RESEARCH ARTICLE

National Science Review
5: 78–87, 2018
doi: 10.1093/nsr/nwx101
Advance access publication 25 August 2017

[1]State Key Laboratory of Plant Genomics and National Center for Plant Gene Research, Institute of Genetics and Developmental Biology, Chinese Academy of Sciences, Beijing 100101, China; [2]Genome Analysis Laboratory of the Ministry of Agriculture, Agricultural Genomics Institute at Shenzhen, Chinese Academy of Agricultural Sciences, Shenzhen 518120, China; [3]The Key Laboratory of Biology and Genetic Resources of Rubber Tree, Rubber Research Institute, Chinese Academy of Tropical Agricultural Sciences, Danzhou 571737, China; [4]Key Laboratory of Plant Molecular Physiology, Institute of Botany, Chinese Academy of Sciences, Beijing 100093, China; [5]Key Laboratory of Biology and Genetic Improvement of Horticultural Crops of the Ministry of Agriculture, Sino-Dutch Joint Laboratory of Horticultural Genomics, Institute of Vegetables and Flowers, Chinese Academy of Agricultural Sciences, Beijing 100081, China and [6]University of the Chinese Academy of Sciences, Beijing 100049, China

*Corresponding authors. E-mails: jyli@genetics.ac.cn; hyu@genetics.ac.cn
†Contributed equally to this work.

Received 28 May 2017; **Revised** 8 August 2017; **Accepted** 14 August 2017

PLANT & ANIMAL SCIENCE

Genome analysis of *Taraxacum kok-saghyz* Rodin provides new insights into rubber biosynthesis

Tao Lin[1,†], Xia Xu[1,†], Jue Ruan[2,†], Shizhong Liu[3], Shigang Wu[2], Xiujuan Shao[2], Xiaobo Wang[2], Lin Gan[3], Bi Qin[3], Yushuang Yang[3], Zhukuan Cheng[1], Suhua Yang[4], Zhonghua Zhang[5], Guosheng Xiong[2], Sanwen Huang[2], Hong Yu[1,*] and Jiayang Li[1,6,*]

ABSTRACT

The Russian dandelion *Taraxacum kok-saghyz* Rodin (TKS), a member of the Composite family and a potential alternative source of natural rubber (NR) and inulin, is an ideal model system for studying rubber biosynthesis. Here we present the draft genome of TKS, the first assembled NR-producing weed plant. The draft TKS genome assembly has a length of 1.29 Gb, containing 46 731 predicted protein-coding genes and 68.56% repeats, in which the LTR-RT elements predominantly contribute to the genome enlargement. We analyzed the heterozygous regions/genes, suggesting its possible involvement in inbreeding depression. Through comparative studies between rubber-producing and non-rubber-producing plants, we found that enzymes of the mevalonate (MVA) pathway and rubber elongation might be critical for rubber biosynthesis, and several key isoforms have been isolated and shown to be predominantly expressed in the latex, indicating their crucial functions in rubber biosynthesis. Moreover, for two important families in rubber elongation, the CPT/CPTL and REF/SRPP families, diverse evolutionary tracks have been revealed. These results provide valuable resources and new insights into the mechanism of NR biosynthesis, and facilitate the development of alternative NR-producing crops.

Keywords: *Taraxacum kok-saghyz*, genomics, rubber biosynthesis, CPT/CPTL, REF/SRPP

INTRODUCTION

Natural rubber (NR) is an isoprenoid polymer synthesized on small dispersed rubber particles in the latex of many plant species. In 2015, the global consumption of NR reached over 12.14 million tons, with a value of ∼17 billion US dollars [1], and demand is still steadily increasing. By now, the Pará rubber tree (*Hevea brasiliensis* L.) is nearly the exclusive source of NR; however, further increase in the yield is severely impaired due to its limited planting area, narrow genetic backgrounds, severe diseases and laborious work requirements [2,3]. Therefore, it is vital to explore an alternative source and a model plant for NR production and research.

There are more than 2500 plant species that could generate NR in the latex, but high-molecular-weight rubber is only identified in a few plants; of these, *Taraxacum kok-saghyz* Rodin (TKS) has drawn special attention since the 1940s [4]. TKS is a perennial plant from the Composite family, originating from the Tian Shan mountain regions in China and Kazakhstan [5], which can be widely planted in high- or low-latitude climates. The root of TKS could produce a large amount of NR (up to ∼20% dry weight), as it shows even higher molecular weight than the Pará rubber tree [6]. In addition, the root of TKS could also produce an abundance of inulin, which can be used for bioethanol production. The advantages of broad planting area, high content and quality of rubber, easy planting and harvest, and short maturation time make TKS an excellent alternative source of NR. Moreover, TKS has a relatively simple genome of an estimated 1.4 Gb [5], forming $2N = 2X = 16$ chromosomes [7], and more importantly it can easily be genetically manipulated. These make

TKS an ideal model plant for NR research. The future of domesticating wild TKS is promising [4], but its high heterozygosity and inbreeding depression due to its self-incompatibility raise serious challenges for generating a complete genome sequence and developing TKS as an NR-producing crop [5].

Understanding the NR biosynthetic pathway is always of great interest. In dandelion and some other rubber-producing plants, several rubber-biosynthesis-related genes have been identified, including *cis*-prenyltransferases (CPTs) that control the NR chain elongation [8,9], small rubber particle proteins (SRPPs) that maintain the stability of the rubber particles [10,11], rubber elongation factors (REF) that improve the CPT activity [12], and rubber transferase activators (RTA or CPTL) that interact with CPTs [13–15]. However, the knowledge is still segmental and the mechanisms of rubber biosynthesis in different rubber-producing plants remain elusive. Therefore, to better understand the biosynthesis and regulation of NR and to accelerate the development of TKS for NR production, here we present the *de novo* genome assembly of TKS line 1151. The draft genome sequence provides new insights into genome evolution, inbreeding depression and genetic mechanisms underlying the biosynthesis of rubber and inulin.

