Published in final edited form as:
*Nat Biotechnol*. 2010 September ; 28(9): 951–956. doi:10.1038/nbt.1674.

# Draft genome sequence of the ricin-producing oilseed castor bean

**Agnes P. Chan**[1,10], **Jonathan Crabtree**[2,10], **Qi Zhao**[1], **Hernan Lorenzi**[1], **Joshua Orvis**[2], **Daniela Puiu**[3], **Admasu Melake-Berhan**[1], **Kristine M. Jones**[2], **Julia Redman**[2], **Grace Chen**[4], **Edgar B. Cahoon**[5], **Melaku Gedil**[6], **Mario Stanke**[7], **Brian J. Haas**[8], **Jennifer R. Wortman**[2], **Claire M. Fraser-Liggett**[2], **Jacques Ravel**[2], and **Pablo D. Rabinowicz**[1,2,9]

[1]J. Craig Venter Institute (JCVI), Rockville, MD 20850, USA

[2]Institute for Genome Sciences (IGS), University of Maryland School of Medicine, Baltimore, MD 21201, USA

[3]Center for Bioinformatics and Computational Biology, University of Maryland, College Park, MD 20742, USA

[4]United States Department of Agriculture, Agricultural Research Service, Western Regional Research Center, Crop Improvement and Utilization, Albany, CA 94710, USA

[5]Center for Plant Science Innovation and Department of Biochemistry, University of Nebraska-Lincoln, Lincoln, NE 68588, USA

[6]International Institute of Tropical Agriculture, Oyo State, Ibadan, Nigeria

[7]Institut für Mikrobiologie und Genetik, Abteilung Bioinformatik, Universität Göttingen, Göttingen, Germany

[8]Broad Institute of the Massachusetts Institute of Technology and Harvard, Cambridge MA 02141, USA

[9]Department of Biochemistry and Molecular Biology, University of Maryland School of Medicine, Baltimore, MD 21201, USA

Correspondence to: Pablo D. Rabinowicz. Correspondence and requests for materials should be addressed to P.D.R. (prabinowicz@som.umaryland.edu).

[10]These authors contributed equally to this work

Note: Supplementary information is available on the Nature Biotechnology website.

**AUTHORS CONTRIBUTIONS**

A.P.C., J.C., H.L., B.J.H., and J.R.W. performed genomic analyses. Q.Z., J.O., and M.S. conducted genome annotation. D.P. worked on the genome assembly. A.M.-B., K.M.J., and J.R. made DNA preparations, library constructions, and closure work. G.C., E.B.C., and M.G. performed manual annotations. C.M.F.-L., J.R. conceived the project. P.D.R. conceived and directed the project.

**COMPETING FINANCIAL INTERESTS**

The authors declare no competing financial interests.

**Availability of data.** Sequence and annotation data has been submitted to GenBank (accession numbers AASG02000001-AASG02059013; and XP_002509419.1-XP_002540639.1), and the annotation data can also be freely accessed through the project's website (http://castorbean.jcvi.org), which includes a genome browser and a BLAST server.

Reprints and permissions information is available at www.npg.nature.com/reprintsandpermissions.

## Abstract

Castor bean (*Ricinus communis*) is an oil crop that belongs to the spurge (Euphorbiaceae) family. Its seeds are the source of castor oil, used for the production of high-quality lubricants due to its high proportion of the unusual fatty acid ricinoleic acid. Castor bean seeds also produce ricin, a highly toxic ribosome inactivating protein, making castor bean relevant for biosafety. We report here the 4.6X draft genome sequence of castor bean, representing the first reported Euphorbiaceae genome sequence. Our analysis shows that most key castor oil metabolism genes are single-copy while the ricin gene family is larger than previously thought. Comparative genomics analysis suggests the presence of an ancient hexaploidization event that is conserved across the dicotyledonous lineage.

Castor bean (*Ricinus communis*) is an oilseed crop that belongs to the family Euphorbiaceae, which is composed of 6,300 species including crops such as cassava (*Manihot esculenta*), rubber tree (*Hevea brasiliensis*), and physic nut (*Jatropha curcas*), as well as the invasive weed leafy spurge (*Euphorbia esula*) and ornamental poinsettias (*Euphorbia pulcherrima*). The castor bean plant is a tropical perennial shrub originated in Africa and cultivated in many tropical and subtropical regions around the world. It can be self- and cross-pollinated and worldwide studies showed low genetic diversity among castor bean germplasm1,2.

Castor bean seed oil contains 90% of the unusual hydroxy-fatty acid ricinoleic acid3. Because of the nearly uniform content of ricinoleic acid in castor oil, and the unique chemical properties that this fatty acid confers to the oil, castor bean is a highly valued oilseed crop for lubricant, cosmetic, medical, and specialty chemical applications. Furthermore, castor bean is a potential biodiesel source, due to its high seed oil content4 and because it can be cultivated in unfavorable environments, making it an appealing crop in tropical developing countries. It is believed that castor oil was first used as an ointment 4,000 years ago in Egypt, from where it spread to other parts of the world, including Greece and Rome, where it was used as a laxative 2,500 years ago5.

An important problem of castor bean as a crop is the high seed content of ricin, an extremely toxic protein6. Ricin is considered one of the deadliest natural poisons when administered intravenously or inhaled as fine particles. Ricin was first isolated more than a century ago7. It has been reportedly used as a weapon6 and attempts to using ricin as a specific immunotoxin for therapeutic purposes in different cancers have been reported8,9. Its biochemical activity has been characterized as a type 2 ribosome-inactivating protein (RIP), composed by two subunits linked by a disulfide bond: a 32 kDa ricin toxin A (RTA) chain that harbors the ribosome-inactivating activity, and a 34 kDa ricin toxin B (RTB) chain, with a galactose-binding lectin domain. RTA is an N-glycosidase that depurinates adenine in a specific residue of the 28S ribosomal RNA10,11. The RTB chain, allows ricin to enter eukaryotic cells by binding to cell surface galactosides and subsequent endocytosis. Other RIPs are common in plants, although they are not toxic because they are usually monomeric and lack a lectin domain. These proteins constitute the type 1 RIPs12.

Ricin is synthesized as a precursor encoding both subunits in the endoplasmic reticulum of endosperm cells, and is translocated and accumulated in protein bodies13. The precursor is proteolytically processed in the endoplasmic reticulum and in the protein bodies, where it is stored as the mature heterodimer.

Ricin is very similar to the *Ricinus communis* agglutinin or RCA14. However, while ricin is a weak hemagglutinin, RCA has low toxicity and a strong hemagglutinin activity. In addition, RCA is a tetrameric protein composed of two RTA- and two RTB-like subunits.

Purification of ricin can be achieved through a relatively simple process, raising biosafety concerns. For this reason, the United States does not extensively produce castor oil, and it is among the world's largest importers of castor oil and its derivatives. Therefore, knowledge of the castor oil metabolism is important to advance towards using castor oil as biofuel, as well as to enable metabolic engineering to obtain safe sources of hydroxy-fatty acids without the complications of ricin.

# RESULTS

## Genome sequencing and annotation

The castor bean genome size has been estimated in 320 Mb by flow cytometry15 and it is distributed in 10 chromosomes but, to our knowledge, no genetic map is available for castor bean. Because of the lack of available genomic information for castor bean, and due its importance for biodefense and as a model system in the Euphorbiaceae family, we set up to generate a draft sequence of the castor bean genome.

We produced approximately 2.1 million high-quality sequence reads from plasmid and fosmid libraries (see Supplementary methods), and used the Celera assembler to build consensus sequences or contigs and to link these contigs into 25,800 scaffolds using the two end-sequences from individual clones (mate-paired reads). The assembly covered the genome approximately 4.6X, spanning 350 Mb, which is consistent with previous genome size estimations. If only the 3,500 scaffolds larger than 2 kb are considered, the assembly spans 325 Mb with an N50 of 0.56 Mb (Table 1).

