-terminal region or beta subunit
PF05345	Putative Ig domain
PF04471	Restriction endonuclease
PF01523	TldD/PmbA family
PF03886	Protein of unknown function
PF02289	Cyclohydrolase
PF08712	Scaffold protein Nfu/NifU N terminal
PF03120	NAD-dependent DNA ligase OB-fold domain
PF06748	Protein of unknown function
PF08879	WRC
PF07506	RepB plasmid partitioning protein
PF01219	Prokaryotic diacylglycerol kinase
PF08442	ATP-grasp domain
PF01887	Protein of unknown function
PF04390	Rare lipoprotein B family

PF04964	Flp/Fap pilin component
PF00135	Carboxylesterase
PF03734	ErfK/YbiS/YcfS/YnhG
PF02872	5'-nucleotidase, C-terminal domain
PF02211	nitrile hydratase, beta subunit
PF09334	tRNA synthetases class I
PF09103	BRCA2, oligonucleotide/oligosaccharide-binding, domain 1
PF05523	WxcM-like, C-terminal
PF01555	DNA methylase
PF08783	DWNN domain
PF09173	Initiation factor eIF2 gamma, C terminal
PF05866	crossover junction endodeoxyribonuclease RusA
PF08802	Cytochrome B6-F complex Fe-S subunit
PF07505	Phage protein Gp37/Gp68
PF08772	Nin one binding
PF08373	RAP domain
PF01522	Polysaccharide deacetylase
PF08264	Anticodon-binding domain
PF08544	GHMP kinases C terminal
PF05673	Protein of unknown function
PF03707	Bacterial signalling protein N terminal repeat
PF02684	lipid-A-disaccharide synthetase
PF08278	DNA primase DnaG DnaB-binding
PF08911	NUP50
PF05930	transcriptional regulator, AlpA family
PF08797	HIRAN domain
PF05359	Domain of Unknown Function
PF06423	GWT1
PF05494	toluene tolerance protein Ttg2D
PF01454	MAGE family
PF02614	glucuronate isomerase
PF08847	Domain of unknown function
PF08666	SAF domain
PF08839	DNA replication factor CDT1 like
PF07045	Protein of unknown function
PF08271	TFIIB zinc-binding
PF08612	TATA-binding related factor
PF08356	EF hand associated
PF08423	Rad51
PF08281	Sigma-70, region 4
PF04014	SpoVT / AbrB like domain
PF02583	Uncharacterized BCR, COG1937
PF09279	Phosphoinositide-specific phospholipase C, efhand-like
PF06320	GCN5-like protein 1
PF08352	Oligopeptide/dipeptide transporter, C-terminal region
PF08661	Replication factor A protein 3
PF08459	UvrC Helix-hairpin-helix N-terminal
PF06146	protein PsiE
PF07152	YaeQ protein
PF06996	Protein of unknown function
PF06114	Domain of unknown function
PF08436	1-deoxy-D-xylulose 5-phosphate reductoisomerase C-terminal
PF08648	Protein of unknown function
PF00016	Ribulose bisphosphate carboxylase large chain, catalytic domain
PF02563	Polysaccharide biosynthesis/export protein
PF07836	DmpG-like communication domain

PF03985	Paf1
PF08477	Miro-like protein
PF08743	Nse4
PF08572	pre-mRNA processing factor 3
PF08547	Complex I intermediate-associated protein 30
PF08719	Domain of unknown function
PF08701	GNL3L/Grn1 putative GTPase
PF08555	Eukaryotic family of unknown function
PF09111	SLIDE
PF01235	Sodium:alanine symporter family
PF07589	Protein of unknown function
PF08245	Mur ligase middle domain
PF08326	Acetyl-CoA carboxylase, central region
PF02082	Transcriptional regulator
PF08387	FBD
PF08243	SPT2 chromatin protein
PF03550	outer membrane lipoprotein LolB
PF07971	glycosyl hydrolase, family 92
PF03717	Penicillin-binding Protein dimerisation domain
PF08774	VRR-NUC domain
PF08711	Transcription elongation factor S-II protein N terminal
PF02350	UDP-N-acetylglucosamine 2-epimerase
PF08235	LNS2
PF06761	ImcF-related
PF08660	Oligosaccharide biosynthesis protein Alg14 like
PF02113	D-Ala-D-Ala carboxypeptidase 3
PF05344	Domain of Unknown Function
PF04476	Protein of unknown function
PF06293	Lipopolysaccharide kinase
PF01628	HrcA protein C terminal domain
PF08443	RimK-like ATP-grasp domain
PF08328	Adenylosuccinate lyase C-terminal
PF02868	Thermolysin metallopeptidase, alpha-helical domain
PF08644	FACT complex subunit
PF08263	Leucine rich repeat N-terminal domain
PF01321	Creatinase
PF03121	Herpesviridae UL52/UL70 DNA primase
PF07987	Nuclear export factor GLE1
PF08375	Proteasome regulatory subunit C-terminal
PF08447	PAS fold
PF09079	CDC6, C terminal
PF08538	Protein of unknown function
PF04524	Protein of unknown function, DUF586
PF02604	addiction module antitoxin, Axe family
PF02368	Bacterial Ig-like domain
PF08626	Transport protein Trs120
PF09297	NADH pyrophosphatase zinc ribbon domain
PF08276	PAN-like domain
PF03428	replication protein C
PF03186	CobD/CbiB protein
PF09269	Domain of unknown function
PF08293	Mitochondrial ribosomal subunit S27
PF05229	Spore Coat Protein U domain
PF08312	cwf21
PF08414	Respiratory burst NADPH oxidase
PF00030	Beta/Gamma crystallin

PF08315	cwf18 pre-mRNA splicing factor
PF08540	Hydroxymethylglutaryl-coenzyme A synthase C terminal
PF03575	Peptidase family S51
PF08288	PIGA
PF08576	Eukaryotic protein of unknown function
PF02421	ferrous iron transport protein B
PF09278	MerR, DNA binding
PF08294	TIM21
PF08621	RPAP1-like, N-terminal
PF09088	MIF4G like
PF08543	Phosphomethylpyrimidine kinase
PF09180	Prolyl-tRNA synthetase, C-terminal
PF09280	XPC-binding domain
PF01695	IstB-like ATP binding protein
PF06559	2'-deoxycytidine 5'-triphosphate deaminase
PF04355	SmpA / OmlA family
PF09192	Actin-fragmin kinase, catalytic
PF08367	Peptidase M16C associated
PF07042	TrfA protein
PF09326	Domain of unknown function
PF08614	Autophagy protein 16
PF05426	Alginate lyase
PF00959	phage lysozyme
PF03412	peptidase, C39 family
PF00296	Luciferase-like monooxygenase
PF03883	Protein of unknown function
PF06812	ImpA-related N-terminal
PF02677	Uncharacterized BCR, COG1636
PF04087	Domain of unknown function
PF08801	Nup133 N terminal like
PF09335	SNARE associated Golgi protein
PF08746	RING-like domain
PF08265	YL1 nuclear protein C-terminal domain
PF09090	MIF4G like
PF09285	Elongation factor P, C-terminal
PF08608	Wyosine base formation
PF08240	Alcohol dehydrogenase GroES-like domain
PF00245	alkaline phosphatase family protein
PF08351	Domain of unknown function
PF08511	COQ9
PF09340	Histone acetyltransferase subunit NuA4
PF08323	Starch synthase catalytic domain
PF06571	Protein of unknown function
PF07745	glycosyl hydrolase, family 53
PF08790	LYAR-type C2HC zinc finger
PF06189	5'-nucleotidase
PF09127	Leukotriene A4 hydrolase, C-terminal
PF08355	EF hand associated
PF08669	Glycine cleavage T-protein C-terminal barrel domain
PF01850	PIN domain
PF08640	U3 small nucleolar RNA-associated protein 6
PF06745	KaiC
PF03631	Ribonuclease BN-like family
PF09138	Urm1
PF05569	peptidase, M56 family
PF08585	Domain of unknown function

PF02278	Polysaccharide lyase family 8, super-sandwich domain
PF02582	Uncharacterized ACR, YagE family COG1723
TIGR02477	diphosphate--fructose-6-phosphate 1-phosphotransferase

Domains shared by castor bean and poplar:

Pfam accession	Description
PF02481	SMF family
PF01478	peptidase, A24
PF02371	Transposase IS116/IS110/IS902 family
PF05532	CsbD-like
PF02653	amino acid or sugar ABC transport systems, permease protein
PF06144	DNA polymerase III, delta subunit
PF00420	NADH-ubiquinone/plastoquinone oxidoreductase chain 4L
PF00115	Cytochrome C and Quinol oxidase polypeptide I
PF02464	competence/damage-inducible protein CinA
PF00437	Type II/IV secretion system protein
PF00528	ABC transporter, permease protein
PF07238	Type IV pilus assembly protein PilZ
PF00589	site-specific recombinase, phage integrase family
PF01551	M23 peptidase domain protein
PF03180	NLPA lipoprotein
PF00196	transcriptional regulator, LuxR family
PF00563	cyclic diguanylate phosphodiesterase
PF01022	transcriptional regulator, ArsR family
PF01032	ABC transporter, iron chelate uptake transporter
PF03992	antibiotic biosynthesis monooxygenase
PF03591	AzlC protein
PF04324	BFD-like [2Fe-2S] binding domain
PF02788	Ribulose bisphosphate carboxylase large chain, N-terminal domain
PF01037	transcriptional regulator, AsnC family
PF04453	Organic solvent tolerance protein
PF02634	FdhD/NarQ family
PF00329	Respiratory-chain NADH dehydrogenase, 30 Kd subunit
PF02496	ABA/WDS induced protein
PF05405	Mitochondrial ATP synthase B chain precursor
PF00421	Photosystem II protein
PF04506	Rft protein
PF03831	PhnA protein
PF07730	Histidine kinase
PF03354	phage terminase, large subunit, putative
PF03466	LysR substrate binding domain
PF01618	transporter, MotA/TolQ/ExbB proton channel family
PF03732	Retrotransposon gag protein
PF04865	baseplate J-like protein
PF03544	Gram-negative bacterial tonB protein
PF00381	PTS HPr component phosphorylation site
PF06912	Protein of unknown function
PF07702	UbiC transcription regulator-associated
PF01276	Orn/Lys/Arg decarboxylase, major domain
PF07244	Surface antigen variable number repeat
PF05157	GSPII_E N-terminal domain
PF01614	transcriptional regulator, IclR family, C-terminal domain
PF05995	Cysteine dioxygenase type I
PF00239	Resolvase, N terminal domain
PF00015	Methyl-accepting chemotaxis protein
PF00264	Common central domain of tyrosinase

PF04350	Pilus assembly protein, PilO
PF00223	Photosystem I psaA and psaB proteins
PF03776	Septum formation topological specificity factor MinE
PF03401	Bordetella uptake gene
PF00593	TonB-dependent receptor
PF01638	putative transcriptional regulator
PF07498	Rho termination factor, N-terminal domain
PF00158	Sigma-54 interaction domain
PF02515	CAIB/BAIF family
PF00665	Integrase core domain
PF06803	Protein of unknown function
PF02205	WH2 motif
PF02954	transcriptional regulator, Fis family
PF00440	transcriptional regulator, TetR family
PF03527	RHS protein
PF02659	Domain of unknown function DUF
PF03929	PepSY-associated TM helix
PF02687	efflux ABC transporter, permease protein
PF02049	flagellar hook-basal body complex protein FliE
PF03711	Orn/Lys/Arg decarboxylase, C-terminal domain
PF05137	Fimbrial assembly protein
PF00990	GGDEF domain
PF04481	Protein of unknown function
PF02322	Cytochrome oxidase subunit II
PF00161	Ribosome inactivating protein
PF03616	Sodium/glutamate symporter
PF07080	Protein of unknown function
PF01702	Queuine tRNA-ribosyltransferase
PF02129	X-Pro dipeptidyl-peptidase
PF06250	Protein of unknown function
PF01558	2-oxoacid:ferredoxin/flavodoxin oxidoreductases, gamma subunit
PF06429	Domain of unknown function
PF07196	Flagellin hook IN motif
PF07660	Secretin and TonB N terminus short domain
PF02655	Domain of unknown function DUF201
PF01514	Secretory protein of YscJ/FliF family
PF00881	nitroreductase family protein
PF00263	Bacterial type II and III secretion system protein
PF02311	AraC-like ligand binding domain
PF00668	Condensation domain
PF07729	FCD domain
PF01584	CheW-like domain
PF07715	TonB-dependent receptor plug domain
PF03958	Bacterial type II/III secretion system short domain
PF02321	outer membrane efflux protein
PF00267	Gram-negative porin
PF02627	carboxymuconolactone decarboxylase family
PF01075	Heptosyltransferase
PF00905	Penicillin binding protein transpeptidase domain
PF02447	GntP family permease
PF01594	Domain of unknown function
PF06838	aluminium resistance protein
PF03928	Domain of unknown function
PF05951	Bacterial protein of unknown function
PF02417	Chromate transporter
PF04932	O-antigen polymerase

PF01527	transposase
PF00165	transcriptional regulator, AraC family
PF03315	Serine dehydratase beta chain
PF00356	transcriptional regulator, lacI family
PF01863	Protein of unknown function
PF05593	RHS Repeat
PF02465	flagellar hook-associated protein 2
PF00497	Bacterial extracellular solute-binding proteins, family 3
PF07963	Prokaryotic N-terminal methylation motif
PF00376	transcriptional regulator, MerR family
PF01272	Prokaryotic transcription elongation factor, GreA/GreB, C-terminal domain
PF00486	Transcriptional regulatory protein, C terminal
PF05840	Bacteriophage replication gene A protein
PF00126	transcriptional regulator, LysR family
PF00652	ricin-type beta-trefoil lectin domain
PF01526	transposase
PF04972	Putative phospholipid-binding domain
PF00669	Bacterial flagellin N-terminus
PF02472	transport energizing protein, ExbD/TolR family
PF03705	CheR methyltransferase, all-alpha domain
PF00124	Photosynthetic reaction center protein
PF02308	putative Mg2+ transporter-C
PF01609	Transposase
PF01547	Bacterial extracellular solute-binding protein
PF01810	translocator protein, LysE family
PF00816	H-NS histone family
PF00975	Thioesterase domain
PF02107	flagellar L-ring protein FlgH
PF01970	Integral membrane protein
PF07805	HipA-like N-terminal domain
PF00392	transcriptional regulator, GntR family
PF01047	transcriptional regulator, MarR family
PF00771	FHIPEP family
PF02392	Ycf4
PF00216	DNA-binding protein HU
PF00346	Respiratory-chain NADH dehydrogenase, 49 Kd subunit
PF01739	CheR methyltransferase, SAM binding domain
PF00116	Cytochrome C oxidase subunit II, periplasmic domain
PF04979	Protein phosphatase inhibitor 2
PF01052	Surface presentation of antigens
PF04879	Molybdopterin oxidoreductase Fe4S4 domain
PF00482	Bacterial type II secretion system protein F domain
PF03743	Bacterial conjugation TrbI-like protein
PF00691	OmpA family
PF03657	Uncharacterised protein family
PF02608	Basic membrane protein
PF00529	auxiliary transport protein, membrane fusion protein
PF04214	Protein of unknown function, DUF
PF04397	LytTr DNA-binding domain
PF04773	sigma factor regulatory protein, FecR/PupR family
PF01957	Nodulation efficiency protein D
PF00672	HAMP domain
PF00460	Flagella basal body rod protein
PF01081	KDPG and KHG aldolase

Domains shared by castor bean and *Arabidopsis*:

Pfam accession	Description
PF07160	Protein of unknown function
PF02445	Quinolinate synthetase A protein
PF06108	Protein of unknown function
PF05206	Protein of unknown function
PF03850	Transcription factor Tfb4
PF02757	YLP motif
PF03853	YjeF-related protein N-terminus
PF04781	Protein of unknown function
PF02091	glycyl-tRNA synthetase, alpha subunit
PF04616	glycosyl hydrolase, family 43
PF07540	Nucleolar complex-associated protein
PF06876	Plant self-incompatibility response
PF00610	Domain found in Dishevelled, Egl-10, and Pleckstrin
PF05462	Slime mold cyclic AMP receptor
PF05097	Protein of unknown function
PF00708	acylphosphatase
PF00852	Fucosyl transferase
PF07172	Glycine rich protein family
PF02551	Acyl-CoA thioesterase
PF04615	Utp14 protein
PF06695	Putative small multi-drug export protein
PF04455	LOR/SDH bifunctional enzyme conserved region
PF06424	PRP1 splicing factor, N-terminal
PF06658	Protein of unknown function
PF04931	DNA polymerase V
PF05871	Eukaryotic protein of unknown function
PF07959	L-fucokinase
PF04547	Protein of unknown function, DUF590
PF06862	Protein of unknown function
PF01927	Protein of unknown function
PF07720	Tetratricopeptide repeat
PF06298	Photosystem II protein Y
PF07842	GC-rich sequence DNA-binding factor-like protein
PF05994	Cytoplasmic Fragile-X interacting family
PF00835	SNAP-25 family
PF05162	Ribosomal protein L41
PF04147	Nop14-like family
PF02223	thymidylate kinase
PF00780	CNH domain
PF03343	SART-1 family
PF04855	SNF5 / SMARCB1 / INI1
PF04005	Hus1-like protein
PF06246	Isy1-like splicing family
PF06549	Protein of unknown function
PF06206	Protein of unknown function
PF04106	Autophagy protein Apg5
PF03439	Supt5 repeat
PF07000	Protein of unknown function
PF07823	Cyclic phosphodiesterase-like protein
PF06963	Ferroportin1
PF08216	Eukaryotic domain of unknown function
PF03060	oxidoreductase, 2-nitropropane dioxygenase family
PF07817	GLE1-like protein
PF05799	Cytochrome c oxidase subunit Vc

PF04922	DIE2/ALG10 family
PF06087	Tyrosyl-DNA phosphodiesterase
PF01197	ribosomal protein L31
PF02104	SURF1 family
PF04780	Protein of unknown function
PF04136	Sec34-like family
PF06644	ATP11 protein
PF00395	S-layer homology domain
PF03754	Domain of unknown function
PF07889	Protein of unknown function

Domains shared by poplar and *Arabidopsis*:

Pfam accession	Description
PF04827	Protein of unknown function
PF02237	Biotin protein ligase C terminal domain
PF02892	BED zinc finger
PF03164	SAND protein
PF07019	Rab5-interacting protein
PF04048	Sec8 exocyst complex component specific domain
PF02040	Arsenical pump membrane protein
PF02357	Transcription termination factor nusG
PF07975	TFIIH C1-like domain
PF06331	REX1 DNA Repair
PF04151	Bacterial pre-peptidase C-terminal domain
PF06735	Protein of unknown function
PF01634	ATP phosphoribosyltransferase
PF04828	Protein of unknown function
PF02095	Extensin-like protein repeat
PF04418	Domain of unknown function
PF04503	Single-stranded DNA binding protein, SSDP
PF03040	CemA protein
PF05251	Uncharacterised protein family
PF05176	ATP10 protein
PF01502	phosphoribosyl-AMP cyclohydrolase
PF01868	Domain of unknown function
PF04180	Low temperature viability protein
PF04128	Partner of SLD five, PSF2
PF05056	Protein of unknown function
PF07708	Tash protein PEST motif
PF01786	Alternative oxidase
PF02602	uroporphyrinogen-III synthase
PF00146	NADH dehydrogenase
PF02943	ferredoxin thioredoxin reductase catalytic beta chain
PF04699	ARP2/3 complex 16 kDa subunit
PF05493	ATP synthase subunit H
PF04114	Gaa1-like, GPI transamidase component
PF03108	MuDR family transposase
PF04050	Up-frameshift suppressor 2
PF04543	Family of unknown function
PF08186	Wound-inducible basic protein family
PF00033	Cytochrome b
PF07297	Dolichol phosphate-mannose biosynthesis regulatory protein
PF04359	Protein of unknown function
PF05254	Uncharacterised protein family
PF05007	Mannosyltransferase
PF03215	Rad17 cell cycle checkpoint protein

PF06859	Bicoid-interacting protein 3
PF00507	NADH-ubiquinone/plastoquinone oxidoreductase, chain 3
PF04172	LrgB-like family
PF02162	XYPPX repeat
PF06155	Protein of unknown function
PF02576	Uncharacterized BCR, YhbC family COG0779
PF05127	Putative ATPase
PF05022	SRP40, C-terminal domain
PF04045	Arp2/3 complex, 34kD subunit p34-Arc
PF05729	NACHT domain

Supplementary Table 3: Ricin/RCA gene family. Gene names are composed of the scaffold number followed by the gene model (m) number. The NCBI lucus tags/gene symbols are also shown. Shadings indicate scaffolds with multiple gene family members. No shading indicates scaffolds with only one ricin-like gene.