RESULTS

Sequencing, assembly and annotation

We carried out *de novo* sequencing and genome assembly of the TKS line 1151 based on the PacBio RSII platform that generates ∼49.76 Gb of genomic sequence with 48-fold coverage and the Illumina HiSeq 2500 platform that produces ∼60.48 Gb of genomic sequence with 58-fold coverage (see Supplementary Fig. S1 and Supplementary Table S1). We first confirmed the karyotype of TKS as $2N = 2X = 16$ (see Supplementary Fig. S2) and estimated the genome size in the range of 1.04 Gb based on a 19-mer analysis (see Supplementary Fig. S3). This is slightly smaller than the size predicted as ∼1.18 Gb by flow cytometry (see Supplementary Fig. S4). To assist the scaffold construction, we also generated an additional ∼123.31 Gb high-quality mate-pair data with insertion sizes from 5000 to 13 000 bp (see Supplementary Fig. S5 and Supplementary Table S1). Finally, we obtained a 1.29 Gb genome assembly, which contains 19 227 scaffolds with N50 sizes of 100.21 kb and 31 965 contigs with N50 sizes of 47.63 kb (see Table 1 and Supplementary Table S2). We estimated that the average heterozygous rate is 4.17 SNPs (Single Nucleotide Polymorphisms) per kb (see Supplementary Figs S6 and S7, and Supplementary Table S3). The pattern of GC content for the assembled TKS genome is similar to those in *Asteraceae*, such as *Cynara cardunculus* var. *scolymus* (globe artichoke), *Conyza canadensis* (horseweed), and *Helianthus annuus* (sunflower) (see Supplementary Fig. S8). To evaluate the assembled genome, we aligned 248 highly core-eukaryotic genes to the assembled genome, and identified high-confidence hits of 238 (95.97%) with full length and 241 (97.18%) with partial length (see Supplementary Table S4). We also gauged the completeness and accuracy of the assembled TKS genome using the 956 Benchmarking Universal Single-Copy Orthologs (BUSCO) gene set [16], and found complete genes of 91.7% (876) and fragmented genes of 2.6% (25) (see supplementary data online). This evidence indicated the high quality and coverage of the assembled TKS genome.

Table 1. Statistics for the *T. kok-saghyz* genome and gene annotation.

Estimate of genome size	1.04 Gb
Number of scaffolds	19 227
Total length of scaffolds	1.29 Gb
N50 of scaffolds	100.21 kb
Longest scaffold	887.06 kb
Number of contigs	31 966
Total length of contigs	1.28 Gb
N50 of contigs	47.63 kb
Longest contig	499.79 kb
GC content	37.29%
Number of genes	46 731
Percentage of gene length	6.77%
Mean gene length	1850.54 bp
Gene density	36.58 per Mb
Mean coding sequence length	1038.64 bp
Mean exon length	244.11 bp
Exon GC content	44.28%
Mean intron length	249.42 bp
Intron GC content	30.76%
Repeat sequence length	875.81 Mb
Percentage of repeat sequence length	68.56%

To facilitate the gene annotation of the assembled TKS genome and construct the gene expression atlas, we performed RNA-seq with three replicates for latex and other 11 representative tissues, generating a total of ∼401 Gb data (see Fig. 1 and Supplementary Table S5). Using the combined method of *ab initio*, assembled transcripts and protein homologues, we predicted 46 731 protein-coding genes (see Table 1 and Supplementary Fig. S9), with ∼82.90% supported by various public protein-function databases (see Supplementary Table S6). In addition, we also identified a large number of non-coding RNAs, including 162 rRNAs, 836 tRNAs, 265 miRNAs, 22 SRPRNAs, 167 snRNAs, 594 snoRNAs and 214 other ncRNAs (see Supplementary Table S7).

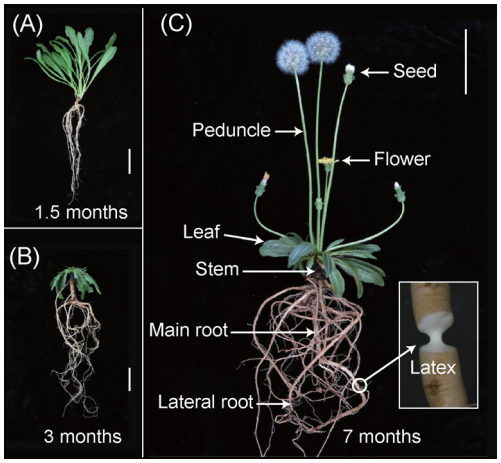


Figure 1. Morphologies of *Taraxacum kok-saghyz* at different growth stages and tissues for transcriptome analysis. Twelve representative tissues were collected for transcriptome analysis, including young and mature leaves, stems and main and lateral roots from plants growing for 1.5 (A) and 3 (B) months, and peduncles, flowers and seeds from plants growing for 7 months (C). All bars = 5 cm.

Globe artichoke is a non-rubber-producing plant and the nearest species to TKS with an assembled genome. A total of 173 genomic synteny regions were found between TKS and globe artichoke, and the largest collinearity region exists between TKS scaffold22 (841 702 bp; 68 genes) and globe artichoke LG9 (18 344 014 bp; 1011 genes), in which 23 genes are matched overall (see Fig. 2A). We also identified two scaffolds (scaffold1261 and scaffold896) containing one CPTL gene and one SRPP gene in the assembled TKS genome (see Supplementary Fig. S10). In the TKS genome assembly, segments on a certain scaffold are collinear to different chromosomes of globe artichoke, indicating that some structural variations occur during genome evolution (see Supplementary Fig. S11). In addition, we found 23 genomic synteny regions and a total length of 2.78 Mb homologous genome sequence between the TKS and *Hevea* genomes. We also identified 51 435 recent segmental duplications with a total length of 26.90 Mb (2.08%) in the assembled TKS genome (see Fig. 2B). However, we should be cautious and these segmental duplications need to be validated by different methods. These genomic results suggest high collinearity with the globe artichoke genome and are useful for studying the TKS evolution and biological functions of rubber biosynthesis.