The genome sequence assembly was searched for repetitive DNA using a combination of sequence alignment to databases of repetitive sequences and RepeatScout to identify repeats *de novo*. Overall, over 50% of the genome was identified as repetitive DNA (excluding low-complexity sequences), most of which could not be associated with known element families. One third of the repetitive elements were retrotransposons, and less than 2% were DNA transposons (Table 2). The most abundant known repeats are long terminal repeat (LTR) elements (22.7% Gypsy-type and 9.5% Copia-type).

Protein coding genes were annotated using multiple gene-prediction programs, homology searches against sequence databases, and the cDNA spliced-alignment tool PASA (Program to Assemble Spliced Alignments). In order to aid the genome annotation, we also generated 52,165 expressed sequence tags (ESTs) from 5 cDNA non-normalized libraries. Using PASA, these and other castor bean cDNA sequences from GenBank could be aligned to 5,491 predicted genes and to 688 genomic regions where no gene had been predicted, allowing the creation of additional gene models. After all gene-prediction programs and homology searches were run, these data were consolidated into consensus gene predictions using the program Evidence Modeler (EVM; see Supplementary methods). EVM showed better sensitivity and specificity than any of the individual gene finders used (Supplementary Table 1). In this way, we identified 31,237 gene models (Table 1). Using TIGR's Paralogous Families pipeline, 58.5% of the castor bean gene models were grouped in 3,020 predicted protein families of at least two members (Supplementary Fig. 1; Supplementary Table 2).

## Polyploidization analysis

Although the castor bean genome assembly is fairly fragmented, it contains several megabase-sized scaffolds. Thus we attempted to investigate the extent of genome duplications in castor bean and contribute to the elucidation of the evolutionary history of the dicotyledonous lineage. Different models have been proposed to explain the origin of genome duplications in dicots. One supports the occurrence of an ancestral hexaploidization event common to all dicots16, while the other model suggests that all dicot genomes share one duplication event17. Analysis of genomic duplications in the castor bean genome brings an opportunity to advance toward resolving this controversy. Thus, we searched for putative paralogous genes using reciprocal best BLAST matches between all castor bean genes. We then selected the 30 pairs of scaffolds that contained the highest numbers of paralogous gene pairs. The 22 unique scaffolds containing those 30 pairs of scaffolds were displayed in a dot plot. This approach led to the identification of 6 triplicated regions (*i.e.* regions for which 2 additional paralogous regions exist in the genome). We also identified 9 duplicated regions (unmarked strings of dots) for which we cannot determine if a third paralogous region exists or not (Fig. 1). We then carried out a more precise and comprehensive search for evidence of genomic triplications by first building Jaccard clusters of paralogous genes using an all-versus-all BLASTP search. We identified and displayed blocks of syntenic genes using Sybil18 and manually inspected the results to identify triplicated regions. With this method, we identified 17 triplicated regions (Supplementary Fig. 2) that included those found using the reciprocal best BLAST matches method. The fact that the triplications were found in multiple groups of scaffolds suggests that the castor bean genome underwent a hexaploidization event.

In order to determine if the triplication of the castor bean genome corresponds to ancestral polyploidization events previously described in the dicot lineage, we compared the castor bean triplicated regions versus the *Arabidopsis*19, poplar20, grapevine16, and papaya21 genomes by generating Jaccard clusters in a pair-wise manner between castor bean and each of the other genomes. Out of the 17 triplications, 8 (including 5 of the 6 triplications identified by reciprocal best BLAST matches) contained blocks of 5 or more syntenic gene pairs between each of the three castor bean regions and all of the other dicot genomes. Castor bean paralogous gene blocks generally showed a one-to-one, one-to-two, and one-to-four relationship with their grapevine, poplar, and *Arabidopsis* orthologues, respectively (Fig. 2 and Supplementary Fig. 3). Some exceptions were observed in the comparison with *Arabidopsis* that were expected due to the further rearrangements that exist in its genome19. Comparison between the castor bean and papaya genomes is less clear due to the fragmentation of both genome assemblies. Our results support the presence of a hexaploidization event common to all dicots, as well as one additional genome duplication in poplar, and two further duplications in the *Arabidopsis* genome.

## The ricin gene family

As the presence of ricin makes castor bean an important subject for biosecurity research, we analyzed the lectin gene family that includes the genes for ricin and RCA. The ricin protein gene codes for three domains: an N-terminal RIP domain and two C-terminal lectin domains. It has been reported that this gene family is composed of 6 to 8 members, detected by Southern-blot hybridization using a ricin cDNA probe22,23. However, the castor bean genome revealed 28 putative genes in the family, including potential pseudogenes or gene fragments. In order to increase the reliability of our analysis of this gene family, sequence gaps or ambiguities inside the ricin-like gene models were subjected to manual finishing work to improve the sequence and assembly quality. In this way, the sequence and assembly of 8 scaffolds was improved and the 28 gene family members (Fig. 3) were contained in a total of 17 scaffolds, each containing 1 to 5 ricin-agglutinin gene family members (Supplementary Table 3). These results suggest that the members of this lectin family tend to be clustered in the castor bean genome. The largest cluster spans 70 kb and includes a group of 5 family members interrupted by one gene that does not belong to the gene family. The other clusters contain 2 or 3 gene family members in regions ranging between 0.7 and 17 kb. Ten scaffolds contained only one gene family member, and 4 of them were longer than 250 kb, suggesting that these four genes were not part of clusters. However it is uncertain if the other 6 scaffolds that contain only one member of the family are part of clusters because they are shorter than 12 kb. Probably, some of these tandem duplications were not discriminated in previous studies using Southern-blot analysis, resulting in an underestimation of the gene family size. Furthermore, although we did not manually curate structural annotation, we found two cases in which adjacent ricin-like gene fragments could belong to pseudogenes that accumulated frame shifts and stop codons (Fig. 3). The length of the different members of the family identified by automatic annotation was variable, ranging from 66 to 584 amino acids. Although some of the shorter genes could be non-functional or pseudogenes, start and stop codons could be predicted, making it difficult to determine their functionality, and 4 of them were truncated due to being at the end of a contig or scaffold. Sequence comparison to ricin and RCA coding sequences in GenBank uncovered one full-length gene model (60629.m00002) identical to the ricin coding sequence and another full-length gene model (60637.m00004) showing 99% identity to RCA coding sequence. These gene models likely correspond to the reported ricin and RCA sequences, respectively. An additional predicted gene (60628.m00003) shows 100% identity to the ricin coding sequence, although presumably, the sequence coding for about 150 of the 576 amino acids is missing from this gene model because it is located at the end of a scaffold. Three other gene models are truncated in a similar way (60626.m00001; 60639.m00003; 60627.m00002) and show 100% identity to the ricin coding sequence, but the available sequences are much shorter (149 to 188 amino acids). Thus, it is uncertain if these genes represent complete identical copies of the ricin-coding gene. The rest of the gene family members showed different degrees of similarity to the ricin or RCA coding sequences. Overall, the proteins coded by 7 genes (including ricin and RCA) out of the 28 family members contain the RIP and the two lectin domains, 9 contain only the RIP domain, and 9 contain one or two lectin domains only (Fig. 3). cDNA alignments showed evidence of expression of the genes coding ricin and RCA as well as one of the homologues (60638.m00018) for which a putatively complete gene was modeled (not shown). Furthermore, evidence of RIP activity has been recently reported for the proteins coded by the 7 full-length ricin-like genes24.