Gene model	NCBI locus	Scaffold length	Protein length	% identity with Ricin	% identity with RCA	Remarks
28842.m000952	RCOM_0344270	550 kbp	282	41	41	
29638.m000512	RCOM_0544700	230 kbp	272	36	36	
29638.m000513	RCOM_0544810	230 kbp	271	34	35	
29791.m000533	RCOM_0792550	250 kbp	572	56	54	
29852.m001982	RCOM_0940160	1.8 Mbp	302	35	37	
29942.m000748	RCOM_1110780	700 kbp	301	36	37	
29942.m000749	RCOM_1110790	700 kbp	431	40	40	
29988.m000125	RCOM_1180940	70 kbp	575	81	78	
29988.m000128	RCOM_1180970	70 kbp	203	82	82	
29988.m000129	RCOM_1180980	70 kbp	140	68	75	
30113.m001449	RCOM_1403770	1.3 Mbp	307	34	35	
36244.m000005	RCOM_1960510	1.4 kbp	226	52	60	
54157.m000007	RCOM_2105270	1.1 kbp	120	67	67	
59846.m000009	RCOM_2152660	1 kbp	66	57	53	
59846.m000010	RCOM_2152670	1 kbp	98	52	54	
60626.m00001	RCOM_2159610	2.2 kbp	149	100	89	5' missing due to end of scaffold
60627.m00002	RCOM_2159710	4.5 kbp	188	100	87	5' missing due to end of scaffold
60628.m00003	RCOM_2159810	2 kbp	420	100	89	3' missing due to end of scaffold
60629.m00002	RCOM_2159910	**12 kbp**	**577**	**100**	**89**	**Ricin precursor**
60637.m00004	RCOM_2160110	**35 kbp**	**576**	**89**	**99**	**Agglutinin precursor, putative**
60637.m00006	RCOM_2160120	35 kbp	584	75	73	
60638.m00018	RCOM_2161680	120 kbp	576	84	86	
60638.m00019	RCOM_2160860	120 kbp	438	92	89	
60638.m00022	RCOM_2160650	120 kbp	304	34	35	
60638.m00023	RCOM_2160640	120 kbp	575	83	79	
60638.m00025	RCOM_2160530	120 kbp	347	37	37	
60639.m00003	RCOM_2161880	12 kbp	188	100	87	5' missing due to end of scaffold
60639.m00004	RCOM_2161890	12 kbp	195	94	85	

Supplementary Table 4: Manual annotation of genes involved in the biosynthesis of fatty acids and triacylglycerols. The automatic annotation of 67 (shown in bold) of the 71 selected gene models was manually updated by expert analysis.

Gene model	NCBI locus	Manual annotation
27568.m000266	RCOM_0040840	**ER glycerol-3-phosphate acyltransferase (GPAT)**
27798.m000585	RCOM_0085190	**α-Carboxyltansferase (α-CT) subunit of Het-ACCase,**
27810.m000646	RCOM_0090060	**Lysophosphatidic acid acyltransferase (LPAT); LPAT2 homolog**
27843.m000160	RCOM_0097860	**Enoyl-ACP reductase precursor (EAR), EC 1.3.1.9**
27985.m000877	RCOM_0138550	**Stearoyl-ACP desaturase**
28035.m000362	RCOM_0146820	**Oleate 12-hydroxylase**
28350.m000105	RCOM_0235670	**ER glycerol-phosphate acyltransferase (GPAT)**
28455.m000368	RCOM_0251360	**Ketoacyl-ACP Synthase III (KAS III), EC 2.3.1.41**
28890.m000006	RCOM_0354800	**β-Carboxyltransferase (β-CT) subunit of Het-ACCase**
29489.m000170	RCOM_0445940	**Phospholipase A2 (cytosolic-type)**
29586.m000620	RCOM_0468100	Phosphatidic acid phosphatase, putative
29613.m000358	RCOM_0503360	**Omega-6 desaturase, endoplasmic reticulum**
29630.m000809	RCOM_0525570	**Biotin carboxyl carrier protein subunit of of Het-ACCase (BCCP2)**
29650.m000277	RCOM_0565650	**Enoyl-ACP reductase**
29660.m000759	RCOM_0577200	Phosphatidic acid phosphatase, putative
29660.m000760	RCOM_0577310	**ER Phosphatidate phosphatase (PAP), EC 3.1.3.4**
29660.m000782	RCOM_0577730	**Palmitoyl-acyl carrier protein thioesterase**
29681.m001360	RCOM_0612610	**Omega-3 fatty acid desaturase, endoplasmic reticulum**
29682.m000581	RCOM_0613570	**Type 2 diacylglycerol acyltransferase (DGAT2)**
29693.m002034	RCOM_0633300	**3-Ketoacyl-ACP synthase I (KAS I), EC 2.3.1.41**
29706.m001305	RCOM_0653990	**Phosphatidylcholine:diacylglycerol acyltransferase, EC 2.3.1.158 (PDAT1)**
29726.m003980	RCOM_0679650	**Acyl carrier protein**
29736.m002070	RCOM_0699160	**ER glycerol-3-phosphate acyltransferase (GPAT)**
29739.m003654	RCOM_0708160	Acyl carrier protein, putative
29739.m003711	RCOM_0710230	**3-Ketoacyl-ACP synthase II (KAS II), EC 2.3.1.41**
29747.m001075	RCOM_0724080	Phosphatidic acid phosphatase, putative
29822.m003441	RCOM_0853360	**ER glycerol-3-phosphate acyltransferase (GPAT)**
29827.m002594	RCOM_0866230	**Acyl-CoA-binding protein**
29840.m000629	RCOM_0893800	**Phospholipase A2-1 (secretory-type), EC 3.1.1.4 (PLA21)**
29841.m002744	RCOM_0894910	**Palmitoyl-acyl carrier protein thioesterase (FATB)**
29841.m002865	RCOM_0900600	**Phosphatidylcholine diacylglycerol cholinephosphotransferase (PDCT)**
29842.m003623	RCOM_0905300	**Phospholipase A2 (cytosolic-type)**
29844.m003365	RCOM_0914000	**Acyl-CoA synthetase (ACS2)**
29848.m004677	RCOM_0925410	**Palmitoyl-acyl carrier protein thioesterase (FATB)**
29851.m002448	RCOM_0938140	**Lysophosphatidic acid acyltransferase (LPAT); LPAT5 homolog**
29851.m002473	RCOM_0938700	**Acyl CoA synthetase**
29889.m003411	RCOM_1004000	**Soluble diacylglycerol acyltransferase/wax synthase**
29908.m005991	RCOM_1033990	**Homomeric acetyl-CoA carboxylase (Hom-ACCase), EC 6.4.1.2**
29912.m005286	RCOM_1046030	**Phospholipid:diacylglycerol acyltransferase 1 (PDAT1)**
29912.m005373	RCOM_1047540	**Type 1 diacylglycerol acyltransferase (DGAT1)**
29912.m005406	RCOM_1048160	**Phospholipase A2-3 (secretory-type), EC 3.1.1.4 (PLA23)**
29917.m001992	RCOM_1064090	**Oleosin1**
29929.m004514	RCOM_1076810	**Stearoyl-ACP desaturase**
29929.m004515	RCOM_1076820	**Stearoyl-ACP desaturase**
29929.m004560	RCOM_1077760	**Biotin carboxyl carrier protein subunit of of Het-ACCase (BCCP1)**
29929.m004732	RCOM_1081890	**3-Ketoacyl-ACP Reductase (KAR), EC 1.1.1.100**
29969.m000267	RCOM_1151840	**ER glycerol-3-phosphate acyltransferase (GPAT)**
29991.m000626	RCOM_1185660	**Phosphatidylcholine : Diacylglycerol Acyltransferase 2, EC 2.3.1.158 (PDAT2)**
30020.m000203	RCOM_1238330	**Stearoyl-ACP desaturase**
30068.m002515	RCOM_1311820	**3-Ketoacyl-ACP synthase I (KAS I), EC 2.3.1.41**

30068.m002660	RCOM_1316770	**Plastidial glycerol-phosphate acyltransferase (GPAT)**
30076.m004616	RCOM_1346070	**Acyl-CoA synthetase**
30113.m001448	RCOM_1403260	**Malonyl-CoA : ACP acyltransferase (MCAAT), EC 2.3.1.39**
30122.m000357	RCOM_1419880	**ER glycerol-3-phosphate acyltransferase 9 (GPAT9)**
30128.m008670	RCOM_1429400	**Acyl-carrier protein (ACP)**
30128.m008777	RCOM_1431520	**Acyl-CoA synthetase (ACS4)**
30138.m003845	RCOM_1464650	**Diacylglycerol cholinephosphotransferase (CPT), EC 2.7.8.2**
30142.m000631	RCOM_1478210	**Phospholipase A22, EC 3.1.1.4 (PLA22)**
30147.m013777	RCOM_1506940	**3-Ketoacyl-ACP reductase**
30147.m014333	RCOM_1502140	**Oleosin2**
30147.m014425	RCOM_1504200	**Acyl carrier protein**
30147.m014468	RCOM_1505510	**Palmitoyl-acyl carrier protein thioesterase (FATB)**
30169.m006432	RCOM_1577620	**Lysophosphatidic acid acyltransferase (LPAT); LPAT2 homolog**
30169.m006433	RCOM_1577630	**Lysophosphatidic acid acyltransferase (LPAT); LPAT2 homolog**
30170.m013990	RCOM_1593580	**Lysophosphatidic acid acyltransferase (LPAT); LPAT4 homolog**
30170.m014002	RCOM_1593790	**Lysophophatidylcholine acyltransferase (LPCAT)-like**
30174.m008615	RCOM_1615780	**ER glycerol-3-phosphate acyltransferase (GPAT)**
30185.m000954	RCOM_1657380	**Biotin carboxylase subunit (BC) of Het-ACCase**
30190.m010831	RCOM_1691890	**Acyl-CoA synthetase (ACS1)**
30200.m000354	RCOM_1712710	**Hydroxyacyl-ACP Dehydrase (HAD), EC 4.2.1**
30217.m000262	RCOM_1750180	**Acyl-ACP Thioesterase A (FATA), EC 3.1.2.14**

Supplementary Table 5: Putative castor bean resistance genes. Gene models annotated as disease resistance genes were compiled. Using our automatic gene naming or BLAST matches to known disease resistance proteins, putative disease resistance gene models were grouped in three classes: NBS-LRR (nucleotide binding-leucine-rich repeat), eLRR (extracellular LRR), and dirigent-like. Shadings indicate scaffolds with multiple disease resistance genes. No shading indicates scaffolds with only one disease resistance gene.

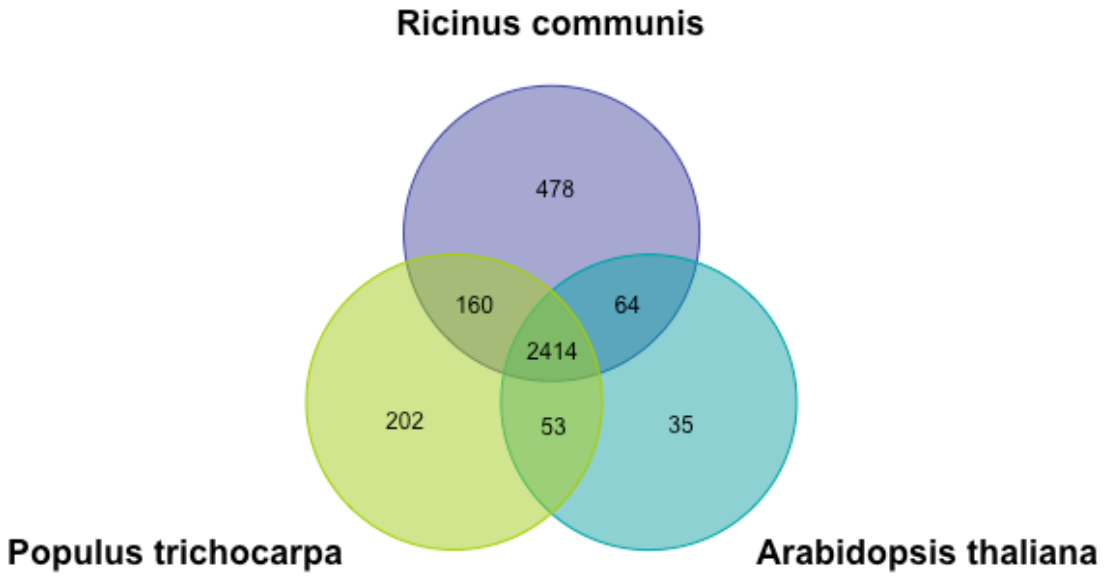
Gene Model	NCBI locus	Automatic annotation	Gene class
27436.m000285	RCOM_0009540	Leucine-rich repeat-containing protein, putative	NBS-LRR
27467.m000171	RCOM_0016080	Leucine-rich repeat-containing protein, putative	NBS-LRR
27504.m000624	RCOM_0024370	Disease resistance protein RPM1, putative	NBS-LRR
27504.m000625	RCOM_0024480	Disease resistance protein RPM1, putative	NBS-LRR
27574.m000228	RCOM_0042270	Disease resistance response protein, putative	dirigent-like
27904.m000211	RCOM_0108660	Disease resistance protein RPM1, putative	NBS-LRR
27956.m000355	RCOM_0127740	Leucine-rich repeat receptor protein kinase EXS precursor, putative	eLRR
28226.m000870	RCOM_0205410	Leucine-rich repeat transmembrane protein kinase, putative	eLRR
28525.m000274	RCOM_0268610	TMV resistance protein N, putative	NBS-LRR
28589.m000049	RCOM_0280040	Disease resistance protein RPS2, putative	NBS-LRR
28592.m000278	RCOM_0280590	Leucine-rich repeat-containing protein, putative	NBS-LRR
28623.m000397	RCOM_0292060	Leucine-rich repeat receptor protein kinase EXS precursor, putative	eLRR
28966.m000543	RCOM_0364920	Leucine-rich repeat-containing protein, putative	NBS-LRR
29168.m000372	RCOM_0393490	Disease resistance protein RPM1, putative	NBS-LRR
29222.m000397	RCOM_0407000	Disease resistance protein RGA2, putative	NBS-LRR
29222.m000399	RCOM_0407120	Leucine-rich repeat-containing protein, putative	NBS-LRR
29577.m000461	RCOM_0455430	Disease resistance protein RPH8A, putative	NBS-LRR
29579.m000196	RCOM_0458060	Leucine-rich repeat-containing protein, putative	NBS-LRR
29585.m000582	RCOM_0464860	Disease resistance protein RPP8, putative	NBS-LRR
29609.m000594	RCOM_0496700	Leucine-rich repeat-containing protein, putative	NBS-LRR
29634.m002084	RCOM_0534550	Disease resistance protein RGA2, putative	NBS-LRR
29666.m001443	RCOM_0585380	Late blight resistance protein R1-A, putative	NBS-LRR
29666.m001444	RCOM_0585390	Disease resistance protein RPH8A, putative	NBS-LRR
29666.m001447	RCOM_0585520	Disease resistance protein RGA2, putative	NBS-LRR
29666.m001448	RCOM_0585530	Disease resistance protein RPS2, putative	NBS-LRR
29676.m001639	RCOM_0602290	Disease resistance protein ADR1, putative	NBS-LRR
29676.m001640	RCOM_0602300	Leucine-rich repeat-containing protein, putative	NBS-LRR
29690.m000408	RCOM_0625620	Leucine-rich repeat-containing protein, putative	NBS-LRR
29693.m002050	RCOM_0633880	Leucine-rich repeat transmembrane protein kinase, putative	eLRR
29702.m000165	RCOM_0645130	Leucine-rich repeat transmembrane protein kinase, putative	eLRR
29716.m000299	RCOM_0662960	Disease resistance protein RPM1, putative	NBS-LRR
29729.m002285	RCOM_0687360	Leucine-rich repeat-containing protein 2, lrrc2, putative	NBS-LRR
29736.m002080	RCOM_0699260	Leucine-rich repeat-containing protein, putative	NBS-LRR
29736.m002081	RCOM_0699270	Leucine-rich repeat-containing protein, putative	NBS-LRR
29756.m000505	RCOM_0739250	Disease resistance response protein, putative	dirigent-like
29756.m000506	RCOM_0739360	Disease resistance response protein, putative	dirigent-like
29757.m000712	RCOM_0740720	Leucine-rich repeat-containing protein, putative	NBS-LRR
29757.m000737	RCOM_0742270	Disease resistance protein RPM1, putative	NBS-LRR
29757.m000746	RCOM_0742650	Leucine-rich repeat-containing protein, putative	NBS-LRR
29761.m000426	RCOM_0751360	Leucine-rich repeat-containing protein, putative	NBS-LRR