Heterozygosity

TKS is reported to be self-incompatible [5], which usually leads to high heterozygosity of the genome and results in inbreeding depression, such as the reduced survival and fertility of offspring. When examining the sequencing depth distribution, we found two peaks (see Supplementary Figs S12 and S13) at ∼18 × depth (the first peak) and ∼33 × depth (the second peak), indicating that the genomic regions with the first peak might be underassembled allelic regions with high heterozygosity. A total of 9241 contigs with half of the sequencing depth were identified as heterozygous genomic regions with a total length of ∼306.47 Mb (see Supplementary Fig. S14). These contigs indeed explain the first peak (∼18 ×) for the sequencing depth, as the depth distribution becomes normal after excluding them (see Supplementary Fig. S12). Of these 9241 contigs, 3261 (∼110.68 Mb in length) were independent contigs, each of which constitutes one scaffold, which might be an allelic copy of the other 5980 integrated contigs (see supplementary data online). Therefore, we aligned these independent contigs to the integrated ones using LASTZ [17]. The results showed that 997 independent contigs (24.21 Mb in length) may be the extra allelic copy because they can be aligned to 427 integrated contigs (28.56 Mb in length) with more than 50% identity (see Supplementary Fig. S15). We examined the functions of 301 genes in these independent contigs (see Supplementary Table S8), and they were significantly enriched in carboxylic acid metabolic pathways, methylation, phosphatase complex, and ion binding and transporter activity pathways by GO analysis (see Supplementary Table S9). These observations indicate that these heterozygous regions or genes are possibly related to inbreeding depression caused by homozygous recessive deleterious effects [18].

LTR-RT elements drive the TKS genome enlargement

We identified 875.81 Mb (68.56% of the assembled genome length) as repetitive sequences in the TKS genome (see Table 1 and Supplementary Table S10), which is higher than the 58.4% in globe artichoke [19], 6.25% in horseweed [20] and 54.4% in danshen [21], but slightly lower than the 71.2% in rubber tree [22]. Among the transposable elements (TE), long terminal repeats (LTRs) were found to be the predominant TE family in TKS, accounting for 40.73% (∼520 Mb) of the assembled genome, and *Copia* (∼260.5 Mb, 20.39%) and *Gypsy*-type (∼252.9 Mb, 19.79%) LTRs were the most abundant TE subfamilies. Complete LTR-RTs were identified for four *Asteraceae* species (TKS, globe artichoke, horseweed and sunflower), one *Lamiaceae* species (danshen) and six other model plants using the LTR-FINDER [23] (see Supplementary Fig. S16 and Supplementary Table S11). The TKS genome assembly contains 6154 complete LTR-RTs, which is the