## Oil metabolism genes

Due to the importance of castor bean as an oilseed, we examined the annotation of 71 gene models that showed similarity to known genes involved in the biosynthesis of fatty acids and triacylglycerols, which in castor bean correspond mainly to ricinoleic acid and triricinolein25. Out of these 71 gene models, the annotation of 67 was manually improved (Supplementary Table 4). Castor bean has not only evolved an oleic acid hydroxylase to synthesize ricinoleic acid, but it also developed the capacity to efficiently accumulate high levels of ricinoleic acid in its seed oil. Therefore, we focused on a few key genes in the ricinoleic acid biosynthetic and metabolic pathways. The oleic acid hydroxylase gene (*FAH*) that produces ricinoleic acid from oleoyl-phosphatidycholine likely evolved from the widely occurring *FAD2* gene for the Δ12-oleic acid desaturase26. BLAST searches of these genes against the entire castor genome confirmed that there is only one copy of each of these genes (28035.m000362 and 29613.m000358, respectively). Among the key enzymes involved in the incorporation of ricinoleic acid into oils are diacylglycerol acyltransferases (DGATs), which catalyze the final step in triacylglycerol assembly. Two classes of endoplasmic reticulum-associated DGATs (DGAT1 and DGAT2) occur in castor bean, as well as a homolog of a soluble DGAT27,28,29. The gene models coding for these enzymes are also single-copy (29912.m005373, 29682.m000581, and 29889.m003411, respectively). In addition to DGAT-coding genes, it is likely that other genes have evolved to maintain high and specific flux of ricinoleic acid from its synthesis on phosphatidylcholine to its storage in triacylglycerols in castor bean seeds. Remarkably, even though ricinoleic acid accounts for nearly 90% of the fatty acid seeds in castor bean seeds, it is present at less than 5% of the fatty acids in phosphatidylcholine30. Although the mechanism for ricinoleic acid flux among lipid classes is not clear, a number of specialized acyltransferase and phosphatidylcholine metabolic enzymes likely participate in these reactions, including phospholipid:diacylglycerol acyltransferase 1 (PDAT1; 29912.m005286)31 and the recently identified phosphatidylcholine:diacylglycerol cholinephosphotransferase32 (PDCT; 29841.m002865). Information on copy number, genomic context, and regulatory regions of these and other metabolic genes will be important for the biotechnological transfer of ricinoleic acid production to established oilseed crops that lack ricin and its associated health risks. In addition, it is likely that the correct combination of specialized metabolic genes identified from the castor bean genome sequence will enable the engineering of triricinolein accumulation to amounts substantially higher than the modest levels achieved to date in model oilseeds33,34.

## Disease resistance genes

In order to contribute to the biotic stress research field in Euphorbiaceae, which is particularly important for cassava35, we compiled the castor bean predicted proteins whose functional annotation was related to disease resistance. One hundred and twenty one potential disease-resistance predicted proteins could be identified (Supplementary Table 5) using our automated annotation pipeline. The majority of these predicted proteins belong to the nucleotide binding-leucine-rich repeat (NBS-LRR) class, followed by the less common extracellular LRR-containing (eLRR) proteins36, and dirigent-like proteins that have been associated with disease resistance37. The castor bean gene models coding for these resistance genes were found distributed in 69 scaffolds and were often found in clusters of genes from the same class, although in some cases (*i.e.* scaffold 30190), different resistance gene classes are found in the same cluster (Supplementary Table 5). These data will be useful for comparative studies on resistance genes in cassava as well as other Euphorbiaceae crops.

## DISCUSSION

Due to the growing biosecurity importance of castor bean as the source of the highly toxic protein ricin38, genomic information becomes crucial for the development of improved diagnostic and forensic methods for ricin detection and cultivar identification for tracing sample origins. Molecular diagnostic methods39 and world-wide analyses of castor bean populations1,2 have been reported and the availability of the castor bean genome sequence will accelerate efforts to advance such studies and technologies.

In addition to its relevance for biosecurity, the castor bean genome information we report here can have implications for the production of biofuels and thus contribute to reducing greenhouse gas production. The industry of castor oil as a biodiesel component is being developed in Brazil4, where the use of biofuels is highly advanced. Furthermore, castor oil can also be used as lubricity additive to replace sulfur-based lubricant components in petroleum diesel, helping to reduce sulfur emissions40.

Unfortunately, the presence of ricin poses a problem for castor bean as widely cultivated oilseed crop. Therefore, considerable effort has been directed to engineering ricinoleic acid production in seeds of the model plant *Arabidopsis* as a prelude to transferring the required genes to an established ricin-free oilseed crop such as soybean. The initial strategy has involved the seed-specific expression of the castor bean *FAH* gene for the FAD2-related Δ12 oleic hydroxylase26, the key enzyme in ricinoleic acid synthesis41,42. However, transgenic expression of *FAH* resulted in the accumulation of ricinoleic acid and other hydroxy fatty acids to only 15 to 20% of the total fatty acids in *Arabidopsis* seeds41,42. Even co-expression of *FAH* with one additional ricinoleic acid metabolic gene, including the castor bean gene for DGAT2, yielded only small increases in ricinoleic acid accumulation in seeds of transgenic *Arabidopsis* that were far less than the levels typically found in castor bean seeds33,43. These results also reflect the modest production of other unusual fatty acids that has been achieved by expression of FAD2 variants such as the Δ12 epoxygenase and fatty acid conjugases in seeds of transgenic plants44,45. These results suggest that expression of a single biosynthetic gene, such as *FAH* alone or together with a gene involved in the metabolism of a given unusual fatty acid is insufficient to reproduce the oil composition observed in castor bean seeds. Thus, additional information on regulatory and metabolic genes is needed to fully transfer high levels of unusual fatty acid production and accumulation to engineered oilseed crops43,46,47. It is envisioned that the castor bean genome sequence and its annotation constitute the foundation for identifying the regulatory and metabolic networks controlling castor oil biosynthesis. In combination with metabolomics studies, these castor bean genome resources will enable metabolic engineering for improving castor oil production in crop plants lacking ricin.

Our analysis of the castor bean genome contributes to the debate on the polyploidization events that occurred in dicotyledonous genomes, supporting the presence of an ancestral hexaploidization event. Extending this type of analyses to cassava will have a great impact in the cassava research community as it will synergize with the recently released genome sequence of cassava (http://www.phytozome.net/cassava), which is an important food and, more recently, industrial crop in poor, tropical countries. It has been proposed that cassava is an allopolyploid48, and preliminary comparative genomics analyses between cassava and castor bean showed evidence of genomic duplications in cassava relative to castor bean (S. Rounsley, *pers. comm.*). These analyses suggest that the allopolyploidization event may have occurred in the cassava genome relatively recently, after the split between the two lineages. Further genome-wide comparative studies will provide insights on the genome evolution of cassava and the Euphorbiaceae family. Such information will help advance cassava breeding, which is a key means for developing countries to generate improved cassava lines with increased levels of stress resistance and nutritional content.

## METHODS

Methods and any associated references are available in the online version of the paper at http://www.nature.com/naturebiotechnology.

## Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

## ACKNOWLEDGEMENTS

This work was supported by the National Institute of Allergy and Infectious Diseases (NIAID), National Institutes of Health, Department of Health and Human Services, under NIAID Contract N01-AI-30071 (*R. communis* subproject) awarded to CMFL, JR, JRW, and PDR; Federal Bureau of Investigation grant J-FBI-04-186 to CMFL, JR, and PDR; and National Science Foundation grant DBI 0701919 to EBC. We thank the Joint Technology Center at JCVI for carrying out all sequencing work, and Ken Wurdack for his assistance with phylogenetics.

## REFERENCES

1. Allan G, et al. Worldwide genotyping of castor bean germplasm (*Ricinus communis L.*) using AFLPs and SSRs. Genet. Resour. Crop. Evol. 2008; 55:365–378.
2. Foster JT, et al. Single nucleotide polymorphisms for assessing genetic diversity in castor bean (Ricinus communis). BMC Plant Biol. 2010; 10:13. [PubMed: 20082707]
3. da Silva Ramos LC, Shogiro Tango J, Savi A, Leal NR. Variability for oil and fatty acid composition in castorbean varieties. J Am Oil Chem Soc. 1984; 61:1841–1843.
4. da Silva Nde L, Maciel MR, Batistella CB, Maciel Filho R. Optimization of biodiesel production from castor oil. Appl. Biochem. Biotechnol. 2006; 129–132:405–414.
5. Scarpa A, Guerci A. Various uses of the castor oil plant (Ricinus communis L.). A review. J. Ethnopharmacol. 1982; 5:117–137. [PubMed: 7035750]
6. Knight B. Ricin--a potent homicidal poison. Br Med J. 1979; 1:350–351. [PubMed: 421122]

7. Lord JM, Roberts LM, Robertus JD. Ricin: structure, mode of action, and some current applications. FASEB J. 1994; 8:201–208. [PubMed: 8119491]