29773.m000283	RCOM_0768900	Disease resistance protein RPS2, putative	NBS-LRR
29801.m003130	RCOM_0812210	Leucine-rich repeat-containing protein, putative	NBS-LRR
29801.m003132	RCOM_0812230	Disease resistance protein RPS2, putative	NBS-LRR
29801.m003134	RCOM_0812250	Leucine-rich repeat-containing protein, putative	NBS-LRR
29805.m001536	RCOM_0820400	Leucine-rich repeat-containing protein, putative	NBS-LRR
29830.m001440	RCOM_0873960	Disease resistance response protein, putative	dirigent-like
29838.m001641	RCOM_0884220	Leucine-rich repeat-containing protein, putative	NBS-LRR
29838.m001666	RCOM_0885080	Disease resistance protein RPS5, putative	NBS-LRR
29841.m002750	RCOM_0895260	Disease resistance protein RPS5, putative	NBS-LRR
29841.m002793	RCOM_0897880	Disease resistance protein RFL1, putative	NBS-LRR
29841.m002815	RCOM_0898290	Disease resistance protein RFL1, putative	NBS-LRR
29841.m002816	RCOM_0898300	Disease resistance protein RPS5, putative	NBS-LRR
29841.m002829	RCOM_0898840	Leucine-rich repeat-containing protein, putative	NBS-LRR
29841.m002830	RCOM_0898850	Disease resistance protein RPS2, putative	NBS-LRR
29841.m002832	RCOM_0898870	Disease resistance protein RFL1, putative	NBS-LRR
29841.m002872	RCOM_0900670	Disease resistance protein RGA2, putative	NBS-LRR
29841.m002919	RCOM_0902160	TMV resistance protein N, putative	NBS-LRR
29872.m000523	RCOM_0979720	Disease resistance protein RFL1, putative	NBS-LRR
29872.m000524	RCOM_0979730	Disease resistance protein RFL1, putative	NBS-LRR
29889.m003298	RCOM_1000560	Disease resistance response protein, putative	dirigent-like
29904.m002928	RCOM_1019160	Leucine-rich repeat-containing protein, putative	NBS-LRR
29910.m000950	RCOM_1045150	TMV resistance protein N, putative	NBS-LRR
29910.m000957	RCOM_1045320	TMV resistance protein N, putative	NBS-LRR
29912.m005378	RCOM_1047690	Leucine-rich repeat-containing protein 2, lrrc2, putative	NBS-LRR
29914.m000195	RCOM_1055490	Disease resistance protein RPM1, putative	NBS-LRR
29929.m004539	RCOM_1077550	TMV resistance protein N, putative	NBS-LRR
29937.m000198	RCOM_1098990	Leucine-rich repeat-containing protein, putative	NBS-LRR
29948.m000709	RCOM_1118640	Disease resistance protein RGA2, putative	NBS-LRR
29950.m001163	RCOM_1122050	Leucine-rich repeat-containing protein, putative	NBS-LRR
29990.m000507	RCOM_1184740	Leucine-rich repeat-containing protein, putative	NBS-LRR
29990.m000508	RCOM_1184850	Leucine-rich repeat-containing protein, putative	NBS-LRR
29994.m000440	RCOM_1195490	Disease resistance protein RPH8A, putative	NBS-LRR
30055.m001596	RCOM_1282160	Leucine rich repeat containing protein kinase, putative	eLRR
30061.m000284	RCOM_1292200	Disease resistance protein RGA2, putative	NBS-LRR
30063.m001400	RCOM_1296800	Leucine-rich repeat-containing protein, putative	NBS-LRR
30063.m001411	RCOM_1298340	Disease resistance protein RPS2, putative	NBS-LRR
30063.m001415	RCOM_1298580	Leucine-rich repeat-containing protein, putative	NBS-LRR
30072.m000956	RCOM_1321720	Leucine-rich repeat protein, putative	eLRR
30074.m001350	RCOM_1329890	TMV resistance protein N, putative	NBS-LRR
30074.m001357	RCOM_1330160	TMV resistance protein N, putative	NBS-LRR
30074.m001359	RCOM_1330180	Leucine-rich repeat-containing protein, putative	NBS-LRR
30074.m001378	RCOM_1331470	Disease resistance protein RPP13, putative	NBS-LRR
30074.m001381	RCOM_1331600	Disease resistance protein RFL1, putative	NBS-LRR
30074.m001382	RCOM_1331610	Disease resistance protein RPS2, putative	NBS-LRR
30074.m001398	RCOM_1333060	Leucine-rich repeat-containing protein 2, lrrc2, putative	NBS-LRR
30074.m001399	RCOM_1333070	Disease resistance protein RPM1, putative	NBS-LRR
30110.m000724	RCOM_1397630	Leucine-rich repeat-containing protein, putative	NBS-LRR
30128.m008658	RCOM_1429080	Leucine-rich repeat-containing protein, putative	NBS-LRR
30131.m007205	RCOM_1452690	Leucine-rich repeat-containing protein, putative	NBS-LRR
30131.m007205	RCOM_1452690	Disease resistance protein RPM1, putative	NBS-LRR

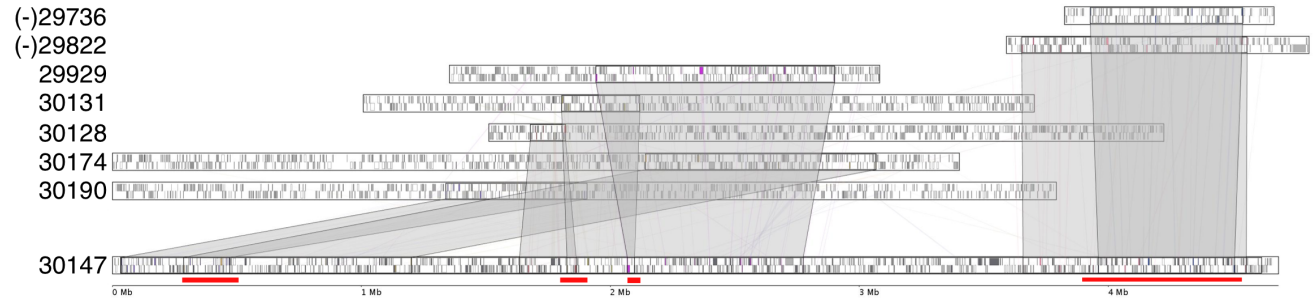
30131.m007209	RCOM_1452730	Disease resistance protein RGA2, putative	NBS-LRR
30131.m007215	RCOM_1452990	Leucine-rich repeat-containing protein, putative	NBS-LRR
30131.m007232	RCOM_1453360	Disease resistance protein RPM1, putative	NBS-LRR
30131.m007235	RCOM_1453390	Disease resistance protein RGA2, putative	NBS-LRR
30131.m007237	RCOM_1453510	Disease resistance protein RPM1, putative	NBS-LRR
30131.m007242	RCOM_1453660	Disease resistance response protein, putative	dirigent-like
30133.m000230	RCOM_1455020	Disease resistance response protein, putative	dirigent-like
30138.m003923	RCOM_1467510	Disease resistance response protein, putative	dirigent-like
30143.m001186	RCOM_1481570	Disease resistance protein RPP8, putative	NBS-LRR
30143.m001201	RCOM_1482430	Disease resistance protein RPP8, putative	NBS-LRR
30146.m003486	RCOM_1487620	Leucine-rich repeat-containing protein, putative	NBS-LRR
30147.m013873	RCOM_1508820	Leucine-rich repeat-containing protein, putative	eLRR
30147.m013889	RCOM_1509380	Disease resistance protein RPP13, putative	NBS-LRR
30147.m014532	RCOM_1506660	Leucine-rich repeat protein, putative	eLRR
30148.m001423	RCOM_1516000	Disease resistance protein RGA2, putative	NBS-LRR
30148.m001424	RCOM_1516110	Leucine-rich repeat-containing protein, putative	NBS-LRR
30169.m006484	RCOM_1579060	Leucine-rich repeat-containing protein, putative	NBS-LRR
30170.m013888	RCOM_1590790	Disease resistance protein RPP13, putative	NBS-LRR
30170.m013933	RCOM_1591970	Leucine-rich repeat-containing protein, putative	NBS-LRR
30190.m010961	RCOM_1678470	Leucine-rich repeat protein, putative	eLRR
30190.m011021	RCOM_1680260	Leucine rich repeat receptor kinase, putative	eLRR
30190.m011025	RCOM_1680500	Leucine rich repeat receptor kinase, putative	eLRR
30190.m011051	RCOM_1681360	Leucine-rich repeat-containing protein, putative	NBS-LRR
30190.m011052	RCOM_1681370	Leucine-rich repeat-containing protein, putative	NBS-LRR
30190.m011060	RCOM_1681450	Leucine-rich repeat transmembrane protein kinase, putative	eLRR
30190.m011348	RCOM_1689510	Disease resistance response protein, putative	dirigent-like
30190.m011349	RCOM_1689520	Disease resistance response protein, putative	dirigent-like
30205.m001595	RCOM_1727110	Leucine-rich repeat-containing protein, putative	NBS-LRR
30226.m002005	RCOM_1772130	Disease resistance protein RPP8, putative	NBS-LRR
30710.m000036	RCOM_1797940	Disease resistance protein RPH8A, putative	NBS-LRR
36675.m000011	RCOM_1965050	Disease resistance response protein, putative	dirigent-like

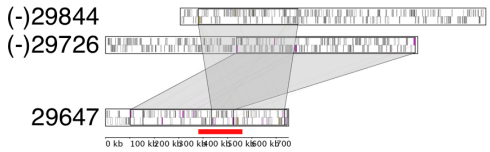
Supplementary figures

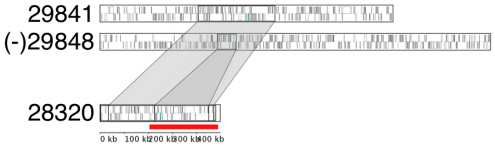
Supplementary Figure 1: Venn diagram showing Specific and shared protein (Pfam) domains between castor bean, poplar, and *Arabidopsis*. The castor bean predicted proteome could be matched to over 3,000 protein domains from Pfam, several of which are not present in *Arabidopsis* or poplar, including secondary metabolism genes. A list of shared and specific domains is shown in Supplementary Table 2.

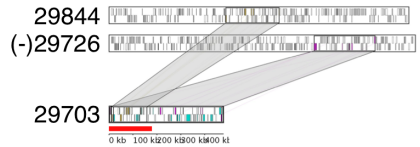


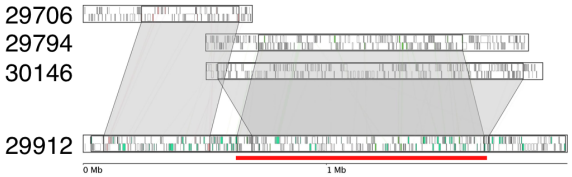
Supplementary Figure 2: Seventeen paralogous triplicated regions identified in the castor bean genome. Images generated using Sybil show regions that contain strings of collinear genes in three different scaffolds. The approximate triplciated regions are shown with a red bar under the alignments. The minus sign in parenthesis indicates reverse orientation of the corresponding scaffold.

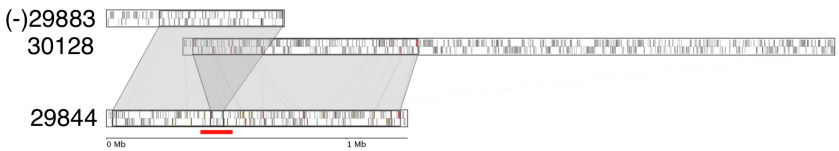






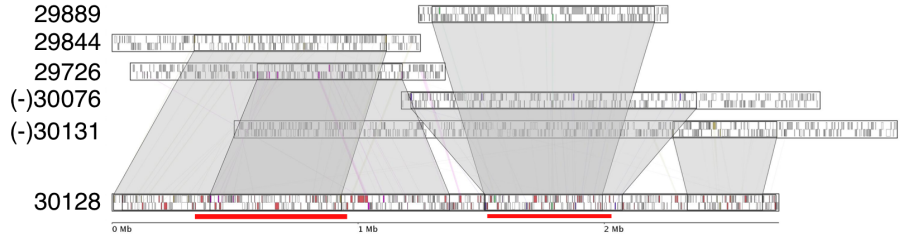


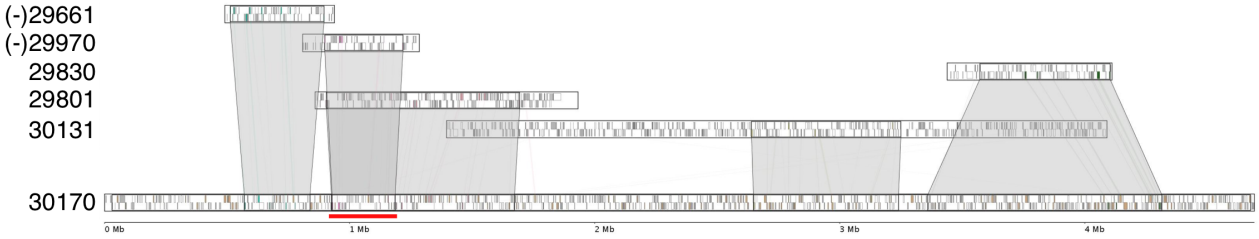


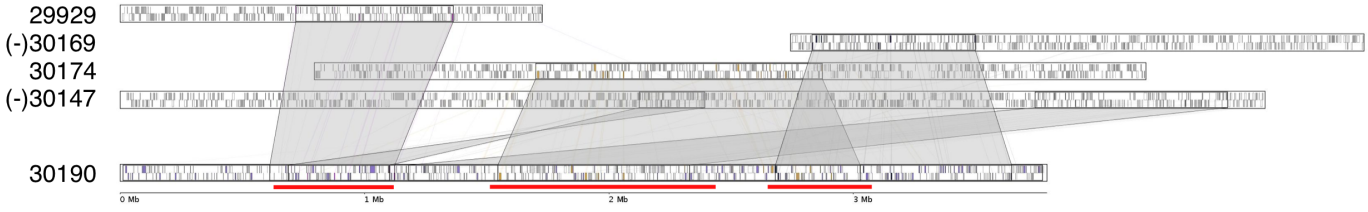



(-)30115
30170
29801
0 Mb
1 Mb


(-)29676
(-)30170
29851
0 Mb
1Mb

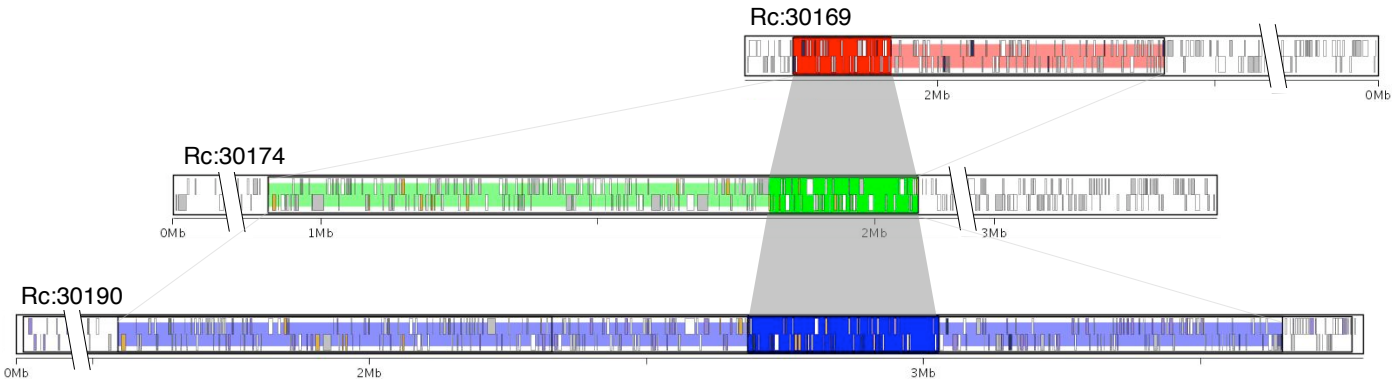

29889
29844
29726
(-)30076
(-)30131
30128
0 Mb
1 Mb
2 Mb

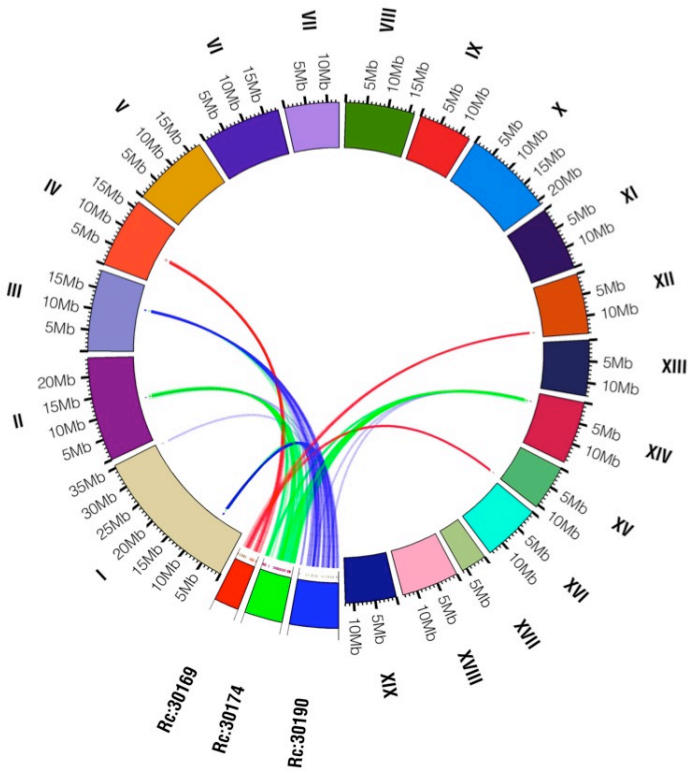

(-)29661
(-)29970
29830
29801
30131
30170
0 Mb
1 Mb
2 Mb
3 Mb
4 Mb

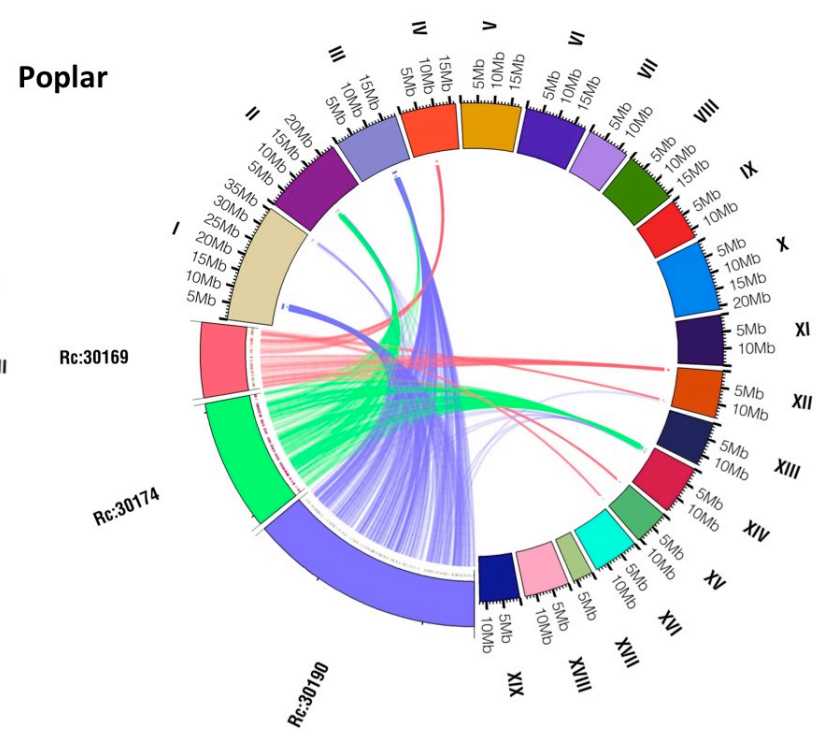

29929
(-)30169
30174
(-)30147
30190
0 Mb
1 Mb
2 Mb
3 Mb

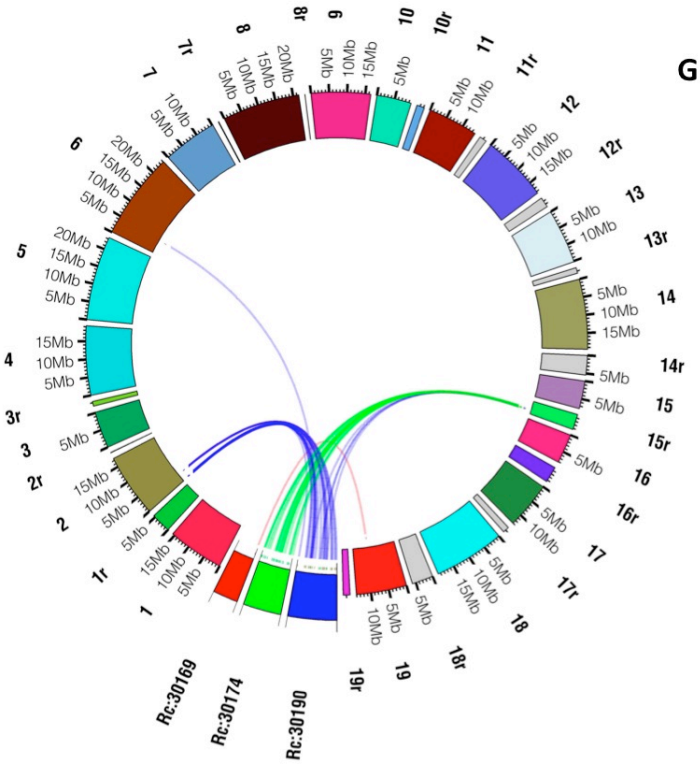
Supplementary Figure 3: Collinearity between paralogous triplicated castor bean genomic regions and their putative orthologues in other dicot genomes. In each section, a castor bean triplicated region is shown along with its corresponding putative orthologous regions in poplar, grapevine, *Arabidopsis*, and papaya (see Fig. 2 in main text for details). Different from Figure 2, in these cases, the extension of the castor bean strings of paralogous genes is different in each of the three scaffolds (see Supplementary Fig. 2). In order to be conservative, only the intersection between the paralogous regions is showed as a shaded area. In addition, the projection of the larger paralogous region onto the shorter ones is shown by gray lines (top panels). Two versions of the Circos images of cross-species comparisons are shown (left: considering only the intersection of the three castor bean paralogous regions; right: considering the projections of the longer paralogous regions).