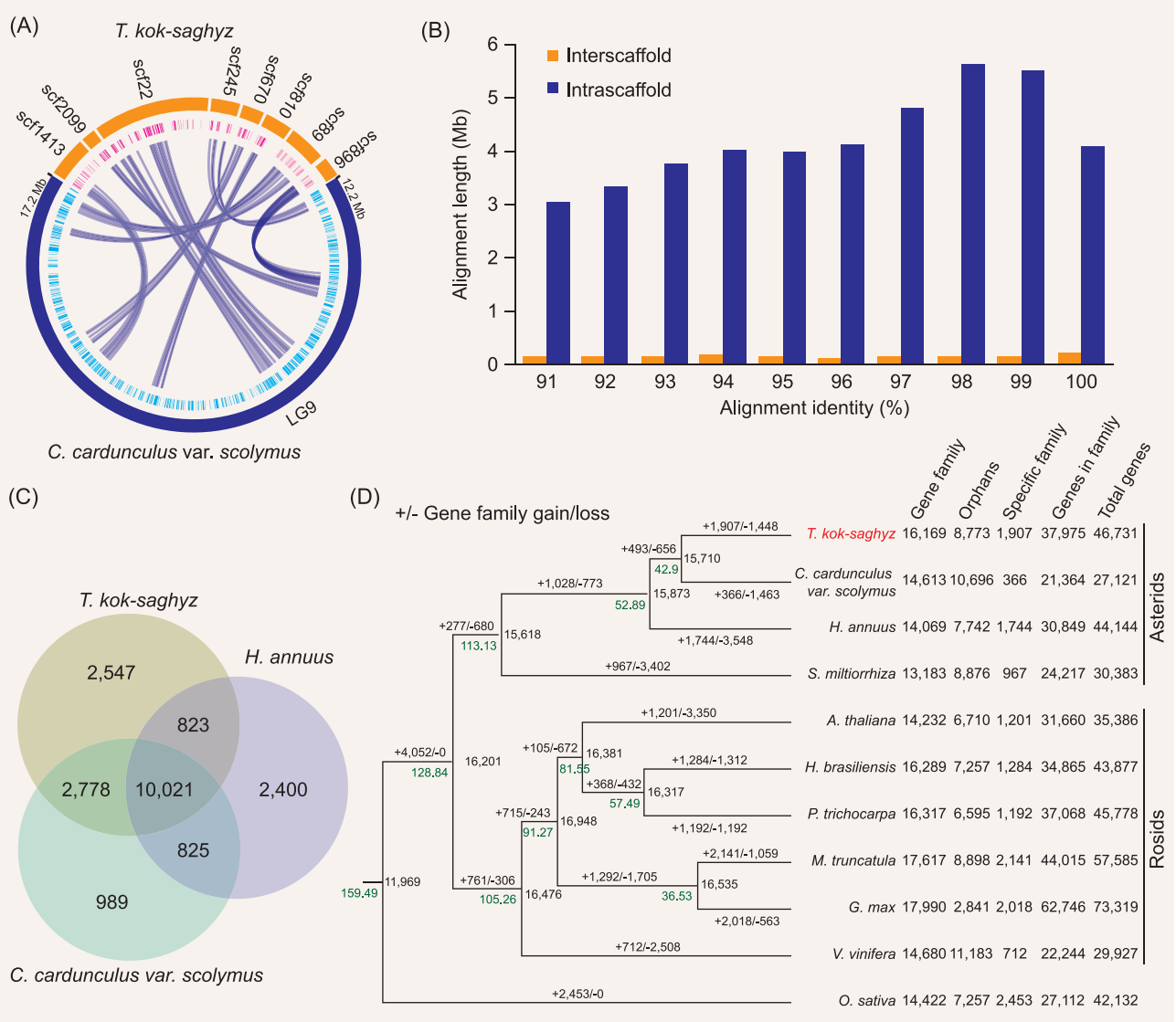


Figure 2. Comparative genomic analysis of *T. kok-saghyz*. (A) Scheme of the largest collinearity region between eight *T. kok-saghyz* scaffolds and *C. cardunculus* var. *scolymus* LG9. The outer circle shows *T. kok-saghyz* scaffolds (orange) and *C. cardunculus* var. *scolymus* LG9 (blue). The inner circle shows the genes corresponding to *T. kok-saghyz* scaffolds (red) and *C. cardunculus* var. *scolymus* LG9 (light blue). The collinear blocks between two species are linked by curved ribbons. (B) Length distribution of the percentage identity of segmental duplication in the *T. kok-saghyz* genome. The *x*-axis represents the identity of sequence alignment, and the *y*-axis represents the total length of the segmental duplications in the corresponding sequence identity. 'Interscaffold' and 'Intrascaffold' show segmental duplications between scaffolds and within scaffolds, respectively. (C) A Venn diagram of gene families among three species. A total of 16 169 gene families were detected in the *T. kok-saghyz* genome, in which 10 021 gene families were shared among the three species, including *T. kok-saghyz*, *C. cardunculus* var. *scolymus* and *H. annuus*. (D) Phylogenetic tree of eleven species based on orthologous of 64 single-gene families. The values above each branch represent the number of gene family gain/loss. The number on the right of each node represents the number of gene families for the common ancestor. The green number to the left of each node represents divergent time (Mya) from the common ancestor. *O. sativa* is used as an outgroup.

highest in all species. The average LTR-RT nucleotide distance (d-value = 0.023) in the TKS genome assembly is significantly lower than the other three *Asteraceae* species (see Supplementary Table S11) with a peak of substitution distance around 0.0025, suggesting that a burst of LTR-RT insertion occurred later than other *Asteraceae* species at ~0.24 Mya (see Supplementary Fig. S17). Taken together, these results suggest that the enlargement of the TKS genome might be driven by LTR insertion.

Genome evolution

Based on protein homologies within and between TKS and ten other species, we identified 16 169

gene families in the assembled TKS genome, 1907 of which were TKS specific and 10 021 were shared with globe artichoke and sunflower (see Fig. 2C). We detected 340 genes in the TKS-specific gene families with tissue-specific expression patterns, and they are significantly enriched in polygalacturonase, responses to stresses or defenses, and cell redox homeostasis by GO analysis (see Supplementary Table S12), which may be associated with biotic and abiotic stress resistance.

A mean family size of 3.9 genes per family was found in 7396 gene families except for 8773 orphan genes, which is at a medium level in the eudicot plants (average size of 3.5 genes per family). To analyze gene family gain–loss events, 64 single-copy gene families were used to construct a maximum likelihood (ML) phylogenetic tree between TKS and ten species (see Fig. 2D). Among these species, TKS has the most gene families in the *Asteraceae* and *Lamiaceae* species. Meanwhile, we detected 3046 gene families that were expanded in TKS compared with globe artichoke, and GO analysis showed significant enrichment ($P < 0.01$) of these genes in protein kinases, ATP binding, macromolecular complex, fatty acid biosynthesis and cell wall organization (see Supplementary Table S13). Notably, we also detected the CPT gene family in these expanded families, suggesting the importance of these expansions in TKS compared to non-rubber-producing plants. The divergence time of the *Asteraceae* species estimated by the nucleotide substitution rate is 52.89 Mya (see Fig. 2D), which is consistent with the 47.5 Mya estimated by the fossil time [24,25].

Rubber and inulin biosynthetic pathways in TKS

NR biosynthesis is initiated by the priming allylic molecules (initiators) and elongated by the *cis*-configuration sequential condensation of isopentenyl pyrophosphate (IPP), an important precursor of NR formed from acetyl-CoA via the cytosolic mevalonate (MVA) pathway or pyruvate and glyceraldehyde 3-phosphate via the plastidic 2-C-methyl-D-erythritol-4-phosphate (MEP) pathway. To investigate rubber biosynthetic pathways of TKS, we performed a homolog analysis of rubber biosynthetic genes (see Supplementary Fig. S18). A total of 102 candidate rubber biosynthesis-related genes were identified in the genome assembly, including 40 genes in 6 processes of the MVA pathway, 23 in 7 processes of the MEP pathway and 19 for initiator synthesis, in which geranyl pyrophosphate (GPP), famesyl pyrophosphate (FPP) and geranylgeranyl pyrophosphate (GGPP) are involved [26]. In addition, 20 genes for 'rubber elongation' on rubber particles were also predicted (see Fig. 3 and Supplementary Table S14). We compared these genes with their homologs in the rubber-producing species (*Hevea*) and non-rubber-producing species (globe artichoke), and found that at least one enzyme exists in each process in all three species, and the gene number is similar for enzymes in the MEP pathway and rubber initiator synthesis, but differs in the MVA pathway and rubber elongation (see Supplementary Table S15). These results suggest that the MVA pathway and rubber elongation-related enzymes might be more important for rubber biosynthesis during genome evolution. Consistent with this, for each process in the MVA pathway in TKS, at least one enzyme shows predominant expression in latex and roots (see Fig. 3), but most genes in the MEP pathway have a medium or low expression level in latex. In *Hevea* it is also reported that the MVA pathway shows latex-biased abundant expression [22,27,28]. These data suggest that the different activities and spatiotemporal expression of these related enzymes may be crucial for their function in different species, and the MVA pathway, rather than the MEP pathway, might be the main source of IPP for rubber biosynthesis in both roots of TKS and inner bark of *Hevea*. Based on this, we found that the two genes, *TkHMGR1* and *TkHMGR2*, which encode 3-hydroxy-3-methylglutaryl-coenyme, a reductase that catalyzes the key step in the MVA pathway, were predominantly expressed in roots, with the highest expression level in latex. We further found that the *TkCPT1*, *TkCPT2* and *TkCPTL1* genes may play a critical role for the elongation of rubber polymers because they encompass the highest expression levels in latex and roots and are homologous with that in *T. brevicorniculatum*. Furthermore, seven of the nine *TkSRPP* genes were found highly expressed in latex and roots. These data demonstrate that the complete genome information together with the gene expression profile could facilitate the identification and isolation of key genes of the rubber biosynthetic pathway.