8. Schnell R, et al. A Phase I study with an anti-CD30 ricin A-chain immunotoxin (Ki-4.dgA) in patients with refractory CD30+ Hodgkin's and non-Hodgkin's lymphoma. Clin. Cancer Res. 2002; 8:1779–1786. [PubMed: 12060617]

9. Fidias P, Grossbard M, Lynch TJ Jr. A phase II study of the immunotoxin N901-blocked ricin in small-cell lung cancer. Clin. Lung Cancer. 2002; 3:219–222. [PubMed: 14662047]

10. Endo Y, Mitsui K, Motizuki M, Tsurugi K. The mechanism of action of ricin and related toxic lectins on eukaryotic ribosomes. The site and the characteristics of the modification in 28 S ribosomal RNA caused by the toxins. J. Biol. Chem. 1987; 262:5908–5912. [PubMed: 3571242]

11. Macbeth MR, Wool IG. Characterization of in vitro and in vivo mutations in non-conserved nucleotides in the ribosomal RNA recognition domain for the ribotoxins ricin and sarcin and the translation elongation factors. J. Mol. Biol. 1999; 285:567–580. [PubMed: 9878430]

12. Lord JM, Hartley MR, Roberts LM. Ribosome inactivating proteins of plants. Semin. Cell Biol. 1991; 2:15–22. [PubMed: 1954339]

13. Lord JM. Synthesis and intracellular transport of lectin and storage protein precursors in endosperm from castor bean. Eur. J. Biochem. 1985; 146:403–409. [PubMed: 3967663]

14. Roberts LM, Lamb FI, Pappin DJ, Lord JM. The primary sequence of Ricinus communis agglutinin. Comparison with ricin. J. Biol. Chem. 1985; 260:15682–15686. [PubMed: 2999130]

15. Arumuganathan K, Earle ED. Nuclear DNA content of some important plant species. Plant Mol. Biol. Rep. 1991; 9:208–218.

16. Jaillon O, et al. The grapevine genome sequence suggests ancestral hexaploidization in major angiosperm phyla. Nature. 2007; 449:463–467. [PubMed: 17721507]

17. Velasco R, et al. A high quality draft consensus sequence of the genome of a heterozygous grapevine variety. PLoS One. 2007; 2:e1326. [PubMed: 18094749]

18. Crabtree J, Angiuoli SV, Wortman JR, White OR. Sybil: methods and software for multiple genome comparison and visualization. Methods Mol. Biol. 2007; 408:93–108. [PubMed: 18314579]

19. The Arabidopsis Genome Initiative. Analysis of the genome sequence of the flowering plant Arabidopsis thaliana. Nature. 2000; 408:796–815. [PubMed: 11130711]

20. Tuskan GA, et al. The genome of black cottonwood, Populus trichocarpa (Torr. & Gray). Science. 2006; 313:1596–1604. [PubMed: 16973872]

21. Ming R, et al. The draft genome of the transgenic tropical fruit tree papaya (Carica papaya Linnaeus). Nature. 2008; 452:991–996. [PubMed: 18432245]

22. Halling KC, et al. Genomic cloning and characterization of a ricin gene from Ricinus communis. Nucleic Acids Res. 1985; 13:8019–8033. [PubMed: 2999712]

23. Tregear JW, Roberts LM. The lectin gene family of Ricinus communis: cloning of a functional ricin gene and three lectin pseudogenes. Plant Mol. Biol. 1992; 18:515–525. [PubMed: 1371405]

24. Leshin J, et al. Characterization of ricin toxin family members from Ricinus communis. Toxicon. 2010; 55:658–661. [PubMed: 19781564]

25. McKeon TA, Chen GQ, Lin JT. Biochemical aspects of castor oil biosynthesis. Biochem. Soc. Tran.s. 2000; 28:972–974.

26. van de Loo FJ, Broun P, Turner S, Somerville C. An oleate 12- hydroxylase from Ricinus communis L. is a fatty acyl desaturase homolog. Proc. Natl. Acad. Sci. USA. 1995; 92:6743–6747. [PubMed: 7624314]

27. He X, Turner C, Chen GQ, Lin JT, McKeon TA. Cloning and characterization of a cDNA encoding diacylglycerol acyltransferase from castor bean. Lipids. 2004; 39:311–318. [PubMed: 15357018]

28. Kroon JT, Wei W, Simon WJ, Slabas AR. Identification and functional expression of a type 2 acyl-CoA:diacylglycerol acyltransferase (DGAT2) in developing castor bean seeds which has high homology to the major triglyceride biosynthetic enzyme of fungi and animals. Phytochemistry. 2006; 67:2541–2549. [PubMed: 17084870]

29. Saha S, Enugutti B, Rajakumari S, Rajasekharan R. Cytosolic triacylglycerol biosynthetic pathway in oilseeds. Molecular cloning and expression of peanut cytosolic diacylglycerol acyltransferase. Plant Physiol. 2006; 141:1533–1543. [PubMed: 16798944]

30. Thomaeus S, Carlsson AS, Stymne S. Distribution of fatty acids in polar and neutral lipids during seed development in Arabidopsis thaliana genetically engineered to produce acetylenic, epoxy and hydroxy fatty acids. Plant Sci. 2001; 161:997–1003.

31. Dahlqvist A, et al. Phospholipid:diacylglycerol acyltransferase: an enzyme that catalyzes the acyl-CoA-independent formation of triacylglycerol in yeast and plants. Proc Natl Acad Sci U S A. 2000; 97:6487–6492. [PubMed: 10829075]

32. Lu C, Xin Z, Ren Z, Miquel M, Browse J. An enzyme regulating triacylglycerol composition is encoded by the ROD1 gene of Arabidopsis. Proc. Natl. Acad. Sci. USA. 2009; 106:18837–18842. [PubMed: 19833868]

33. Burgal J, et al. Metabolic engineering of hydroxy fatty acid production in plants: RcDGAT2 drives dramatic increases in ricinoleate levels in seed oil. Plant Biotechnol. J. 2008; 6:819–831. [PubMed: 18643899]

34. Cahoon EB, et al. Engineering oilseeds for sustainable production of industrial and nutritional feedstocks: solving bottlenecks in fatty acid flux. Curr. Opin. Plant. Biol. 2007; 10:236–244. [PubMed: 17434788]

35. Hillocks RJ, Jennings DL. Cassava brown streak disease: a review of present knowledge and research needs. International Journal of Pest Management. 2003; 49:225–234.

36. van Ooijen G, van den Burg HA, Cornelissen BJ, Takken FL. Structure and function of resistance proteins in solanaceous plants. Annu. Rev. Phytopathol. 2007; 45:43–72. [PubMed: 17367271]

37. Fristensky B, Horovitz D, Hadwiger LA. cDNA sequences for pea disease resistance response genes. Plant Mol. Biol. 1988; 11:713–715. [PubMed: 24272504]

38. Musshoff F, Madea B. Ricin poisoning and forensic toxicology. Drug Test Anal. 2009; 1:184–191. [PubMed: 20355196]

39. Audi J, Belson M, Patel M, Schier J, Osterloh J. Ricin poisoning: a comprehensive review. JAMA. 2005; 294:2342–2351. [PubMed: 16278363]

40. Goodrum JW, Geller DP. Influence of fatty acid methyl esters from hydroxylated vegetable oils on diesel fuel lubricity. Bioresour. Technol. 2005; 96:851–855. [PubMed: 15607199]

41. Broun P, Somerville C. Accumulation of ricinoleic, lesquerolic, and densipolic acids in seeds of transgenic Arabidopsis plants that express a fatty acyl hydroxylase cDNA from castor bean. Plant Physiol. 1997; 113:933–942. [PubMed: 9085577]

42. Smith MA, Moon H, Chowrira G, Kunst L. Heterologous expression of a fatty acid hydroxylase gene in developing seeds of Arabidopsis thaliana. Planta. 2003; 217:507–516. [PubMed: 14520576]

43. Lu C, Fulda M, Wallis JG, Browse J. A high-throughput screen for genes from castor that boost hydroxy fatty acid accumulation in seed oils of transgenic Arabidopsis. Plant J. 2006; 45:847–856. [PubMed: 16460516]

44. Li R, Yu K, Hatanaka T, Hildebrand DF. Vernonia DGATs increase accumulation of epoxy fatty acids in oil. Plant Biotechnol. J. 2010; 8:184–195. [PubMed: 20078841]

45. Cahoon EB, et al. Conjugated fatty acids accumulate to high levels in phospholipids of metabolically engineered soybean and Arabidopsis seeds. Phytochemistry. 2006; 67:1166–1176. [PubMed: 16762380]

46. Cernac A, Benning C. WRINKLED1 encodes an AP2/EREB domain protein involved in the control of storage compound biosynthesis in Arabidopsis. Plant J. 2004; 40:575–585. [PubMed: 15500472]

47. Thelen J, Ohlrogge J. Metabolic engineering of fatty acid biosynthesis in plants. Metab. Eng. 2002; 4:12–21. [PubMed: 11800570]

48. Umanah EE, Hartmann RW. Chromosome numbers and karyotypes of some Manihot species. Am. Soc. Hortic. Sci. 1973; 98:272–274.