A

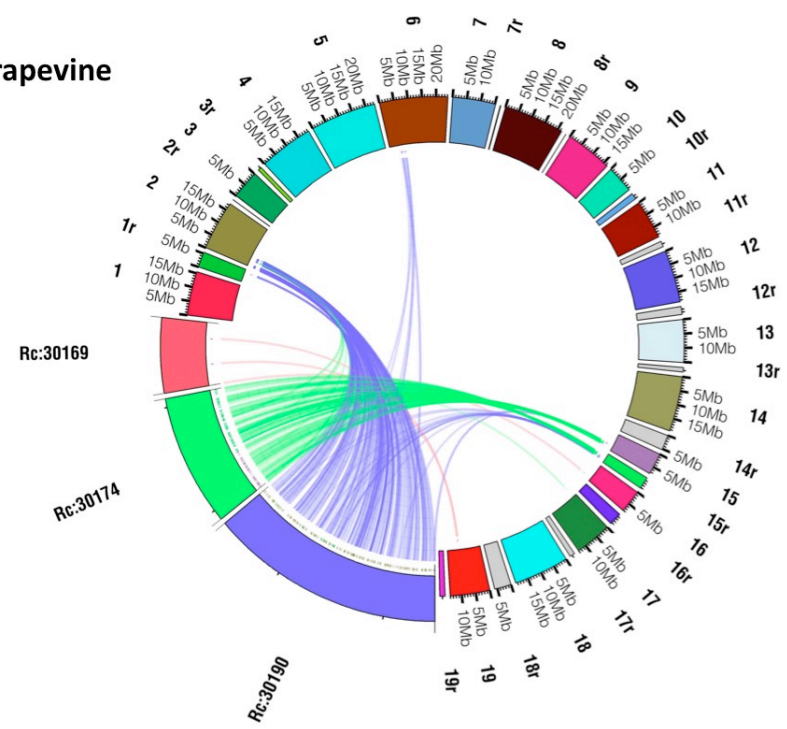


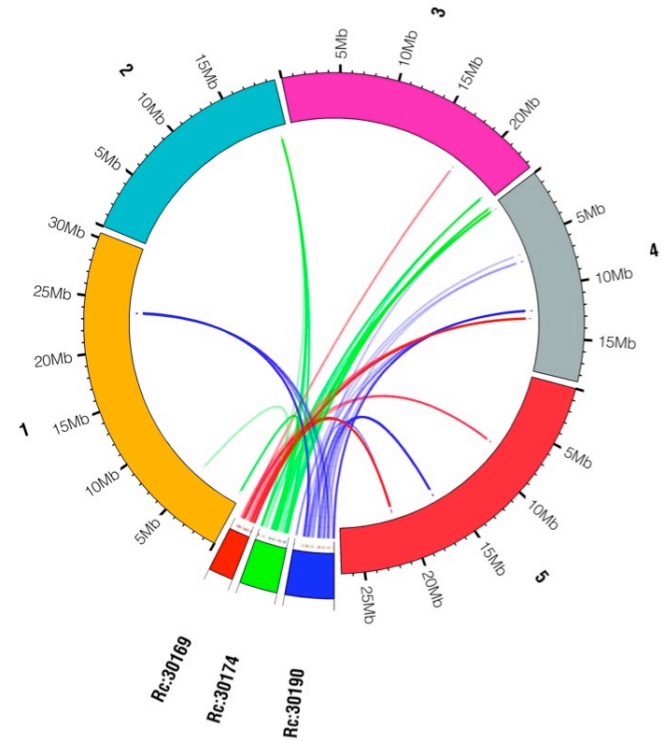




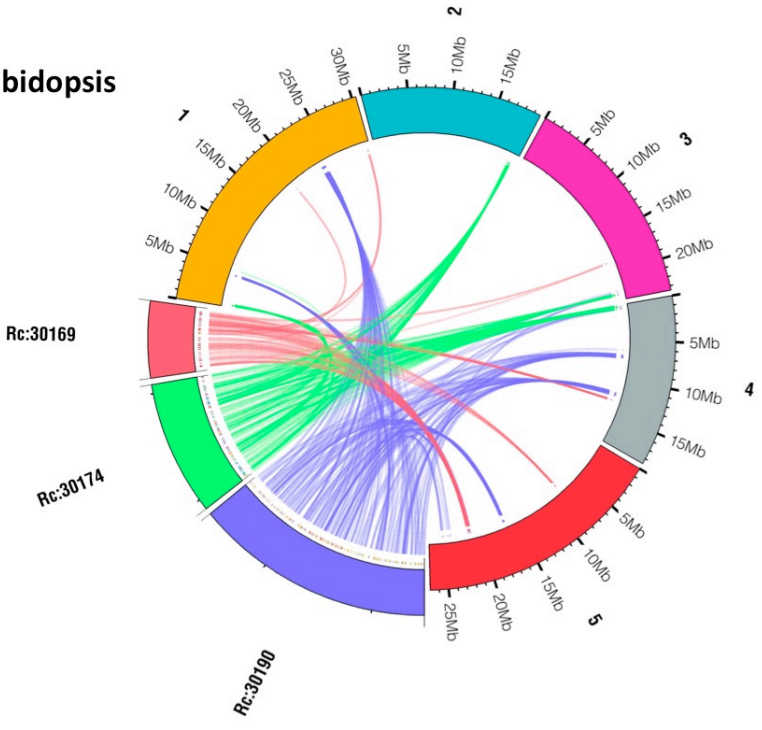


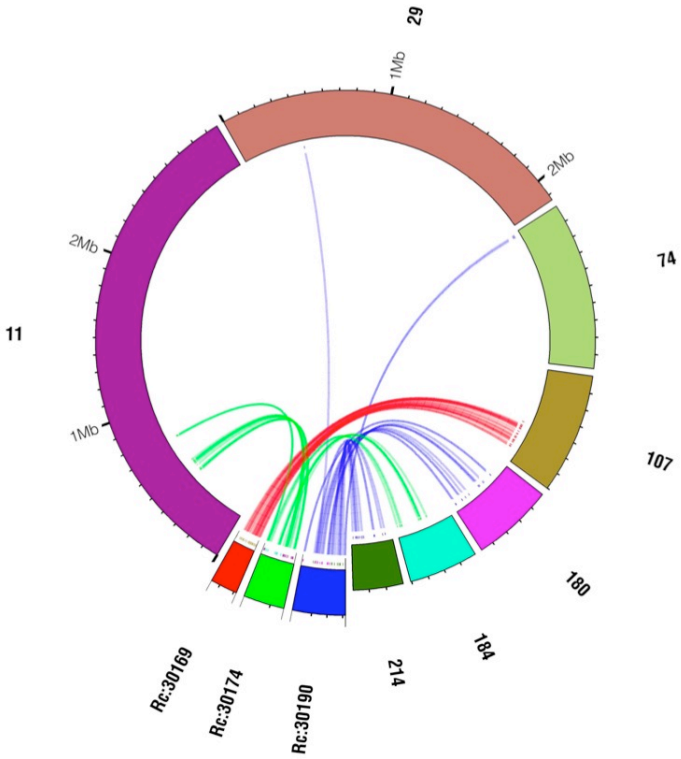
Grapevine



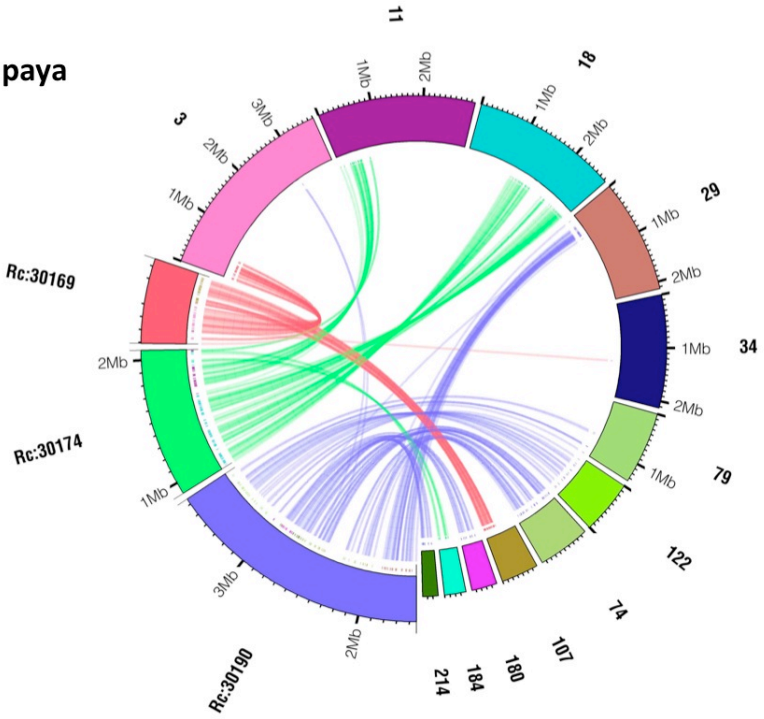


Arabidopsis

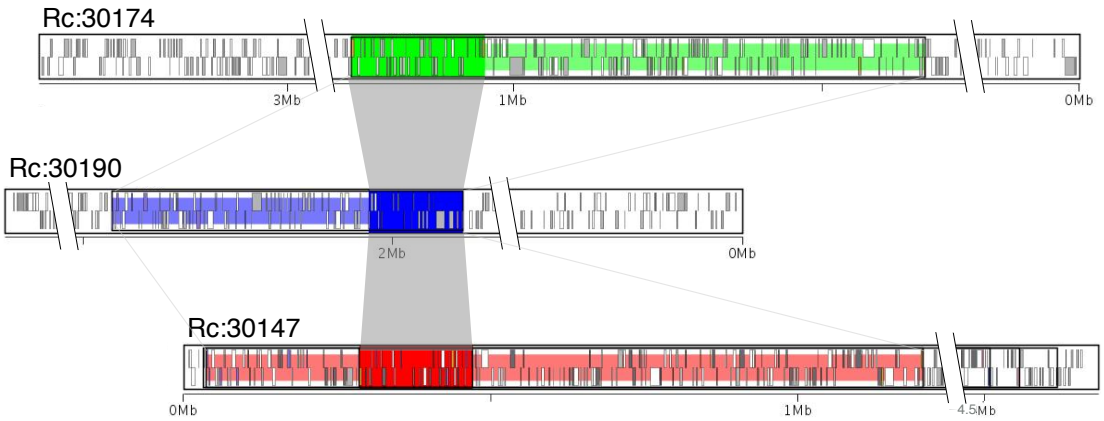


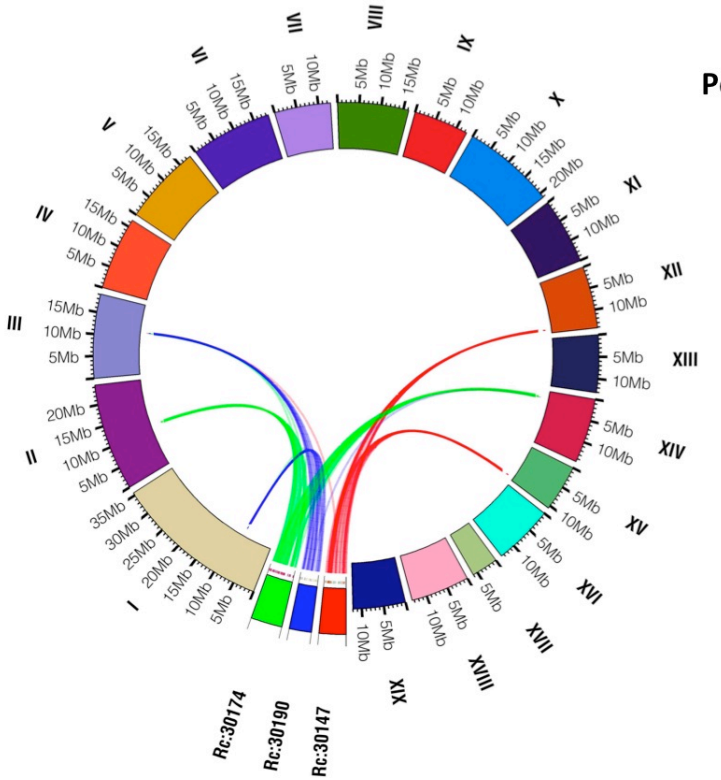

29
1Mb
2Mb
74
11
2Mb
107
1Mb
180
Rc:30169
Rc:30174
Rc:30190
214
184

Papaya

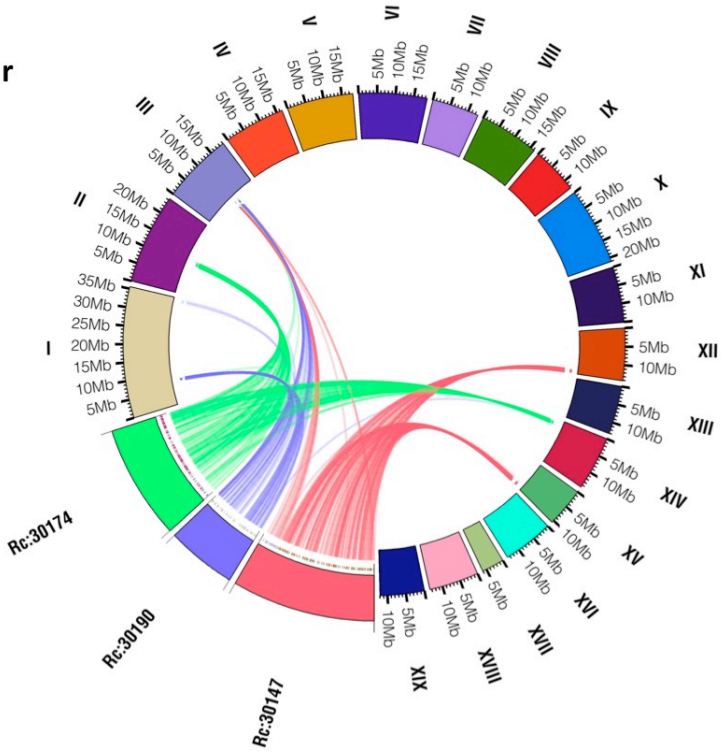

11
1Mb
2Mb
18
3
3Mb
2Mb
1Mb
29
Rc:30169
2Mb
34
Rc:30174
1Mb
79
3Mb
122
2Mb
74
Rc:30190
107
214
184
180