In addition to NR, abundant secondary metabolites, especially inulin, are also important economic outcomes of TKS. In the assembled genome, 15 candidate genes encoding the proteins required for biosynthesis and metabolism of inulin were detected (see Supplementary Fig. S19), including six sucrose transporters (SUCs), two fructosyltransferases, and seven fructan 1-exohydrolases (1-FEHs). Both the two essential genes encoding sucrose:sucrose 1-fructoslytransferase (1-SST) and fructan:fructan 1-fructosyltransferase (1-FFT) were found strongly expressed in roots, but most of the *1-FEH* genes were not highly expressed.

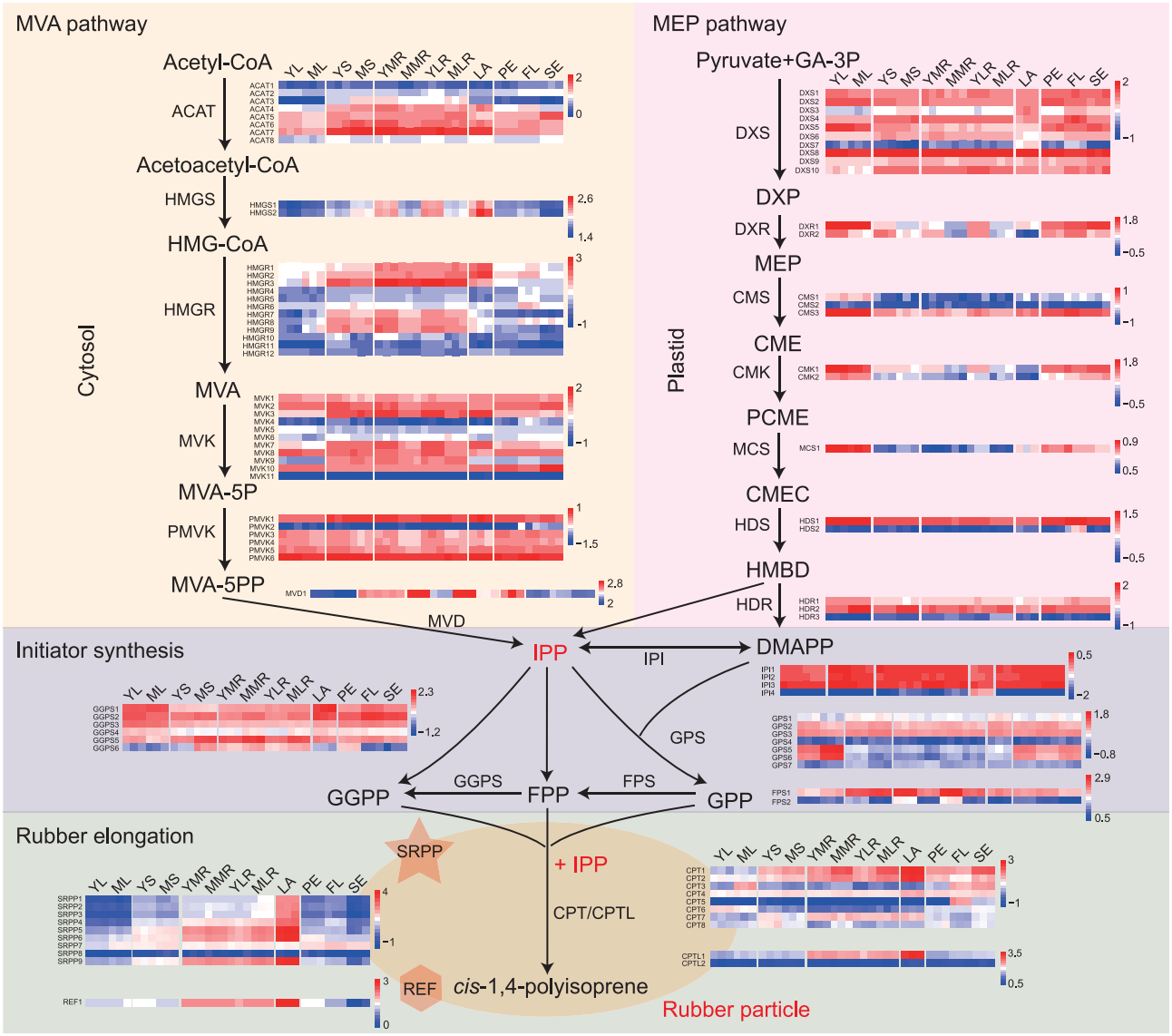
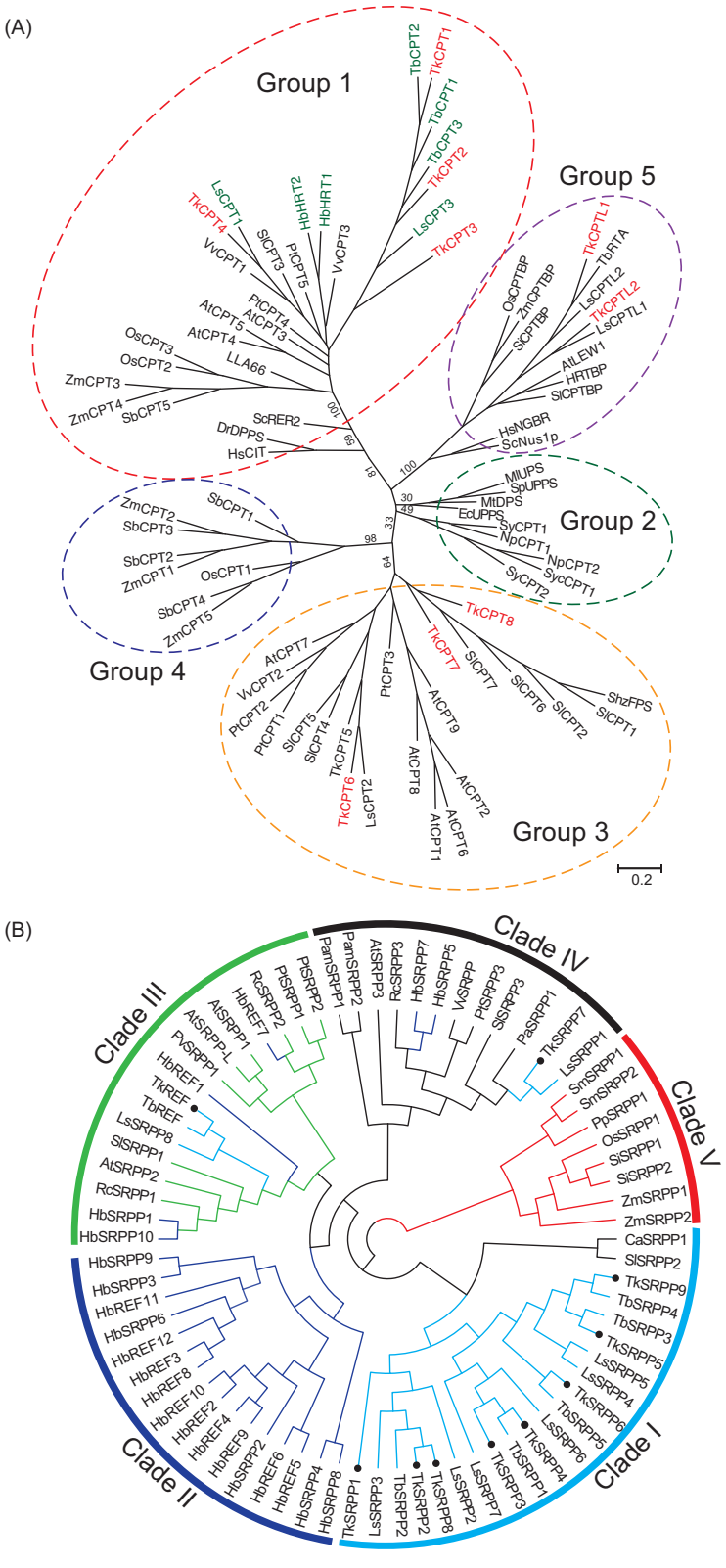