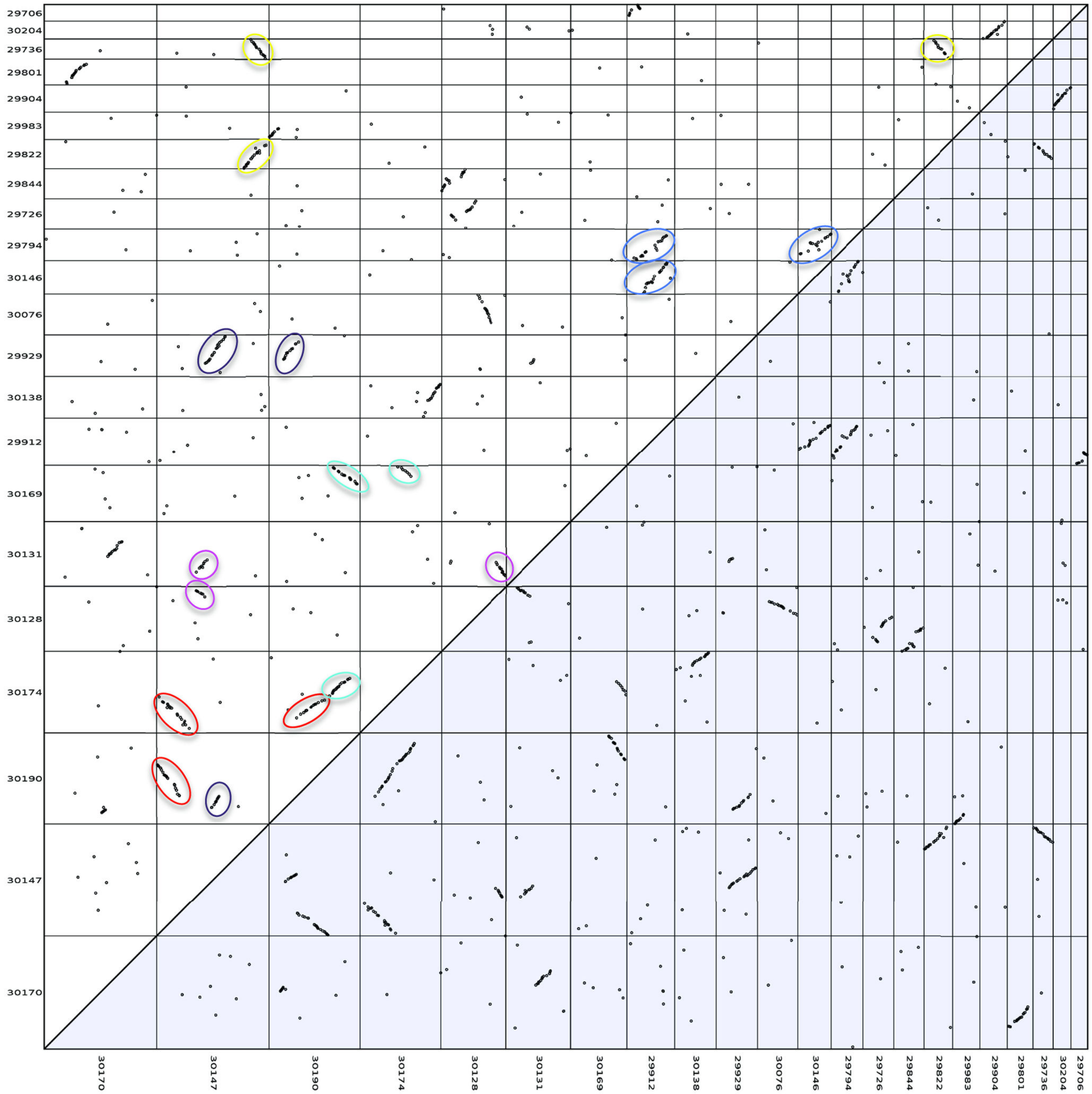
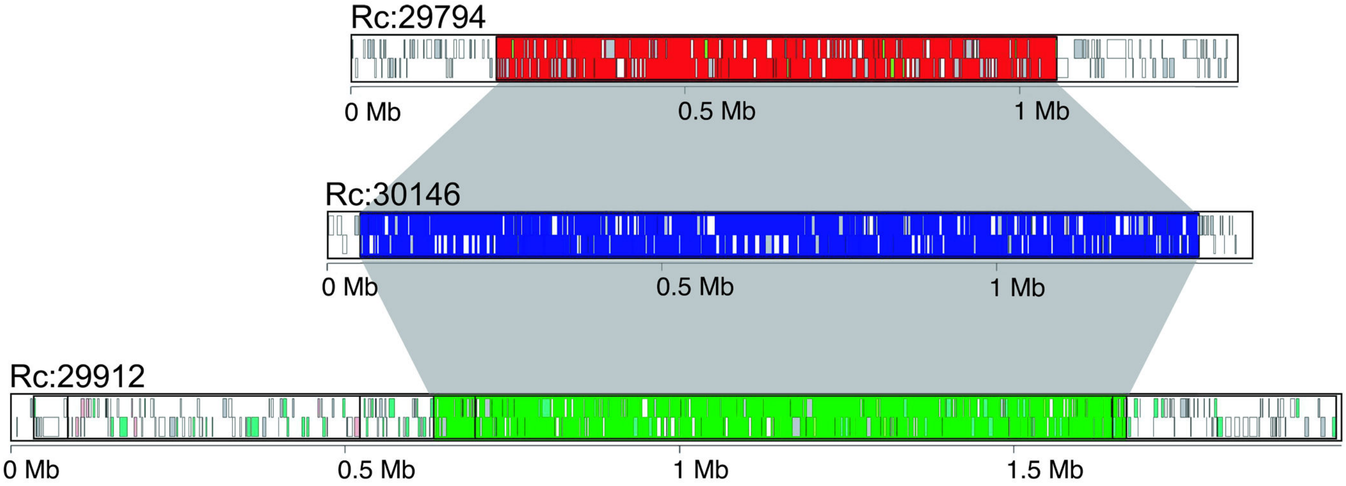
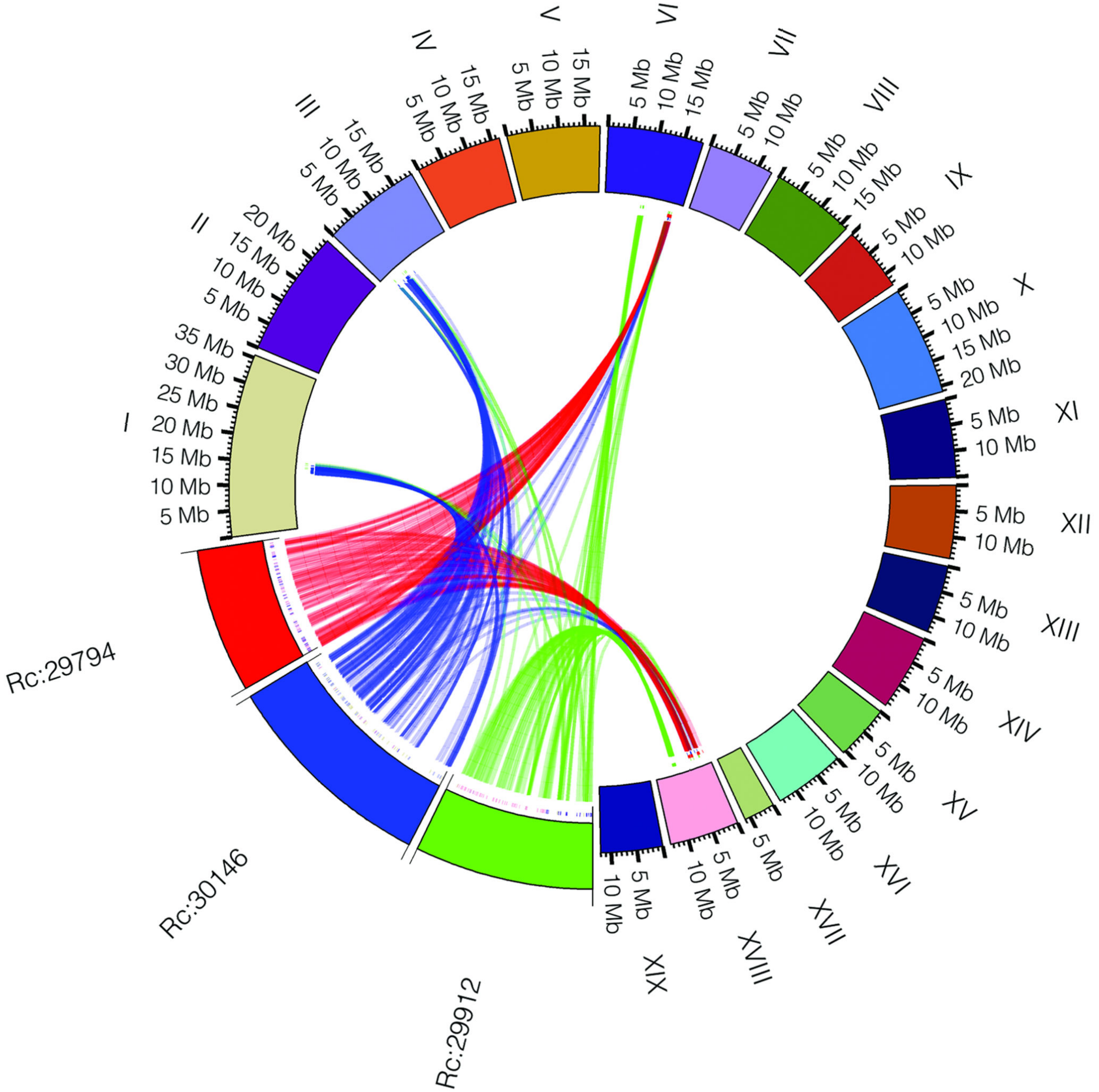
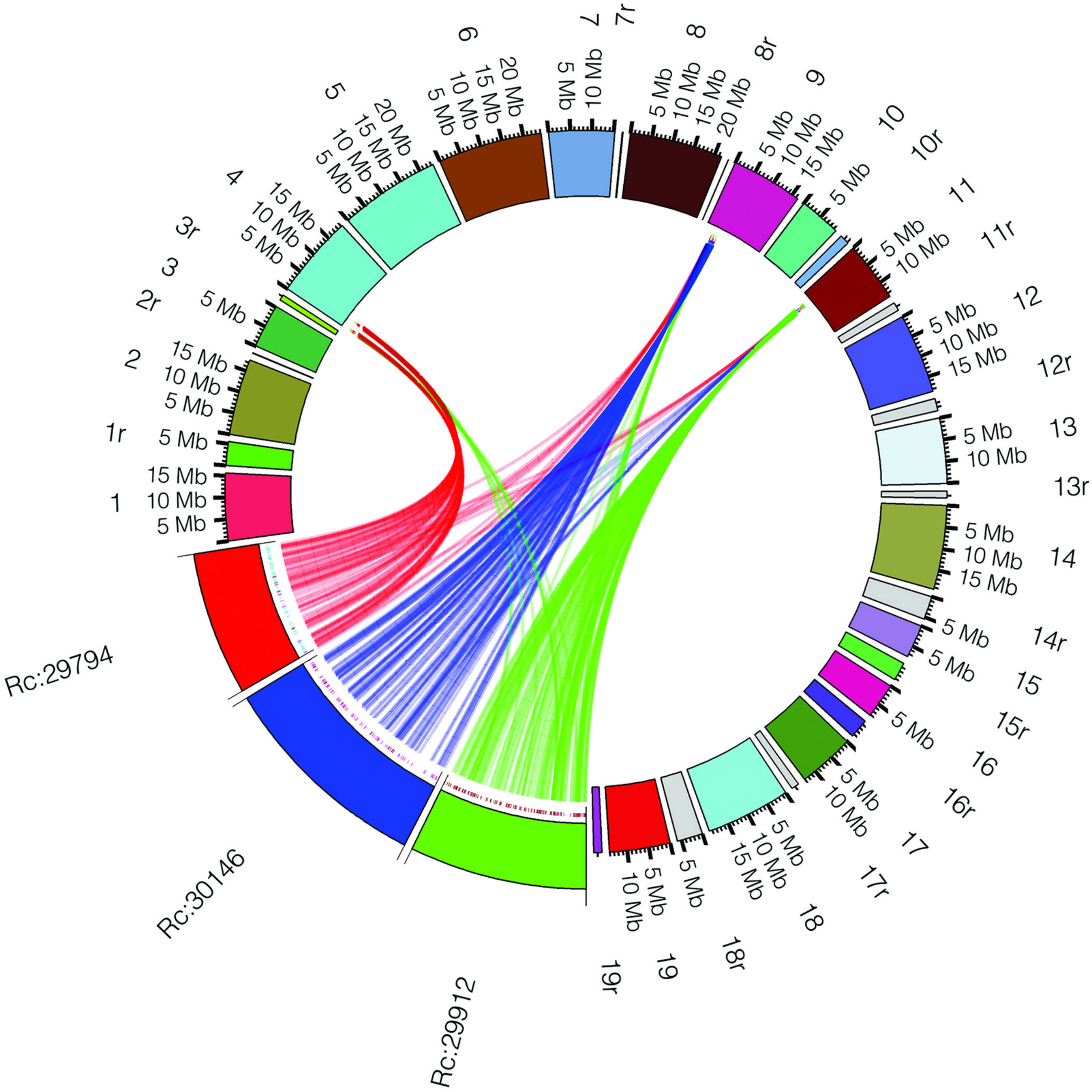