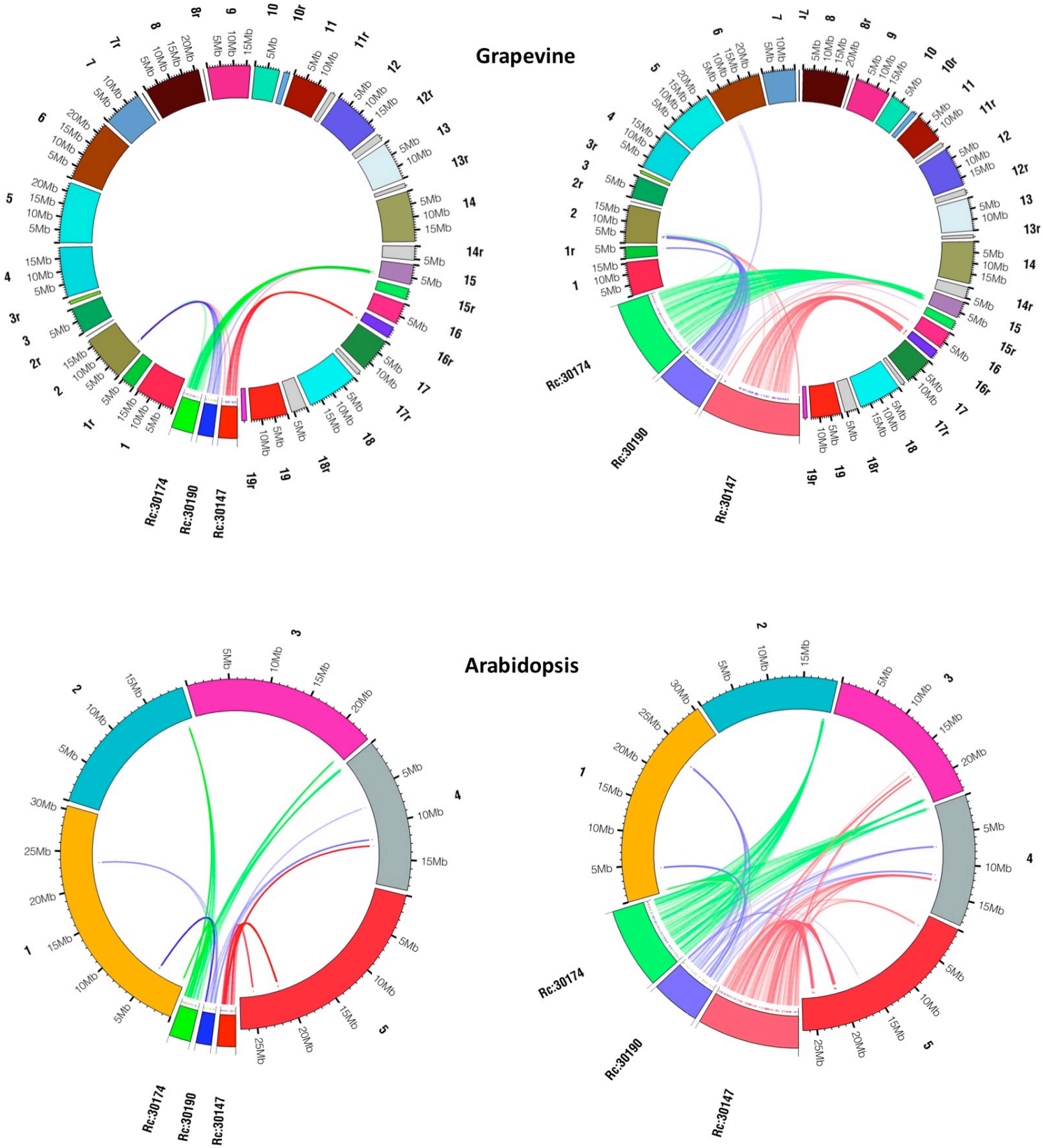
B

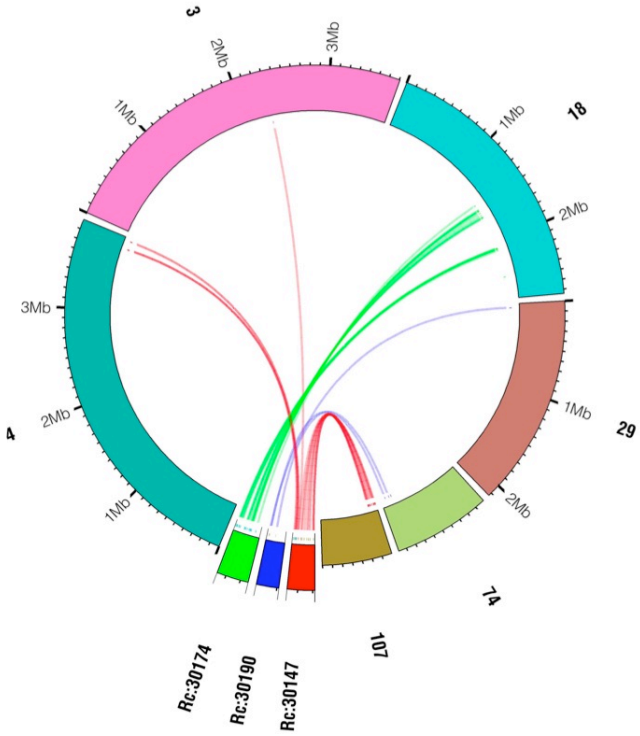

Rc:30174
3Mb
1Mb
0Mb
Rc:30190
2Mb
0Mb
Rc:30147
0Mb
1Mb
4.5Mb


Rc:30174
Rc:30190
Rc:30147

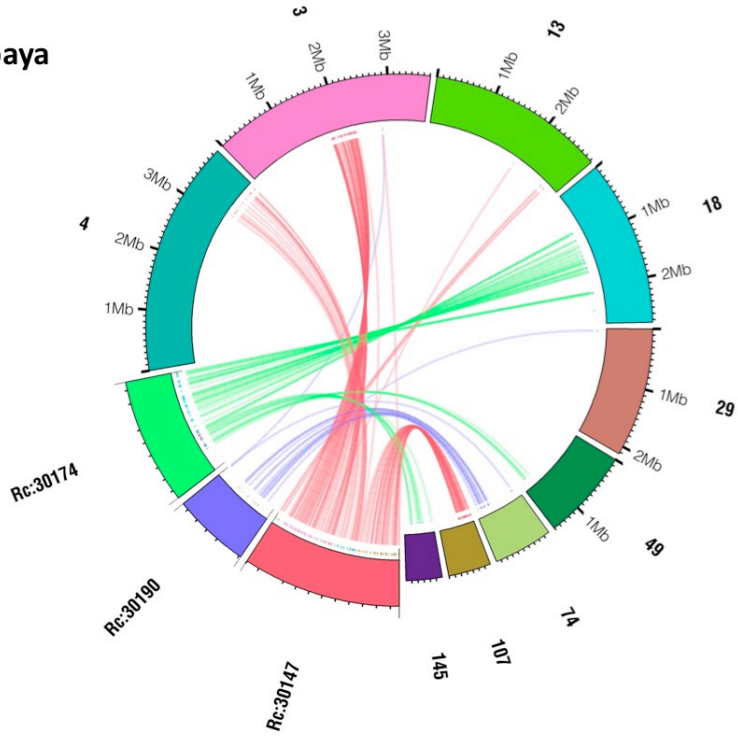
Poplar


Rc:30174
Rc:30190
Rc:30147

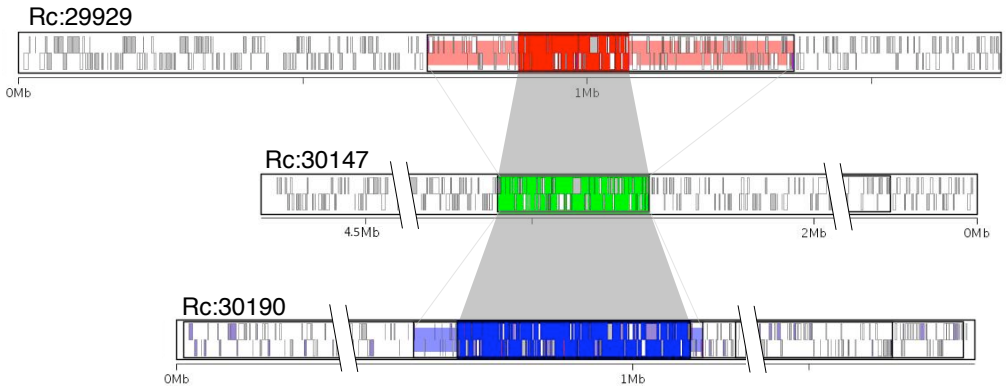

Grapevine
Arabidopsis
Rc:30174
Rc:30190
Rc:30147

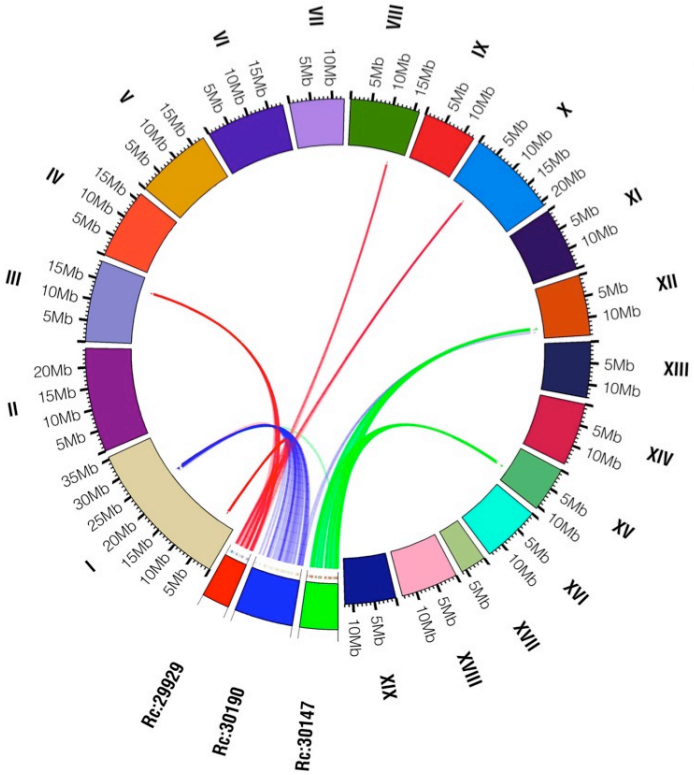

3
2Mb
3Mb
1Mb
18
2Mb
1Mb
29
2Mb
74
107
Rc:30174
Rc:30190
Rc:30147
4
2Mb
3Mb
1Mb

Papaya

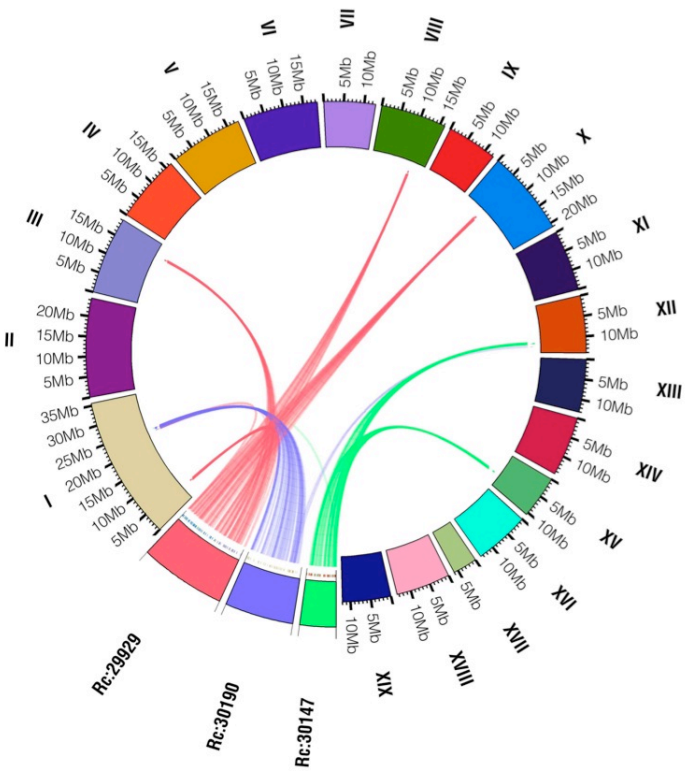

3
2Mb
3Mb
1Mb
13
1Mb
2Mb
18
1Mb
2Mb
29
1Mb
2Mb
49
1Mb
74
107
145
Rc:30147
Rc:30190
Rc:30174
4
2Mb
3Mb
1Mb

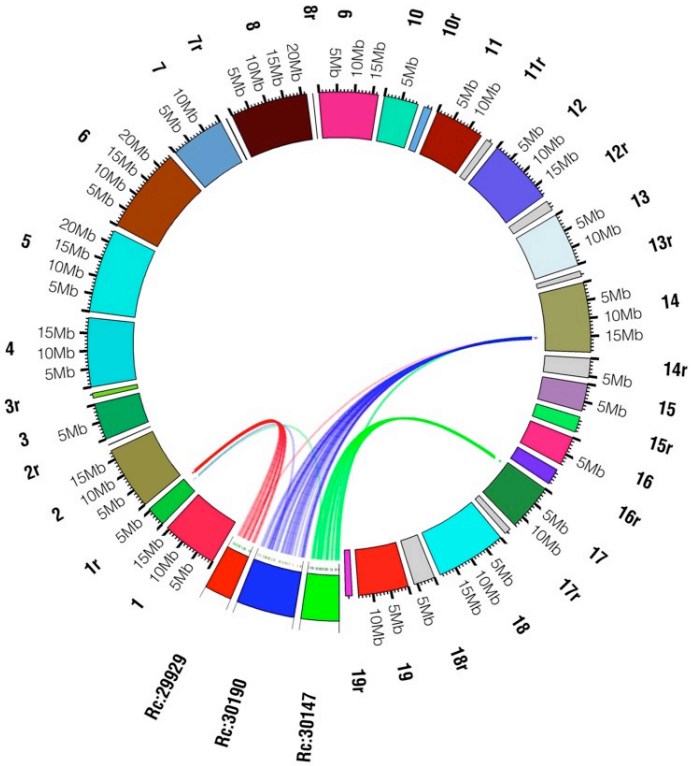
C


Rc:29929
0Mb
1Mb
Rc:30147
4.5Mb
2Mb
0Mb
Rc:30190
0Mb
1Mb

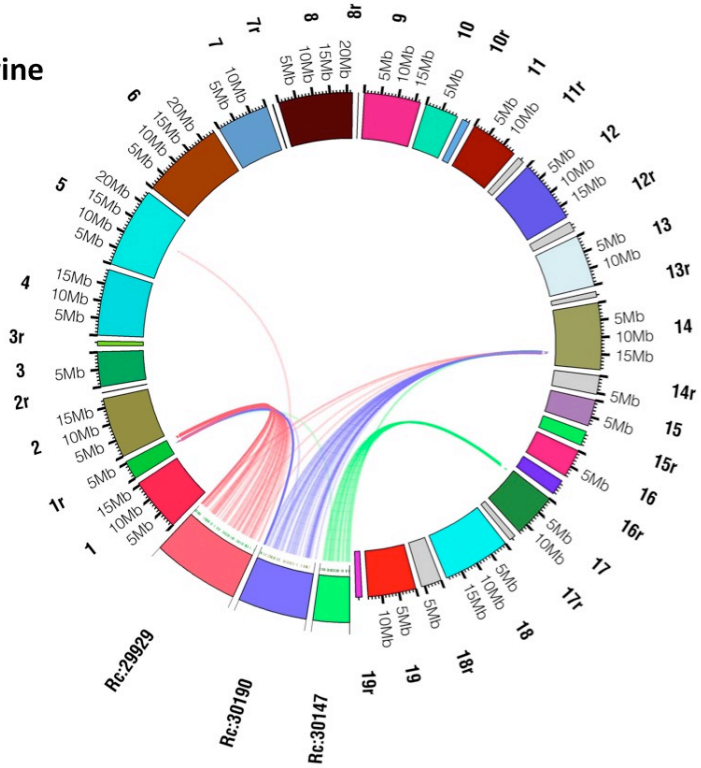

Rc:29929
Rc:30190
Rc:30147

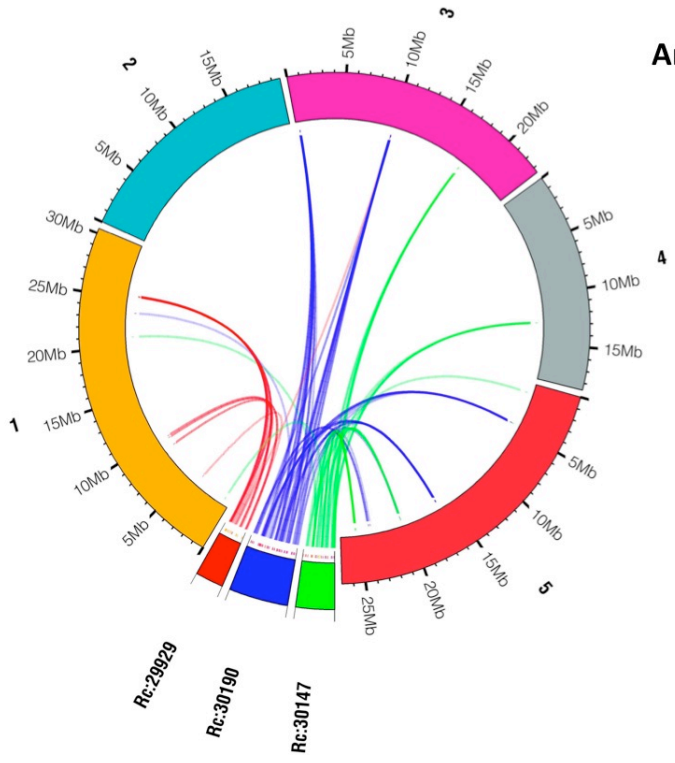
Poplar


Rc:29929
Rc:30190
Rc:30147

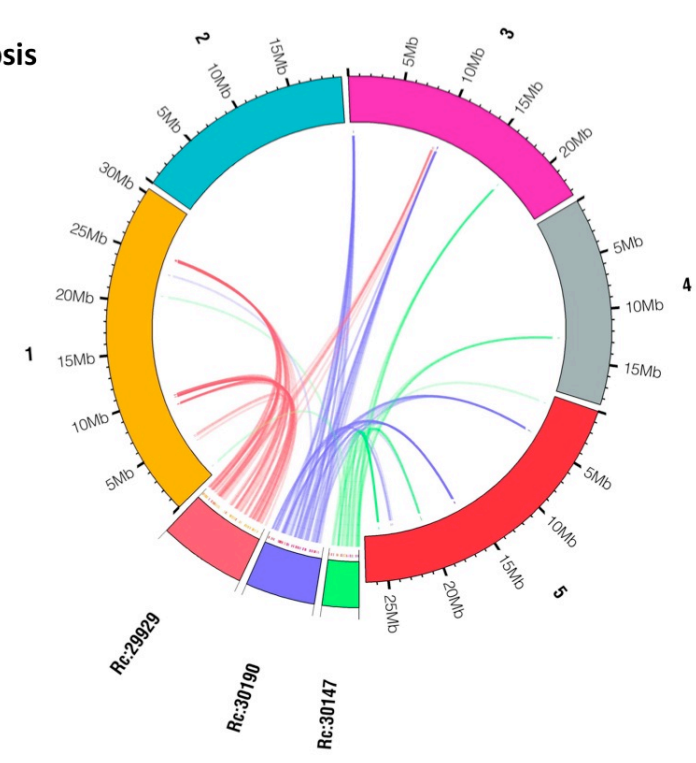


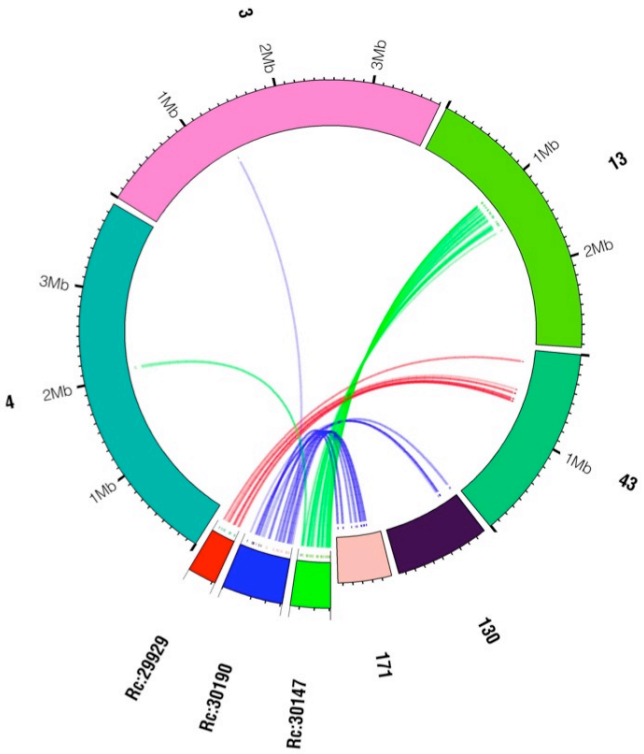
Grapevine



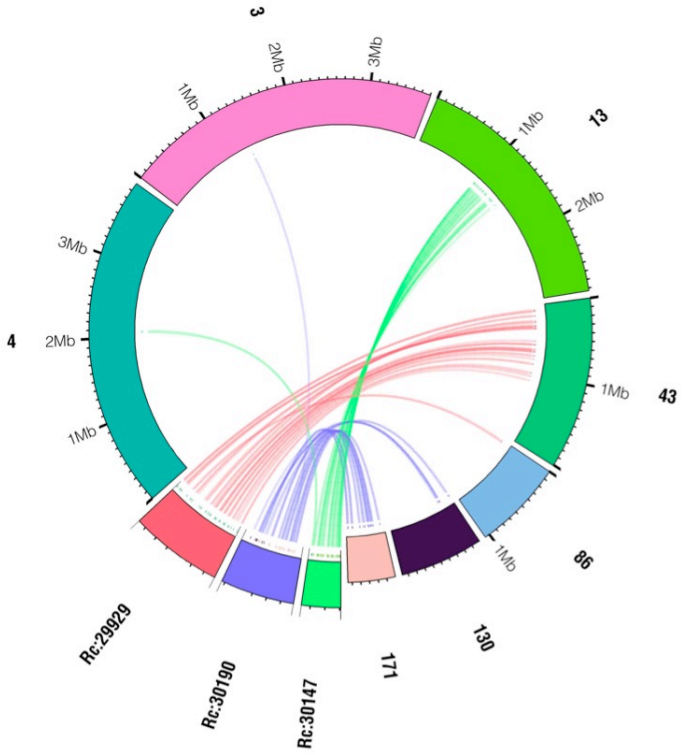


Arabidopsis

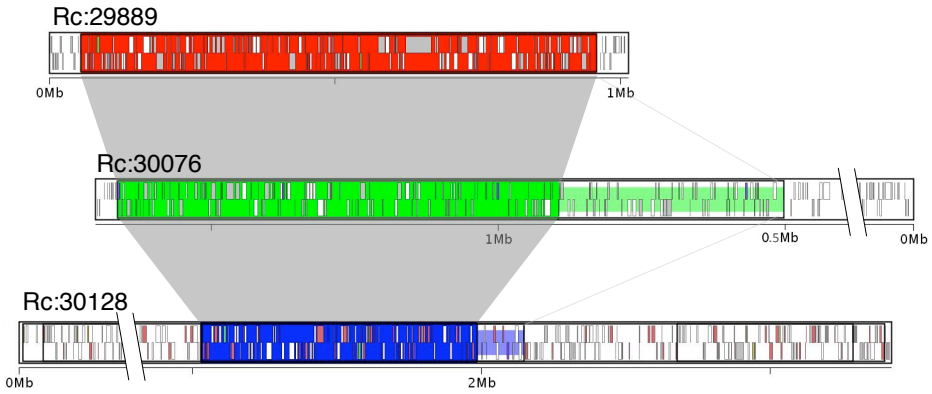


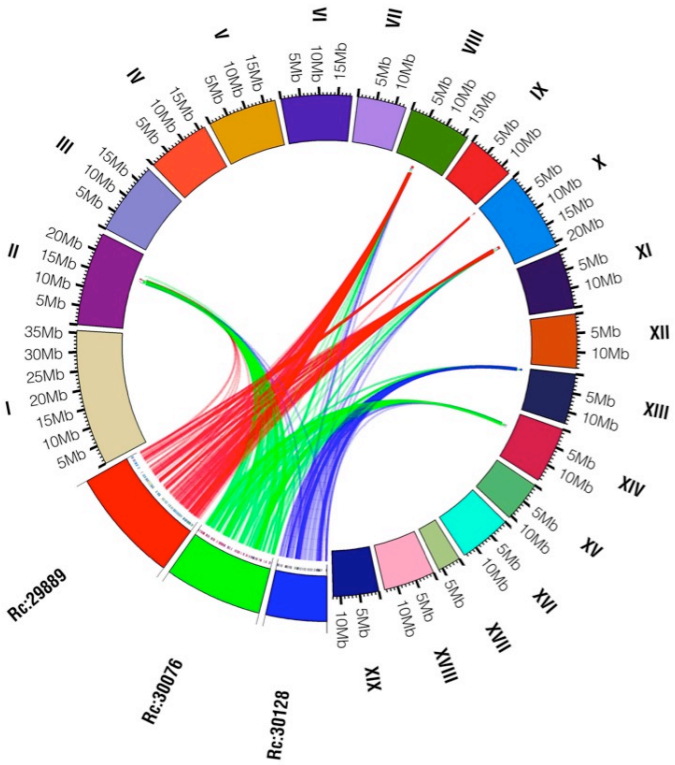

3
2Mb
3Mb
1Mb
13
1Mb
2Mb
43
1Mb
130
171
Rc:30147
Rc:30190
Rc:29929
1Mb
4
2Mb
3Mb

Papaya

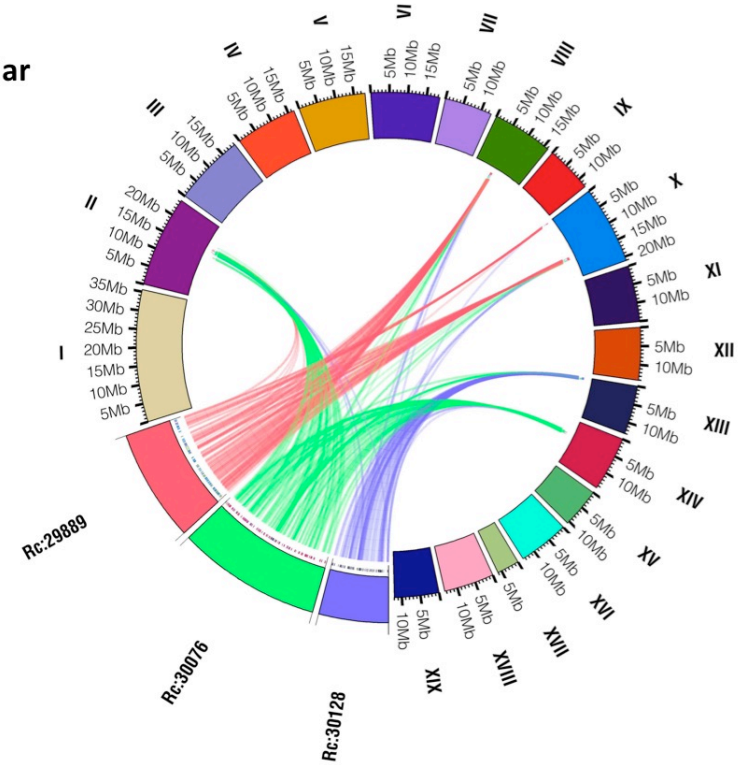

3
2Mb
3Mb
1Mb
13
1Mb
2Mb
43
1Mb
86
1Mb
130
171
Rc:30147
Rc:30190
Rc:29929
1Mb
4
2Mb
3Mb