Figure 3. Rubber biosynthetic pathway in *T. kok-saghyz*. Expression levels for each gene in latex and eleven other tissues were shown by heatmap using $\log_{10}$(RPKM). RPKM, reads per kilobase per million mapped reads. YL, young leaf; ML, mature leaf; YS, young stem; MS, mature stem; YMR, young main root; MMR, mature main root; YLR, young lateral root; MLR, mature lateral root; LA, latex; PE, peduncle; FL, flower; SE, seed.

The *CPT/CPTL* gene family in TKS

We identified eight *TkCPT* and two *TkCPTL* genes for 'rubber elongation' on rubber particles in the assembled TKS genome. Comparative analysis revealed that the *CPT* family genes possess five highly conserved regions, Region I to Region V (see Supplementary Fig. S20), in which the conserved Asp residue in Region I and the five conserved residues (F93, S94, R243, R249 and E307) in Regions II to V are essential for catalysis and substrate recognition in the rubber elongation, respectively [29,30]. In addition, the eight *TkCPT* genes could be further classified into two types (see Supplementary Figs S20 and S21), in which type I is highly conserved between Regions III and IV, whereas type II is diverse from them.

The phylogenetic tree divides the *CPT/CPTL* gene family from eighteen species into five groups (see Fig. 4A), and Group 3 and Group 4 bear dicot- and monocot-specific *CPT* genes, respectively. All four type I *CPT* genes in the assembled TKS genome were situated within Group 1, which also contains the genes of three *T. brevicorniculatum CPT*s, two lettuce *CPT*s and two *Hevea CPT*s (see Fig. 4A), suggesting that this clade is very important for NR biosynthesis and the rubber chain elongation pathway is conserved among different rubber-producing species. Furthermore, we found that both *TkCPT1*



and *TkCPT2* are predominantly expressed in latex (see Fig. 3), indicating that they may be responsible for rubber biosynthesis in TKS. Interestingly, among the four type II TKS *CPT* genes in Group 3, *TkCPT7* is highly expressed in roots and latex, but it contains deletions of conserved F93 and S94 residues and a substitution of R for S at the 249 position, suggesting that *TkCPT7* may have other diverse functions in roots and latex.

Recently, the study of *T. brevicorniculatum* [13], lettuce [15] and *Hevea* [14] has shown that *CPTL* genes are essential for rubber biosynthesis. We observed a distinct clade (Group 5) that includes 13 *CPTLs* from plants and human/yeast (see Fig. 4A), indicating that the divergence of *CPT* and *CPTL* genes is prior to the split of the plant and animal lineages. The protein sequences of CPTLs have low homology to the CPTs, and only a part of three conserved regions (III, IV and V) were found in the CPTLs (see Supplementary Fig. S22), suggesting that these CPTLs may have lost the catalytic functions. Indeed, in lettuce, *LsCPTL2* shows no catalytic activity *in vitro* [15]. Meanwhile, two other conserved regions were detected in the CPTLs but were lost in CPTs; these were predicted to be a transmembrane domain [31]. In TKS, *TkCPTL1* was expressed highest in latex, but its homologous *TkCPTL2* showed low expression levels in all tissues (see Fig. 3). Further research on *TkCPTL1* will provide new insights in understanding the NR biosynthetic pathway.