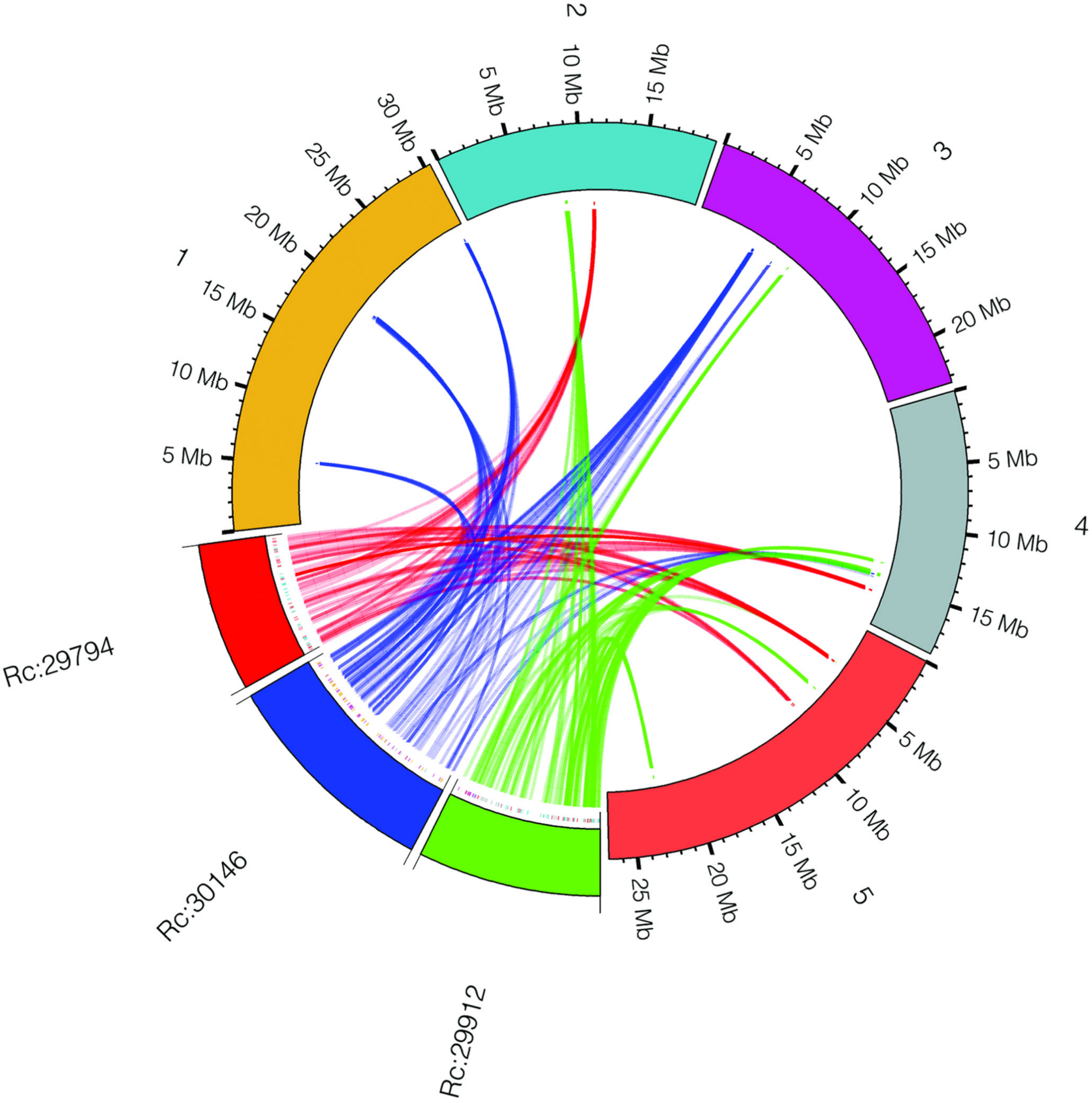
**Figure 1. Reciprocal best BLAST matches between castor bean genes**