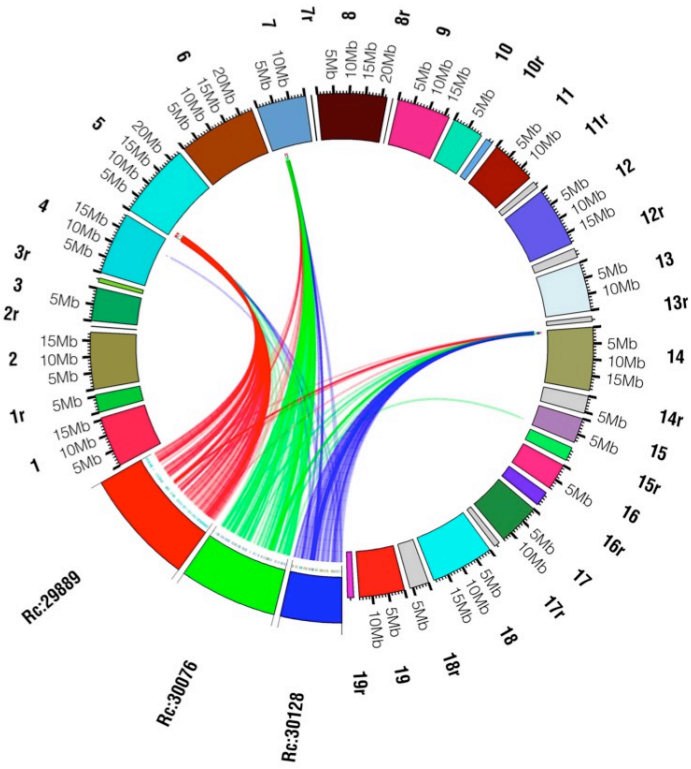
D


Rc:29889
0Mb
1Mb
Rc:30076
1Mb
0.5Mb
0Mb
Rc:30128
0Mb
2Mb

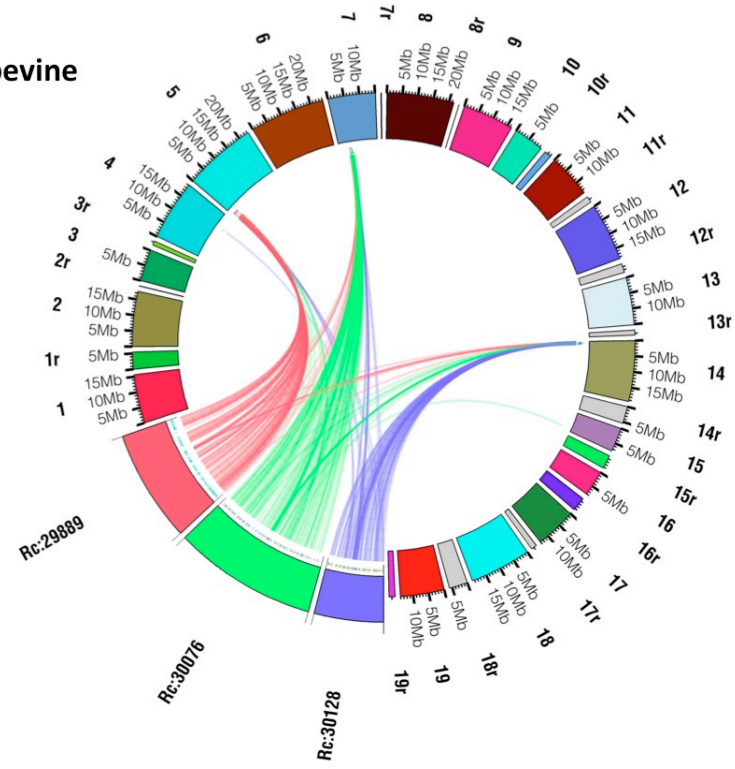

Rc:29889
Rc:30076
Rc:30128

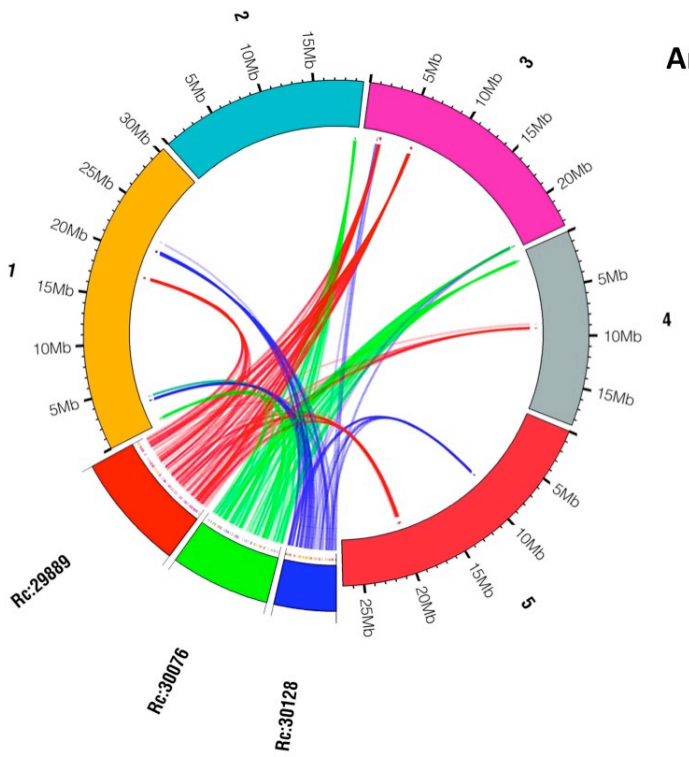
Poplar


Rc:29889
Rc:30076
Rc:30128

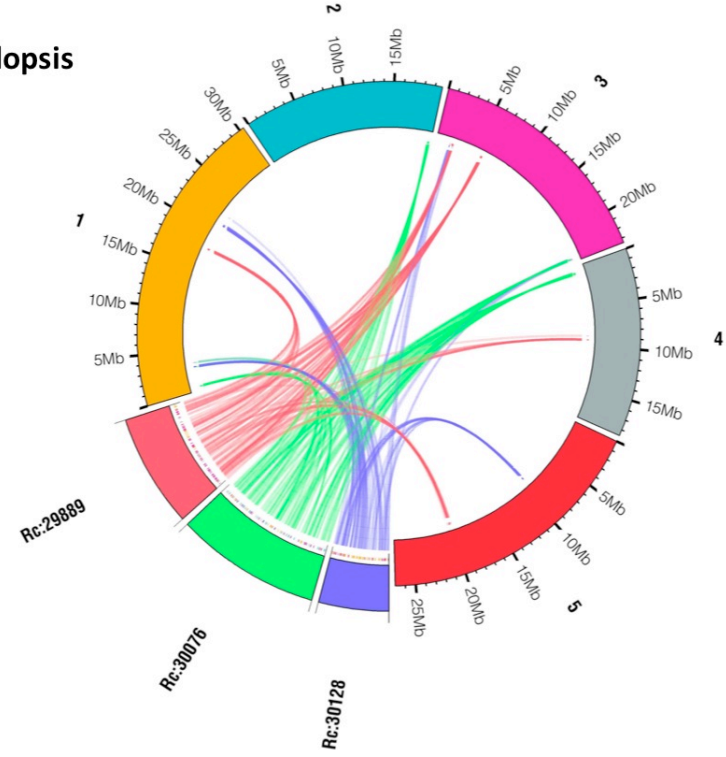


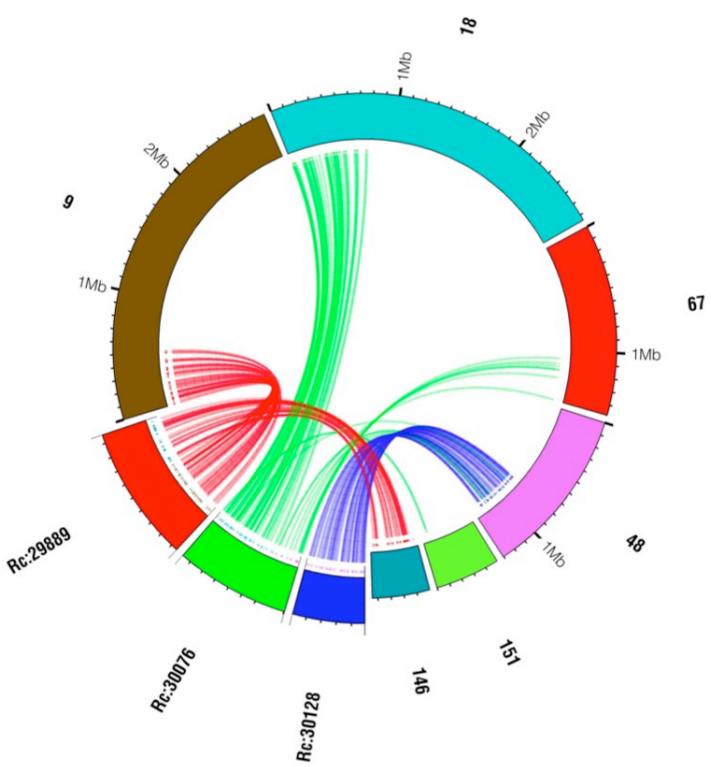
Grapevine



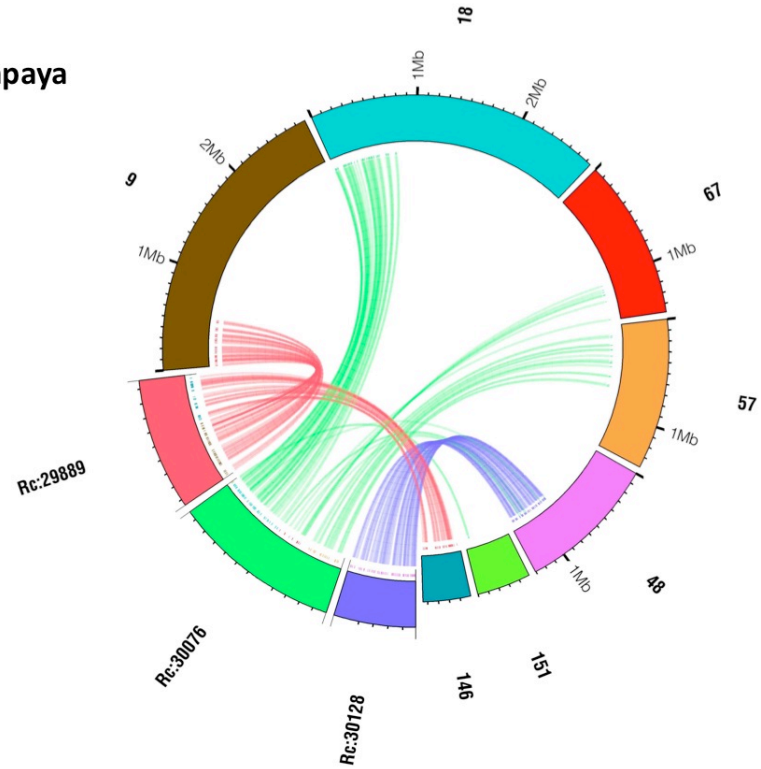


Arabidopsis

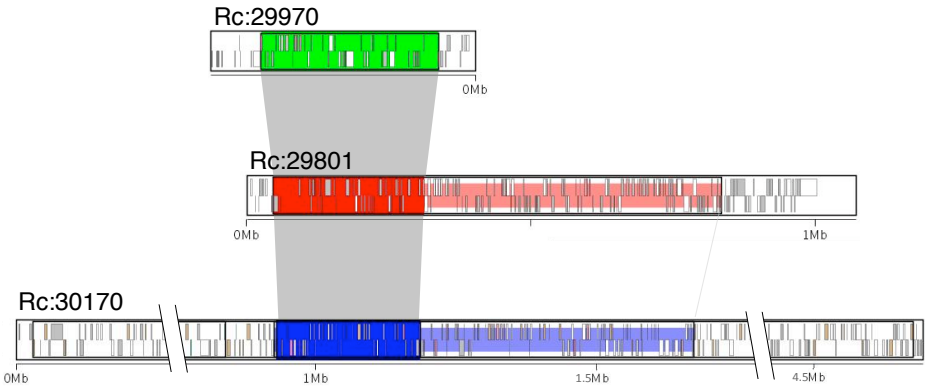


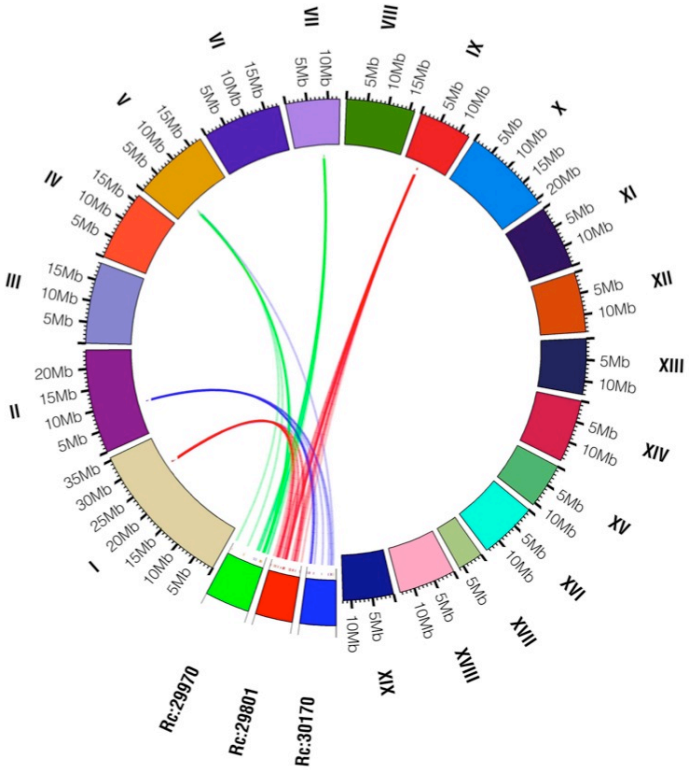

18
1Mb
2Mb
9
1Mb
2Mb
67
1Mb
48
1Mb
Rc:29889
Rc:30076
Rc:30128
146
151

Papaya

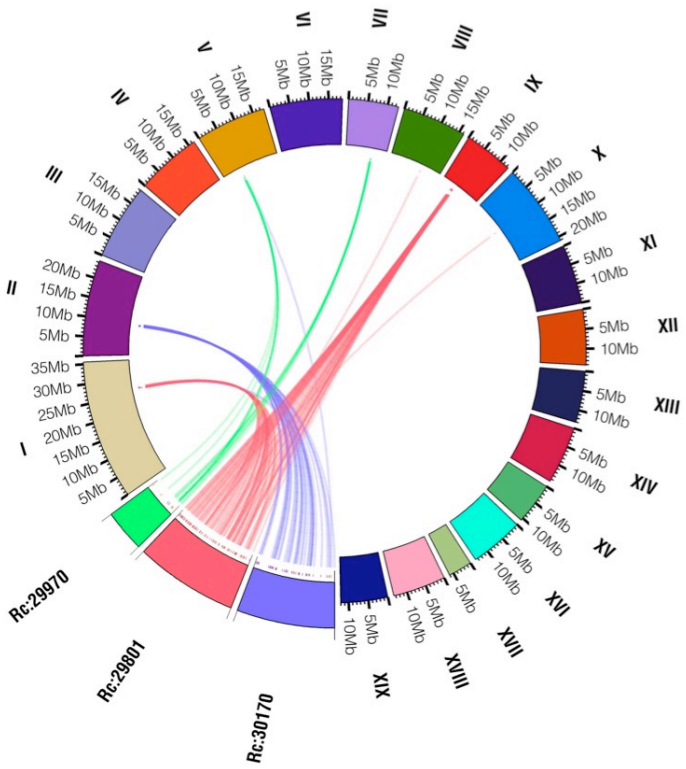

18
1Mb
2Mb
9
1Mb
2Mb
67
1Mb
57
1Mb
48
1Mb
Rc:29889
Rc:30076
Rc:30128
146
151

