ONLINE METHODS

Identification of genome duplications

A total of 167,984 predicted polypeptides from *Ricinus communis*, *Vitis vinifera*, *Populus trichocarpa*, *Arabidopsis thaliana*, and *Carica papaya* were subjected to an all-*vs*.-all BLASTP analysis using WU-BLASTP 2.0MP, with the default BLOSUM62 substitution matrix, no low-complexity sequence filter, and an Evalue cutoff of 10^{-5}. The castor bean subset of the BLAST results was analyzed to extract 5,536 pairs of castor genes that are reciprocal best hits and reside on distinct sequence contigs.

Each of the 721 (of 25,828) castor scaffolds with at least 5 annotated protein-coding genes was examined for runs of 5 or more genes that are collinear and are reciprocal best hits of collinear genes in another castor bean scaffold. Images were generated from these results and inspected manually for the presence of regions that appear to be triplicated in the castor bean genome, on the basis of overlapping runs of collinear matching gene pairs. The regions thus determined were also cross-checked against dot plots showing the relative positions of the paralogous gene pairs.

To further analyze these putative triplications four sets of Jaccard[49] orthologous (protein) clusters[18] were computed between castor and each of the four other genomes: Jaccard clusters were first defined within each genome by taking all BLASTP matches with E value <= 10^{-10}, >= 80% identity, and >=70% sequence coverage and then forming clusters by transitively merging all pairs of proteins with Jaccard coefficient >= 0.6. In the second step, pairs of Jaccard clusters in distinct genomes were merged if each contained a protein with a best hit in the other cluster, taking into consideration only BLASTP matches with E value <=10^{-10} and >=70% sequence coverage (but imposing no other restriction on percent identity). For each triplication, Circos[50] was used to display the three castor regions and any collinear cluster matches between genes in those regions and those in the respective target genomes.

49. Jaccard, P. The distribution of the flora in the alpine zone. *New Phytol.* **11**, 37-50 (1912).

50. Krzywinski, M. *et al.* Circos: an information aesthetic for comparative genomics. *Genome Res.* **19**, 1639-1645 (2009).