Strings of paralogous genes that correspond to triplicated regions are highlighted in the same color. The 30 pairs of scaffolds that contained the highest numbers of paralogous gene pairs are shown.


Rc:29794
0 Mb
0.5 Mb
1 Mb
Rc:30146
0 Mb
0.5 Mb
1 Mb
Rc:29912
0 Mb
0.5 Mb
1 Mb
1.5 Mb


I
5 Mb
10 Mb
15 Mb
20 Mb
25 Mb
30 Mb
35 Mb
II
5 Mb
10 Mb
15 Mb
20 Mb
III
5 Mb
10 Mb
15 Mb
IV
5 Mb
10 Mb
15 Mb
V
5 Mb
10 Mb
15 Mb
VI
5 Mb
10 Mb
15 Mb
VII
5 Mb
10 Mb
VIII
5 Mb
10 Mb
15 Mb
IX
5 Mb
10 Mb
X
5 Mb
10 Mb
15 Mb
20 Mb
XI
5 Mb
10 Mb
XII
5 Mb
10 Mb
XIII
5 Mb
10 Mb
XIV
5 Mb
10 Mb
XV
5 Mb
10 Mb
XVI
5 Mb
10 Mb
XVII
5 Mb
XVIII
5 Mb
10 Mb
XIX
5 Mb
10 Mb
Rc:29912
Rc:30146
Rc:29794


1
1r
2
2r
3
3r
4
5
6
7
7r
8
8r
9
10
10r
11
11r
12
12r
13
13r
14
14r
15
15r
16
16r
17
17r
18
18r
19
19r
5 Mb
10 Mb
15 Mb
20 Mb
Rc:29794
Rc:30146
Rc:29912


1
5 Mb
10 Mb
15 Mb
20 Mb
25 Mb
30 Mb
2
5 Mb
10 Mb
15 Mb
3
5 Mb
10 Mb
15 Mb
20 Mb
4
5 Mb
10 Mb
15 Mb
5
5 Mb
10 Mb
15 Mb
20 Mb
25 Mb
Rc:29794
Rc:30146
Rc:29912

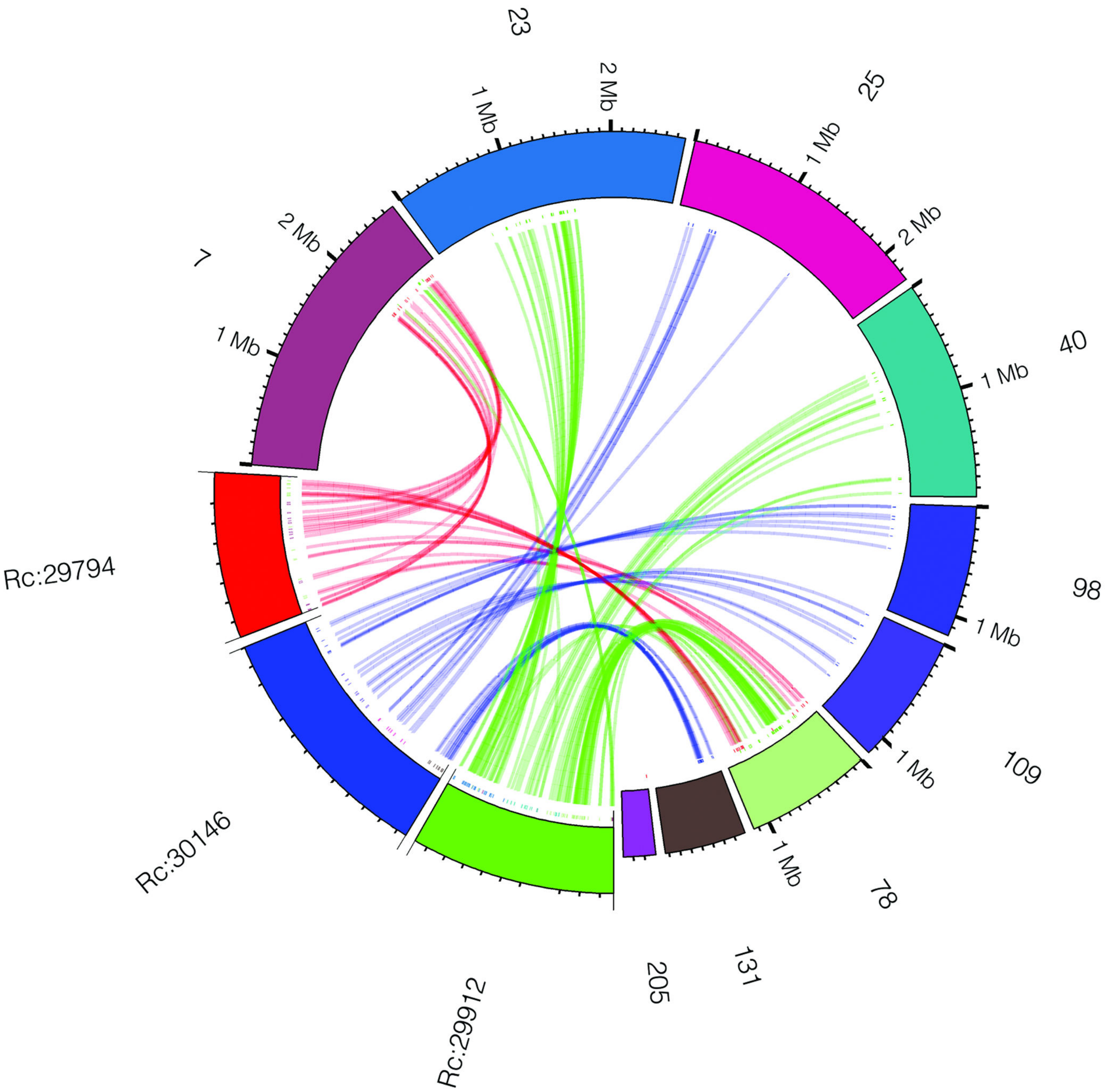


**Figure 2. Collinearity between three paralogous castor bean genomic regions and their putative orthologues in other dicot genomes**