Figure 4. Phylogenetic analysis of predicted REF/SRPP and CPT/CPTL gene families from TKS and other species. (A) Identification of the TKS CPT/CPTL gene family. A total of 80 CPT/CPTL proteins from rubber-producing and non-rubber-producing plants, as well as representative CPT proteins from bacteria, yeast and animal species were classified into five groups. Genes in red indicate CPT and CPTL from the TKS genome and genes in green are reported functional CPT genes in different rubber-producing plants. Branch lengths represent the number of amino acid substitutions per position. (B) A total of 74 REF/SRPP proteins were assigned into five clades, which are labeled with different colors, including eudicots (Clades I, II, III and IV) and monocots (Clade V). Black dots represent SRPP and REF genes from TKS. Tk, *Taraxacum kok-saghyz*; Tb, *Taraxacum brevicorniculatum*; Ls, *Lactuca sativa*; Ca, *Capsicum annuum*; Sl, *Solanum lycopersicum*; Hb, *Hevea brasiliensis*; At, *Arabidopsis thaliana*; Rc, *Ricinus communis*; Pv, *Phaseolus vulgaris*; Pt, *Populus trichocarpa*; Pam, *Persea americana*; Vv, *Vitis riparia*; Pa, *Parthenium argentatum*; Sm, *Selaginella moellendorffii*; Pp, *Physcomitrella patens*; Os, *Oryza sativa*; Si, *Setaria italica*; Zm, *Zea mays*. Dr, *Danio rerio*; Ec, *Escherichia coli*; Hs, *Homo sapiens*; Mt, *Mycobacterium tuberculosis*; Np, *Nostoc punctiforme*; Sc, *Saccharomyces cerevisiae*; Sp, *Streptococcus pneumoniae*; Sy, *Synechococcus*; Syc, *Synechocystis*.

The *REF/SRPP* gene family in TKS

REF/SRPP family proteins are important components for the biogenesis and stability of rubber particles. In the assembled TKS genome, we detected a family consisting of one *TkREF* and nine *TkSRPP* genes. Based on the phylogenetic tree of REF/SRPP proteins from rubber-producing and non-rubber-producing plants, we found that these proteins could be assigned into five clades (see Fig. 4B), including Clades I, II, III and IV from the eudicots, and Clade V from the monocots and bryophyte/lycophyte, suggesting that the *REF/SRPP* genes of Asterids and Rosids may originate from a common ancestor. In the eudicots, some *REF/SRPP* genes from rubber-producing plants were interspersed with other SRPP-like genes from non-rubber-producing plants. However, we found that most of the TKS and *Hevea SRPP* genes belong to two separate clades, which may suggest that the mechanism of rubber particle stabilization is different in the two species because of independent evolution in rubber biosynthesis. Similar results were also observed in previous studies [22,32].

Most of the *REF/SRPP* genes in the assembled TKS genome exhibit specific high expression levels in latex and roots (see Fig. 3). Four SRPP isoforms (*TkSRPP1*, *TkSRPP2*, *TkSRPP3* and *TkSRPP4*) have considerably higher RPKM values (>140) in latex than in other tissues (RPKM values of ∼6.23 on average), and another four isoforms have strikingly high RPKM values of >100 in roots and >2700 in latex, which is over 55 times higher than those in non-root tissues (RPKM values of ∼11.65 on average), suggesting that these *REF/SRPP* genes may play important roles in the biosynthesis of NR in TKS.

DISCUSSION

In this study, we present a draft sequence of *T. kok-saghyz*, which is a potential alternative resource for NR production and an ideal model system for NR research. TKS is the first assembled genome sequence of a rubber-producing wild weed plant. The advent of the TKS genome will facilitate and accelerate understanding of the molecular mechanisms underlying rubber biosynthesis by the cloning of key genes, genome-wide association studies and genetic improvement of TKS breeding. The TKS genome consists of a higher proportion of repetitive content (∼70%) and heterozygous rate (∼1%), making it difficult to assemble the genome. Indeed, using the Illumina sequencing data, we only obtained ∼4.2 kb length of the contig N50. Similar difficulties were also observed for the barley genome, containing ∼84% repetitive DNA, and the rubber tree genome, containing ∼78% repetitive DNA, the assembled genomes of which had ∼1.4 kb and ∼3.0 kb lengths of the contig N50, respectively [33,34]. Recently, in several species, the PacBio sequencing data has helped to greatly improve the contiguity compared to the Illumina sequencing data; e.g. the contig N50 length of the sunflower genome increased from 21–25 kb to 399 kb (∼20-fold) [35], and the contig N50 length of the maize genome increased from 40 kb to 1.2 Mb (∼30-fold) [36]. We therefore used long PacBio reads to assemble the TKS genome, and produced a much better assembled genome, which achieved a greater than 13-fold increase in the contig N50 length and more complete LTRs than other model plants. Furthermore, the PacBio assembly also helped us to get a fine gene set, which covered 91.7% genes with full length in BUSCO evaluation. However, further improvement of the TKS genome assembly is still needed. Next-generation DNA sequencing technologies offer new and powerful tools for the construction of genetic linkages. In particular, a high-density genetic linkage map is an effective strategy to validate assemblies and assign short sequences to chromosomal linkage groups. Therefore, the construction of a high-resolution genetic linkage map to improve the assembled TKS genome will be an important goal in future.