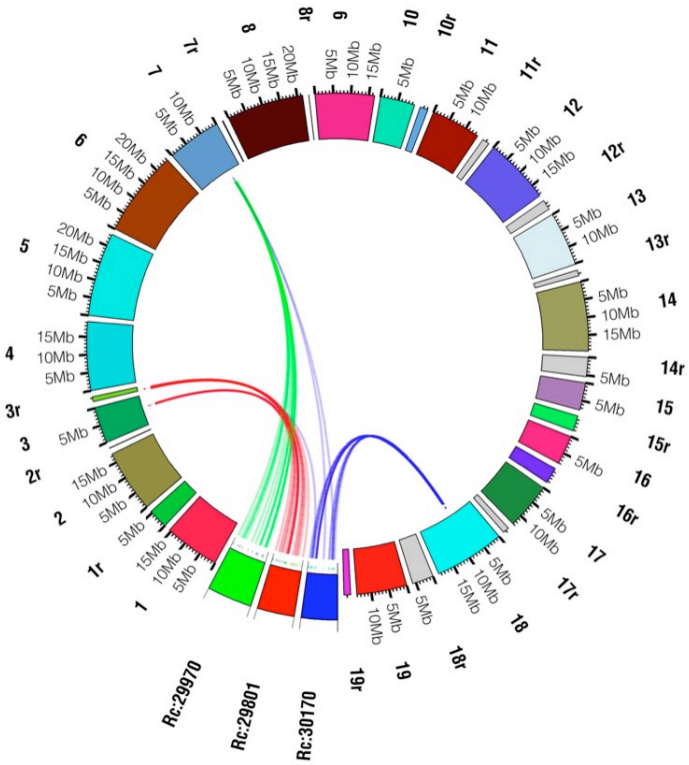
E


Rc:29970
0Mb
Rc:29801
0Mb
1Mb
Rc:30170
0Mb
1Mb
1.5Mb
4.5Mb

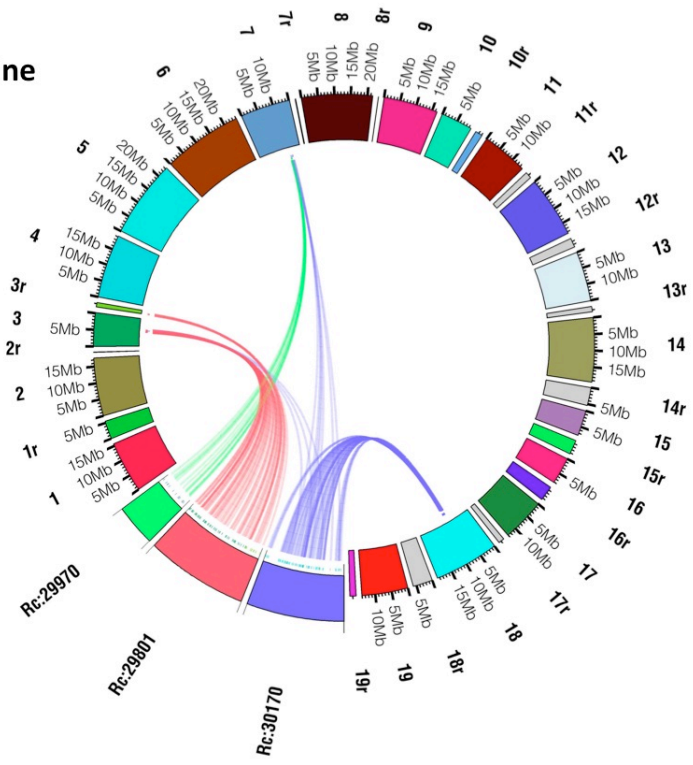

Rc:29970
Rc:29801
Rc:30170

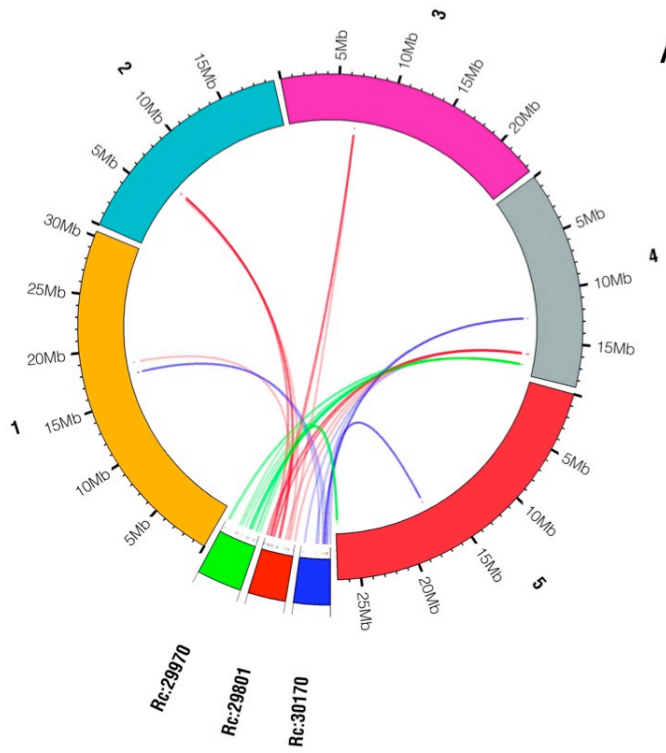
Poplar


Rc:29970
Rc:29801
Rc:30170

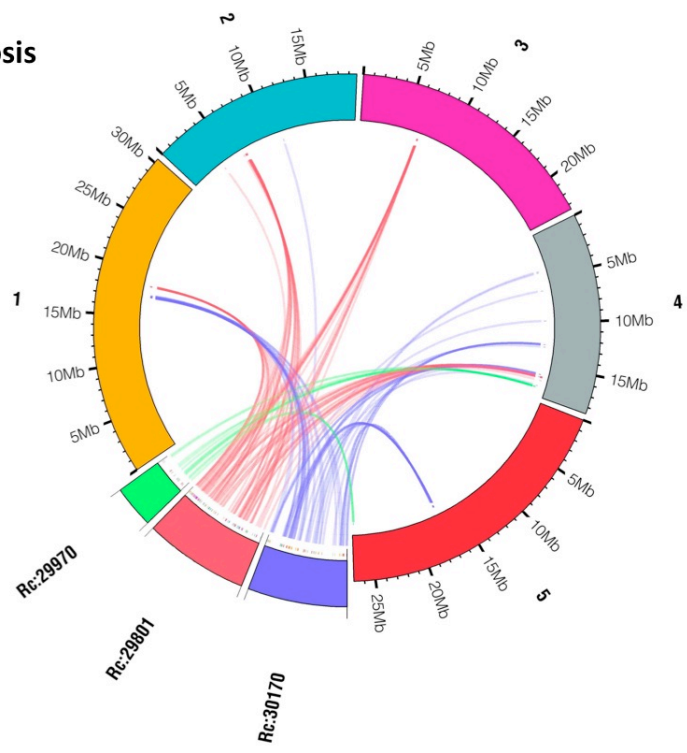


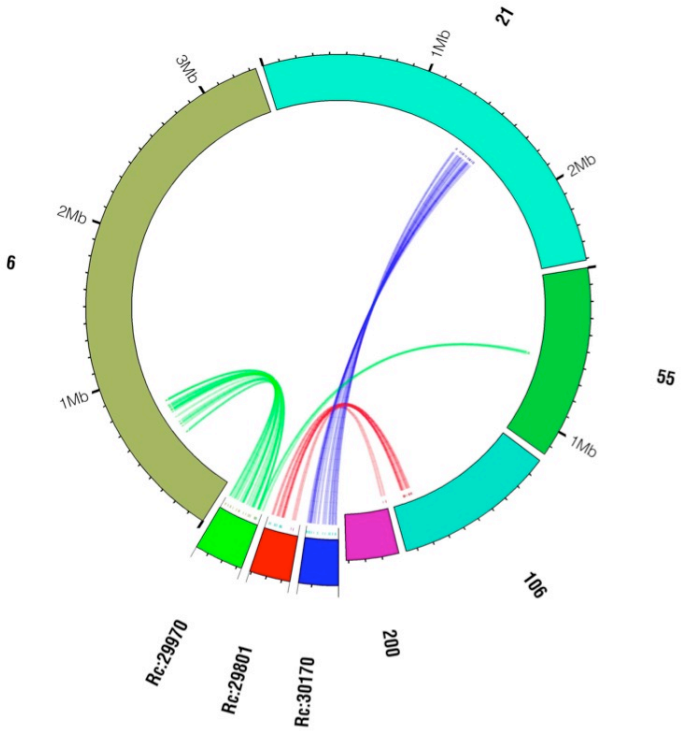
Grapevine



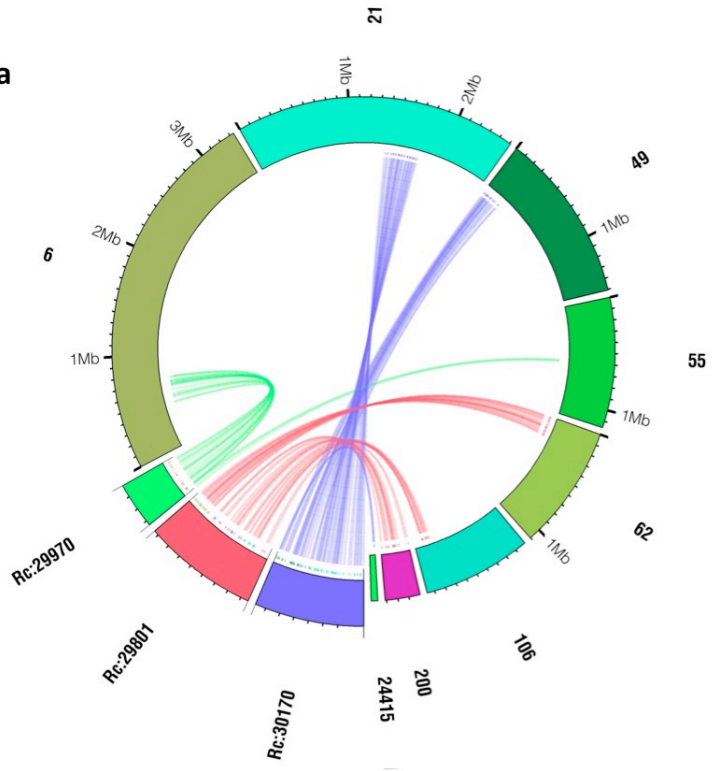


Arabidopsis

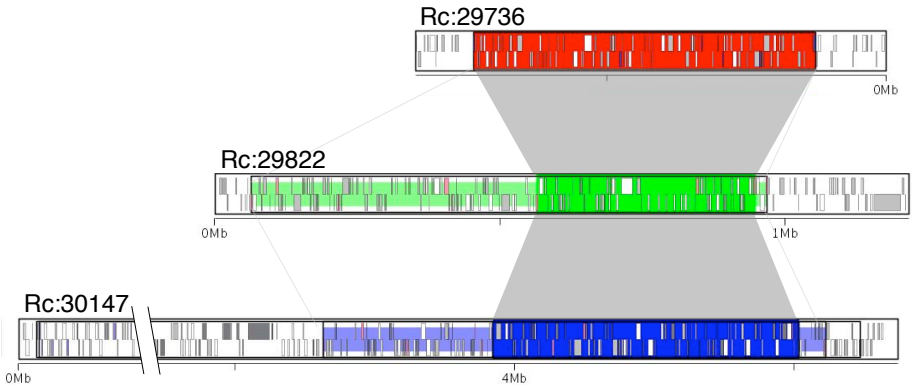


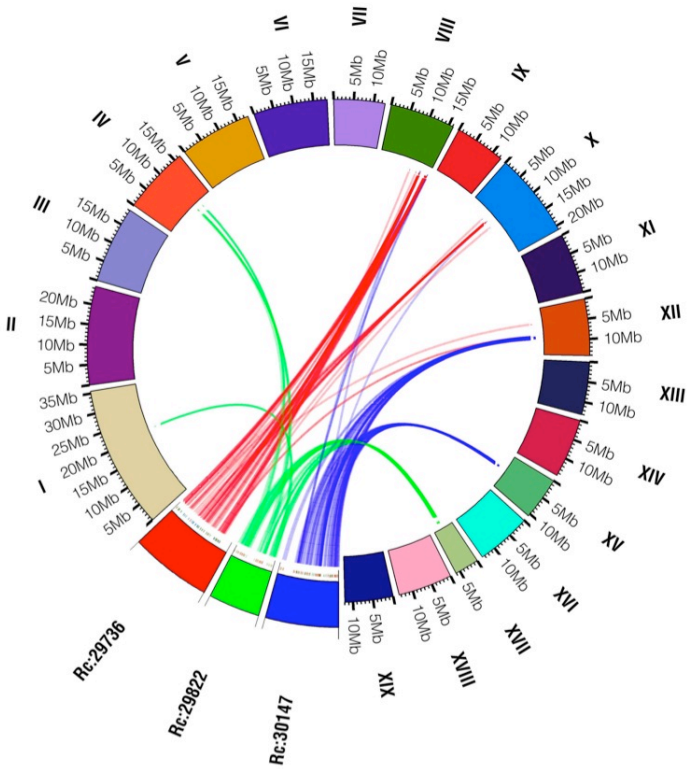

21
1Mb
2Mb
3Mb
6
2Mb
1Mb
55
1Mb
106
200
Rc:30170
Rc:29801
Rc:29970

Papaya

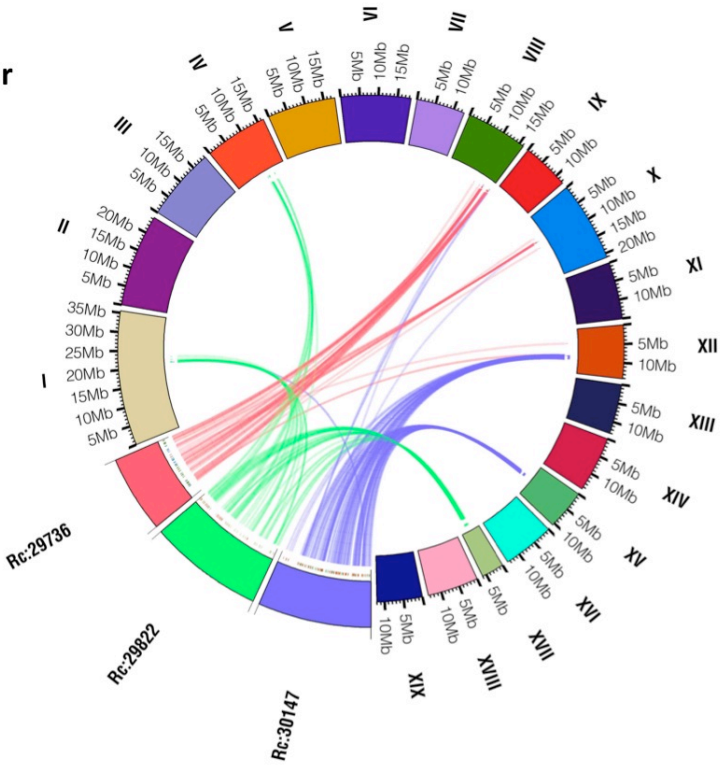

21
1Mb
2Mb
3Mb
49
6
2Mb
1Mb
55
1Mb
62
1Mb
106
200
24415
Rc:30170
Rc:29801
Rc:29970