**a**) An example of a conserved paralogous triplication in the castor bean genome. **b–e**) Putative orthologous gene pairs are shown as colored lines connecting the castor bean scaffolds (noted as Rc:scaffold number) to chromosomes or scaffolds in the other dicot genome. In most cases, one copy of the paralogous castor bean genes corresponds to two genes in poplar (**b**), one gene in grapevine (**c**), and four genes in *Arabidopsis* (**d**). The castor bean-papaya relationship (**e**) is inconclusive. Numbers around the circles correspond to linkage group numbers (**b**), chromosome numbers (**c** and **d**), or scaffold numbers (**e**).

Grapevine scaffolds that were mapped to chromosomes but their exact location is unknown are noted with an "r" (random). The size of the castor bean genomic regions is proportional in all circles. Additional castor bean paralogous regions and their corresponding orthologues from other dicots are shown in Supplementary Figure 3.

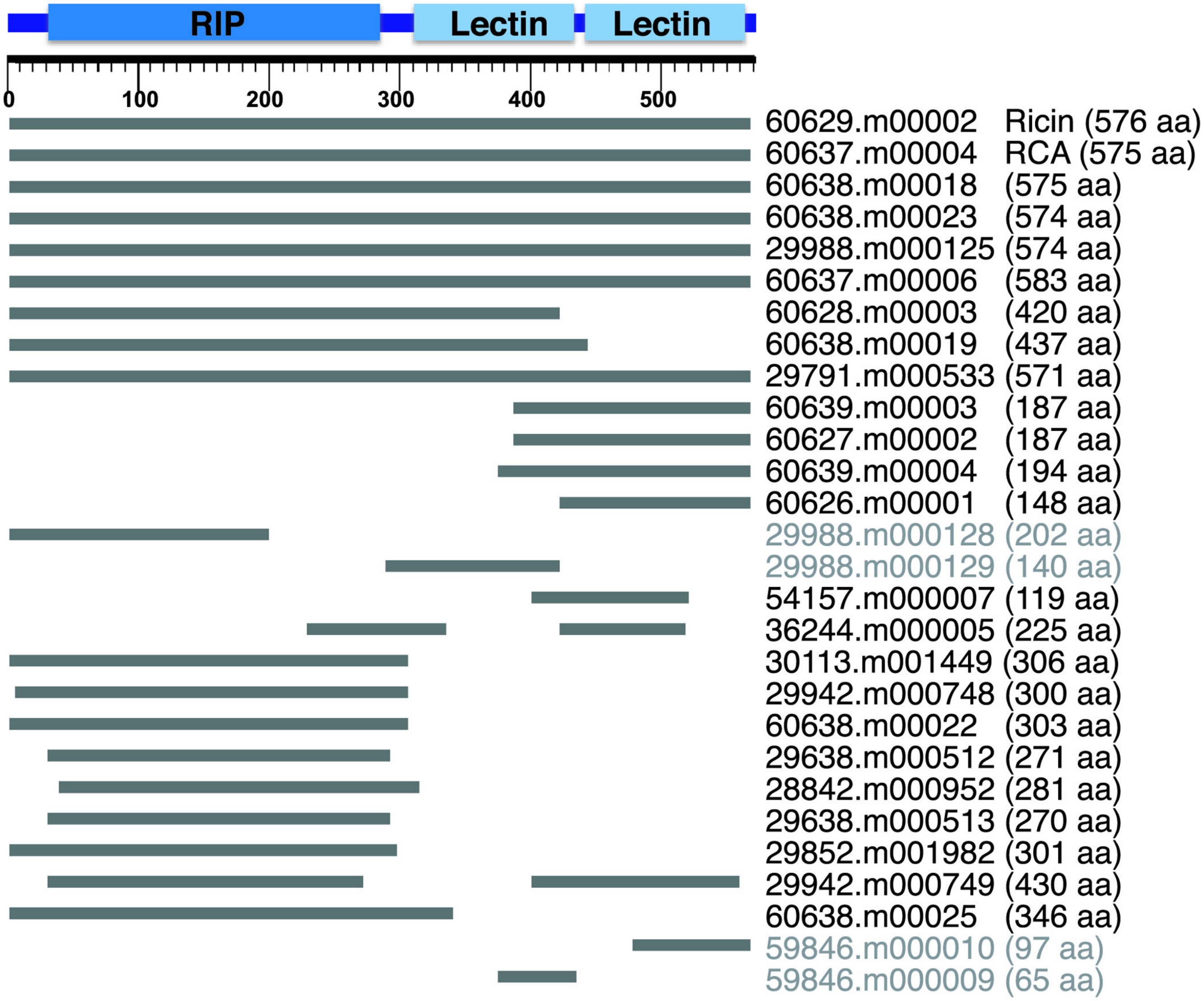


**Figure 3. Schematic representation of the members of the ricin/RCA lectin gene family in castor bean**

Ricin protein domains are represented at the top by blue boxes, and gray boxes represent protein sequences from this gene family aligned to the ricin precursor protein sequence used as reference. The ruler indicates the amino acid coordinates. The ricin and RCA genes are indicated and the amino acid sequence length for each gene model is shown in parenthesis. Pairs of adjacent gene models that could belong to a single pseudogene are shown in gray.

## Table 1

Genome assembly and annotation statistics

| | All scaffolds | Scaffolds longer than 2 kb |
|---|---|---|
| Fold genome coverage | 4.59 | 4.59 |
| Number of scaffolds | 25,828 | 3,500 |
| Total span | 350.6 Mb | 325.5 Mb |
| N50 (scaffolds) | 496.5 kb | 561.4 kb |
| Largest scaffold | 4.7 Mb | 4.7 Mb |
| Average scaffold length | 14 kb | 93 kb |
| Number of contigs | 54,000 | 24,500 |
| Largest contig | 190 kb | 190 kb |
| Average contig length | 6 kb | 13 kb |
| N50 (contigs) | 21,1 kb | |
| GC content | 32.5% | |
| Gene models | 31,237 | |
| Gene density | 11,220 bp/gene | |
| Mean gene length | 2,258.6 bp | |
| Mean coding sequence length | 1,004.2 bp | |
| Longest gene | 15,849 bp | |
| Mean number of exons per gene | 4.2 | |
| Mean exon length | 251 bp | |
| Longest exon | 6,590 bp | |
| GC content in exons | 44.5% | |
| Mean intron length | 381 bp | |
| Longest intron | 33,291 bp | |
| GC content in introns | 31.8% | |
| Mean intergenic region length | 6,846 bp | |
| Longest intergenic region | 691,597 bp | |
| GC content in intergenic regions | 30.7% | |

## Table 2

Classification of repetitive sequences

| | Length occupied (bp) | % of total repeats | % of genome |
|---|---|---|---|
| Retrotransposons | 61,199,930.00 | 36.07 | 18.16 |
| Gypsy | 38,595,566 | 22.75 | 11.45 |
| Copia | 16,078,721 | 9.48 | 4.77 |
| Line | 465,220 | 0.27 | 0.14 |
| Sine | 1,867 | 0.00 | 0.00 |
| Other | 6,058,556 | 3.57 | 1.80 |
| Unclassified elements | 105,387,872 | 62.12 | 31.26 |
| DNA transposons | 3,065,391 | 1.81 | 0.91 |
| Total transposable elements | 169,653,193 | 25.33 | 50.33 |
| Low complexity sequences | 6,348,051 | 0.95 | 1.88 |