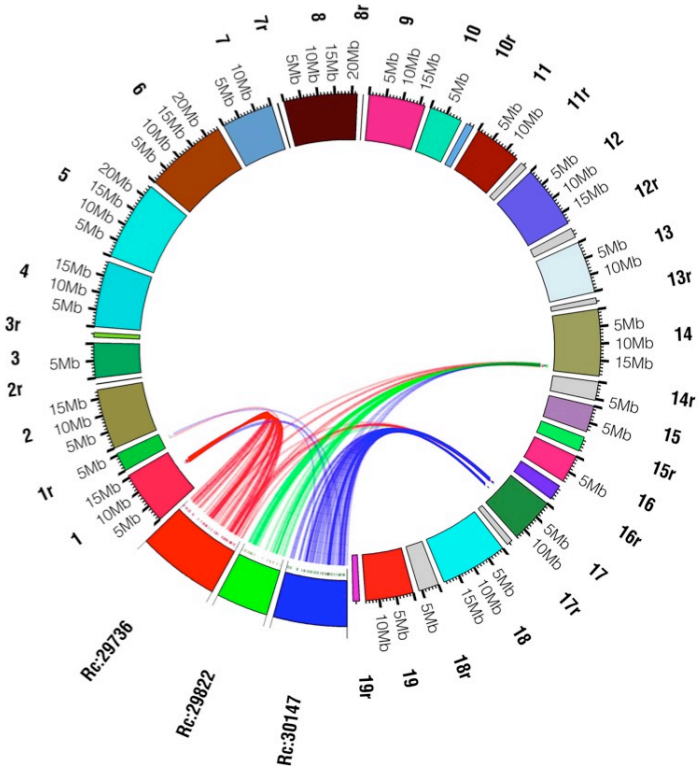
F


Rc:29736
0Mb
Rc:29822
0Mb
1Mb
Rc:30147
0Mb
4Mb

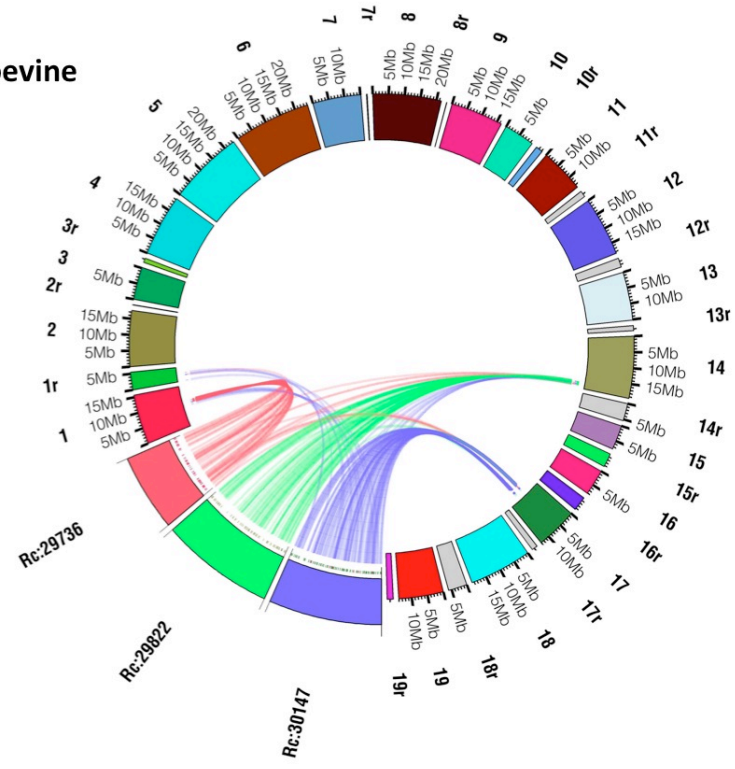

Rc:29736
Rc:29822
Rc:30147

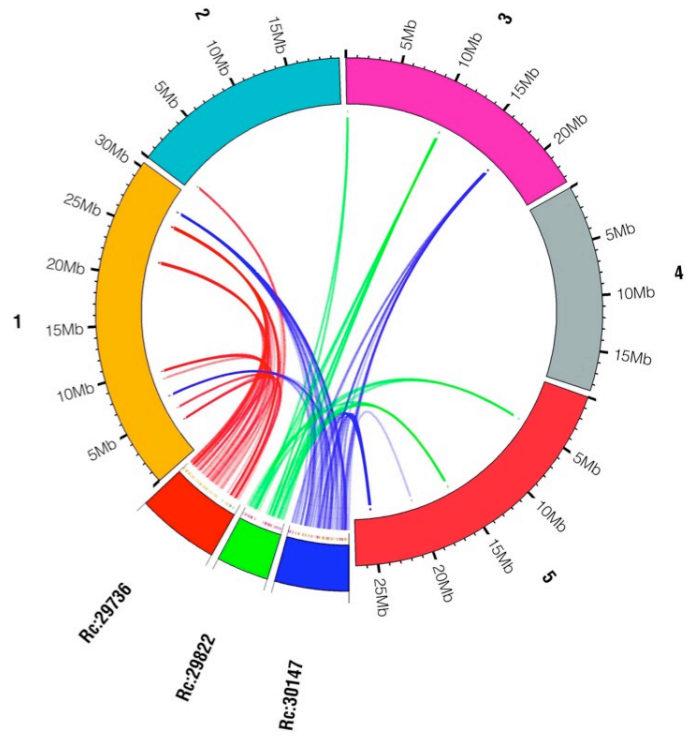
Poplar


Rc:29736
Rc:29822
Rc:30147

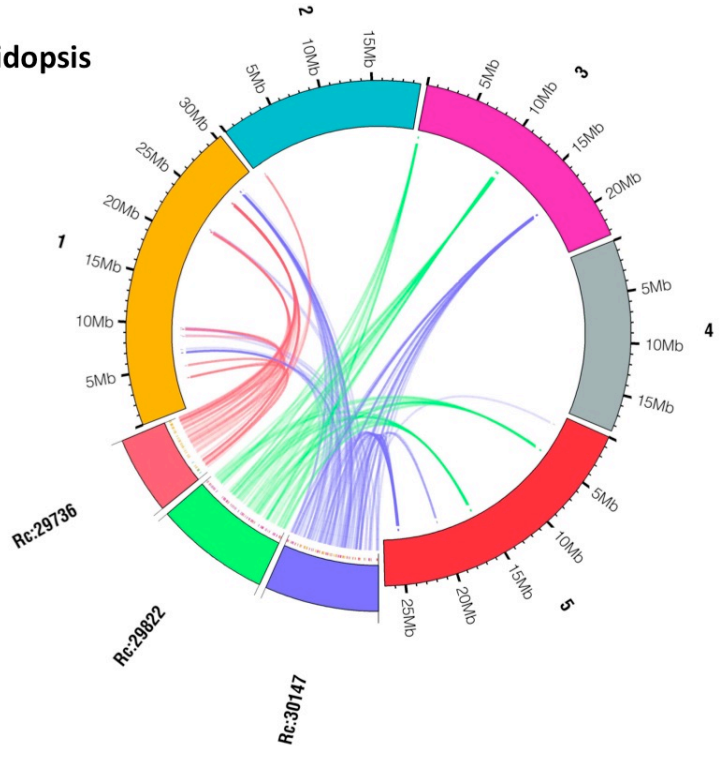


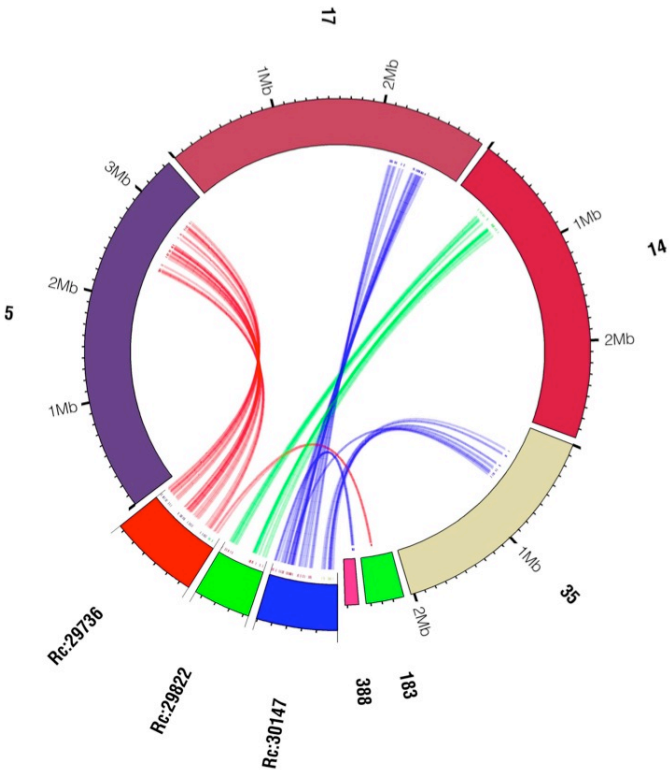
Grapevine



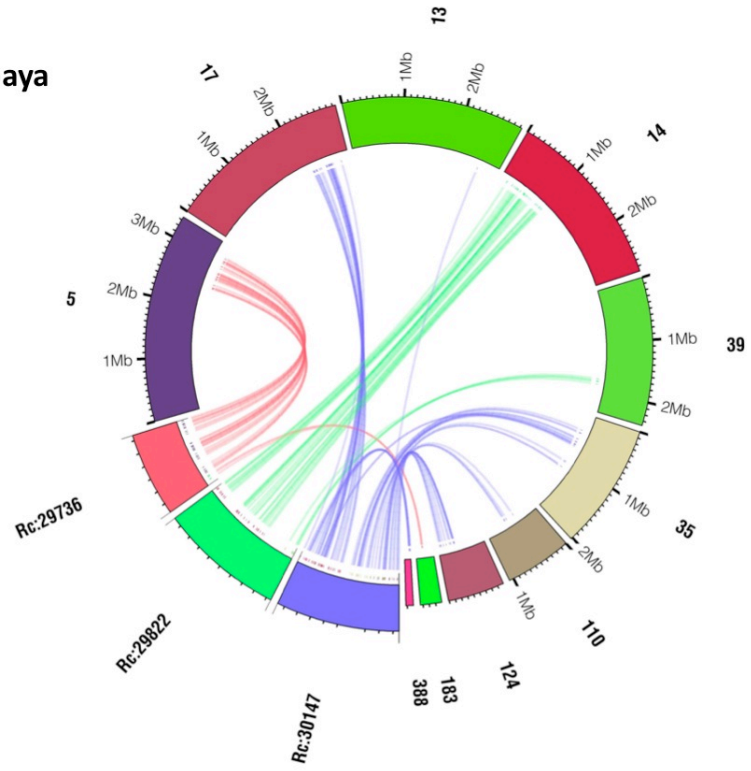


Arabidopsis

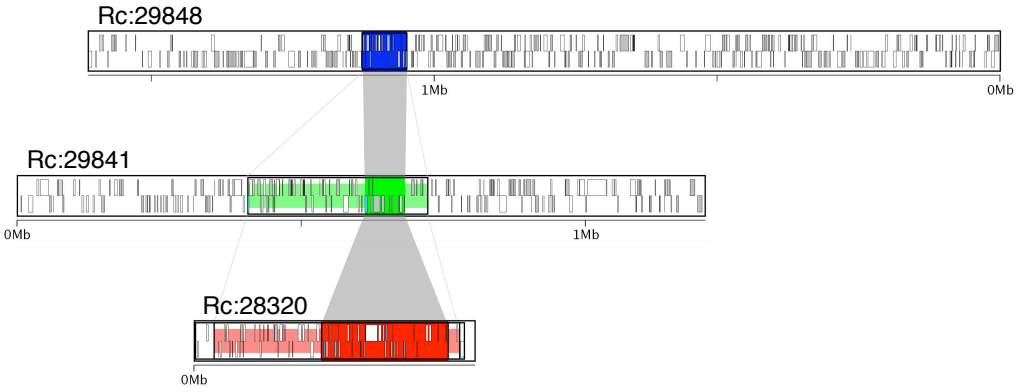


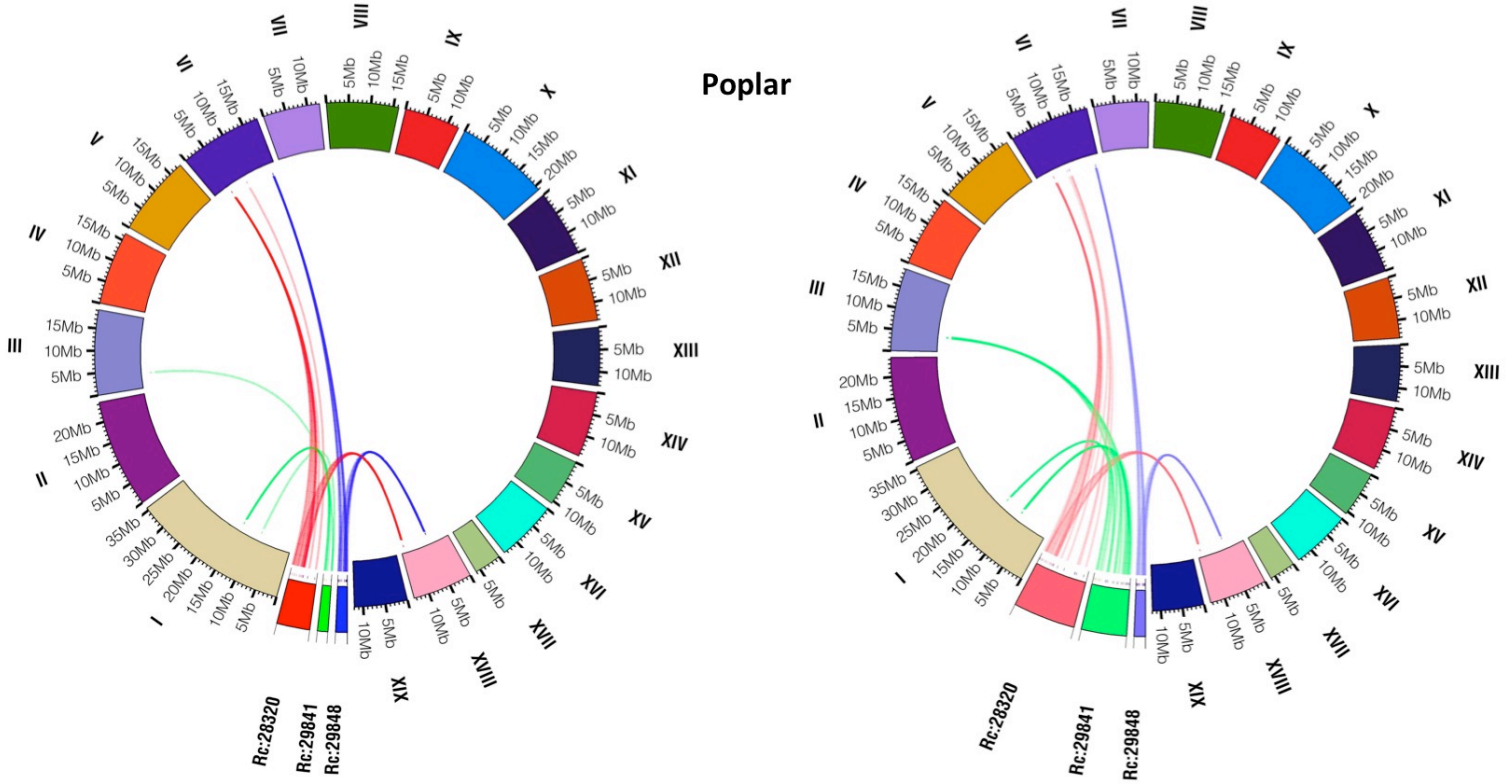

17
1Mb
2Mb
3Mb
5
2Mb
1Mb
14
2Mb
1Mb
35
2Mb
Rc:29736
Rc:29822
Rc:30147
388
183

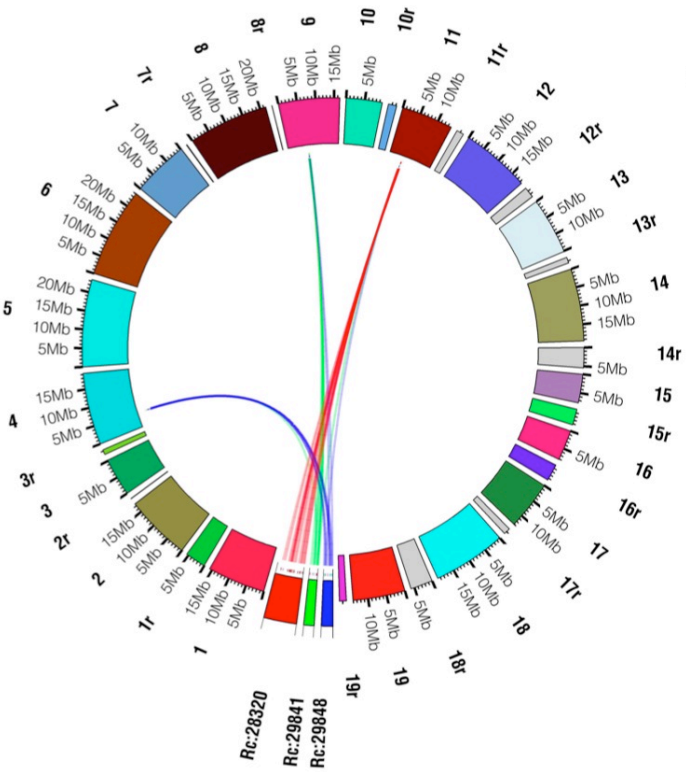
Papaya


13
17
2Mb
1Mb
2Mb
1Mb
14
3Mb
2Mb
5
1Mb
2Mb
1Mb
39
2Mb
1Mb
35
Rc:29736
2Mb
1Mb
110
Rc:29822
124
388
183
Rc:30147

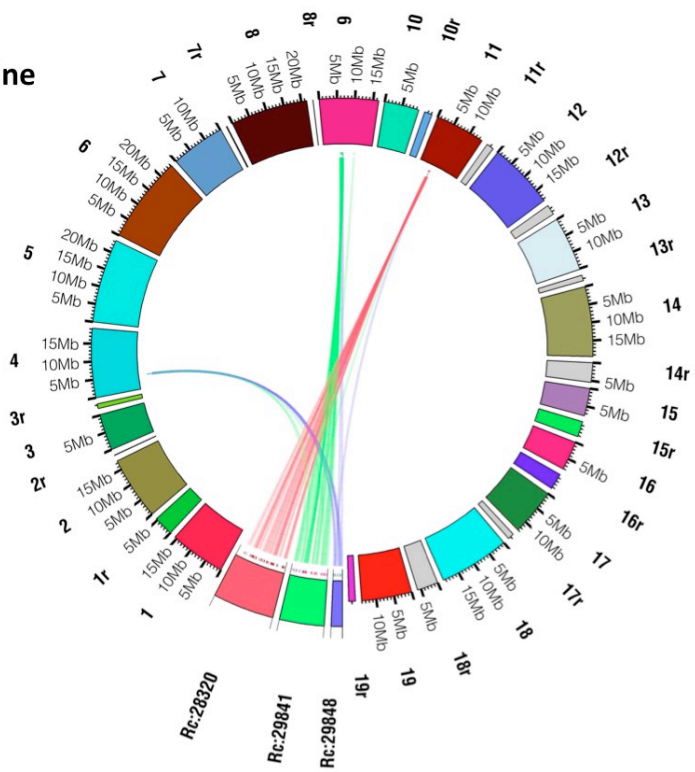
G

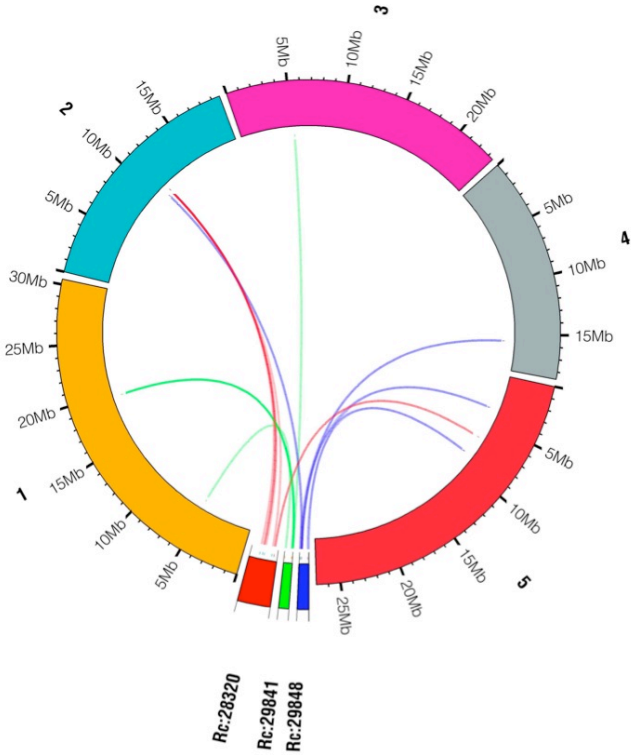

Rc:29848
1Mb
0Mb
Rc:29841
0Mb
1Mb
Rc:28320
0Mb


Poplar
Rc:28320
Rc:29841
Rc:29848

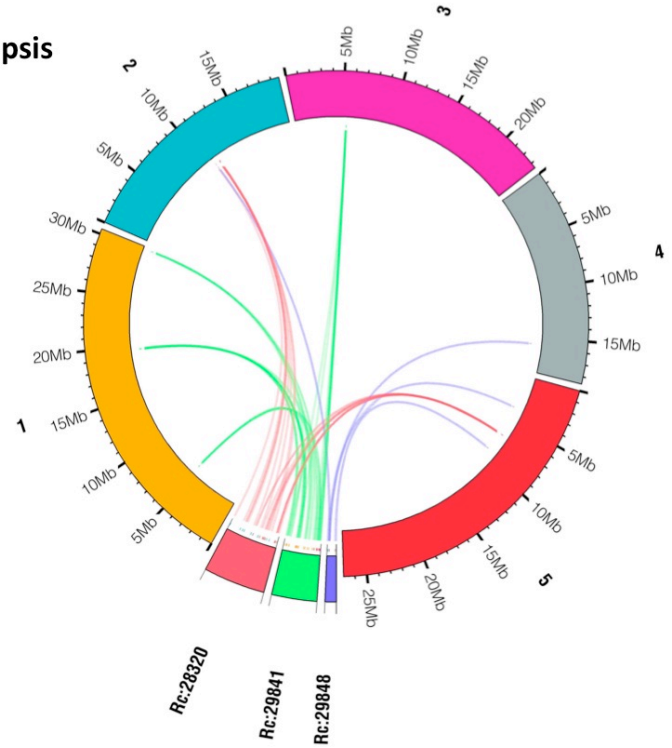


Grapevine

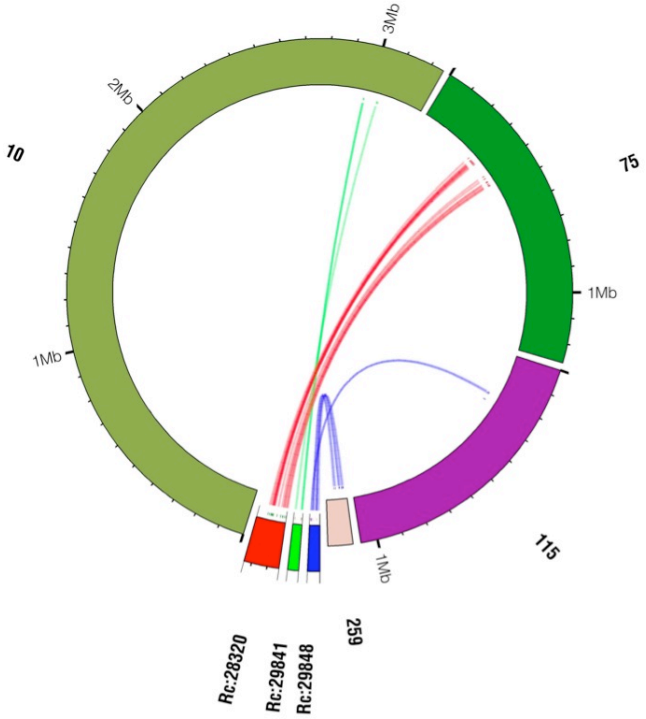


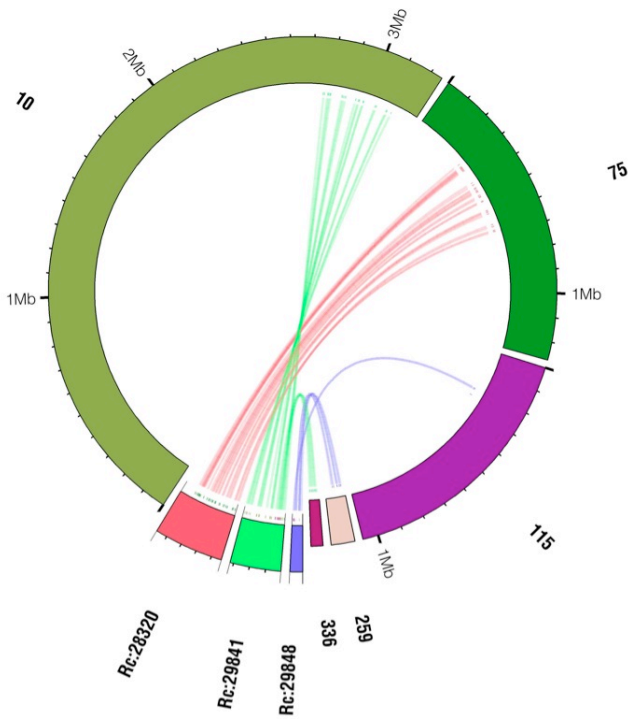


Arabidopsis



Papaya


3Mb
2Mb
10
75
1Mb
1Mb
115
1Mb
259
Rc:28320
Rc:29841
Rc:29848


3Mb
2Mb
10
75
1Mb
1Mb
115
1Mb
336
259
Rc:28320
Rc:29841
Rc:29848