Based on the comparative analysis of *CPT/CPTL* and *REF/SRPP* gene families, we found several clues for common and unique evolution and mechanisms for rubber biosynthesis in rubber-producing plants. First, the divergence of *CPT* and *CPTL* genes is prior to the split of the plant and animal lineages, and for rubber-producing plants the *CPT* genes were clustered into two groups by their functions rather than species (see Fig. 4A), suggesting a similar mechanism for the rubber chain elongation in these species. Second, we detected very high sequence identity (more than 98%) among REF/SRPP homologs from TKS and *T. brevicorniculatum* (see Fig. 4B), indicating their close relatives between genomes [37]. Third, no expansion of the *REF/SRPP* gene family was found in the assembled TKS genome, which is different from the *Hevea* genome [22]. Similarly, the phylogenetic analysis of *REF/SRPP* gene family showed that most of the *TkSRPP*, *TbSRPP* and *LsSRPP* genes belong to the same clade, but most of the *Hevea SRPP* genes are assigned to a different clade, suggesting that the gene function and mechanism of rubber particle stabilization might be diversified during evolution between TKS and *Hevea*. These evolutionary traces raise new research interests and require further molecular and genetic evidence. For example, the rubber content in the *TbREF-RNAi* plant was lower than that in the wild type [12], but silencing its highest homologous gene *LsSRPP8* in lettuce showed

no effect on its rubber biosynthesis [38], indicating different molecular mechanisms of rubber particle stabilization in these two species. The assembled genome of *T. kok-saghyz* will provide fundamental information to facilitate the molecular dissection of NR biosynthesis and regulation, contribute to a better understanding of the adaptation of TKS to environmental stresses and provide a more powerful tool to develop a new crop in producing NR.

METHODS

The dandelion (*Taraxacum kok-saghyz* Rodin) line 1151 was used for genome sequencing and assembly. The genome assembly was sequenced using the whole-genome shotgun approach with PacBio RSII platforms and assembled with smartdenovo. Detailed information on materials, sequencing, assembly, annotation, and genome and transcriptome analyses is given in the supplementary data online. The *T. kok-saghyz* genome, gene annotation, and nucleotide and protein sequences were deposited in the Genome Warehouse (GWH; http://bigd.big.ac.cn/gwh/) under the accession number PRJCA000437.

ACKNOWLEDGEMENT

We thank Prof. Fangpu Han (Institute of Genetics and Developmental Biology, Chinese Academy of Sciences) for providing maize B73 seeds and Prof. Chengzhi Liang for providing the computing platform. T.L., X.X. and H.Y. designed and performed experiments, analyzed data and wrote the paper. T.L., X.X., J.R., S.W., X.S., X.W., Z.Z., G.X. and S.H. analyzed data. S.L., L.G., B.Q., Y.Y., Z.C. and S.Y. performed experiments. J.L. and H.Y. conceived and supervised the project, designed research, analyzed data and wrote the paper.

FUNDING

This work was supported by grants from the Institute of Genetics and Developmental Biology, National Postdoctoral Program for Innovative Talents (BX201600191) and State Key Laboratory of Plant Genomics.

Conflict of interest statement. None declared.

SUPPORTING INFORMATION

Supplementary data are available at *NSR* online.

Editor's note: the commentaries from recommender and reviewers can be referred to:
doi: 10.1093/nsr/nwx101a
doi: 10.1093/nsr/nwx101b
doi: 10.1093/nsr/nwx101c

REFERENCES

1. IRSG. *International Rubber Study Group: Statistical Summary of World Rubber Situation*. http://www.rubberstudy.com/statistics.aspx (10 December 2016, date last accessed).
2. Clément-Demange A, Priyadarshan PM and Thuy Hoa TT *et al.* Hevea rubber breeding and genetics. In: Jules Janick (ed). *Plant Breeding Reviews*. Wiley, 2007, 177–283.
3. Nair KPP (ed). Rubber (*Hevea brasiliensis*). In: *The Agronomy and Economy of Important Tree Crops of the Developing World*. Amsterdam: Elsevier Science, 2010, 237–73.
4. Whaley WG and Bowen JS. Russian dandelion (kok-saghyz). An emergency source of natural rubber. *Misc Publ US Dept Agric* 1947; **618**: 1–212.
5. Kirschner J, Štěpánek J and Černý T *et al.* Available *ex situ* germplasm of the potential rubber crop *Taraxacum koksaghyz* belongs to a poor rubber producer, *T. brevicorniculatum* (Compositae–Crepidinae). *Genet Resour Crop Evol* 2012; **60**: 455–71.
6. van Beilen JB, and Poirier Y. Establishment of new crops for the production of natural rubber. *Trends Biotechnol* 2007; **25**: 522–9.
7. Luo S, Feng W, and Wu X. The study of the *Taraxacum koksaghyz* Rodin. *Sci China* 1951; **2**: 373–9.
8. Post J, van Deenen N, and Fricke J *et al.* Laticifer-specific cis-prenyltransferase silencing affects the rubber, triterpene, and inulin content of *Taraxacum brevicorniculatum*. *Plant Physiol* 2012; **158**: 1406–17.
9. Schmidt T, Hillebrand A, and Wurbs D *et al.* Molecular cloning and characterization of rubber biosynthetic genes from *Taraxacum koksaghyz*. *Plant Mol Biol Rep* 2009; **28**: 277–84.
10. Hillebrand A, Wurbs D, and Post J *et al.* Down-regulation of small rubber particle protein expression affects integrity of rubber particles and rubber content in *Taraxacum brevicorniculatum*. *PLoS One* 2012; **7**: e41874.
11. Schmidt T, Lenders M, and Hillebrand A *et al.* Characterization of rubber particles and rubber chain elongation in *Taraxacum koksaghyz*. *BMC Biochem* 2010; **11**: 11.
12. Laibach N, Hillebrand A, and Twyman RM *et al.* Identification of a *Taraxacum brevicorniculatum* rubber elongation factor protein that is localized on rubber particles and promotes rubber biosynthesis. *Plant J* 2015; **82**: 609–20.
13. Epping J, van Deenen N, and Niephaus *et al.* A rubber transferase activator is necessary for natural rubber biosynthesis in dandelion. *Nat Plants* 2015; **1**: 15048.
14. Yamashita S, Yamaguchi H, and Waki T *et al.* Identification and reconstitution of the rubber biosynthetic machinery on rubber particles from *Hevea brasiliensis*. *eLife* 2016; **5**: e19022.
15. Qu Y, Chakrabarty R, and Tran HT *et al.* A lettuce (*Lactuca sativa*) homolog of human Nogo-B receptor interacts with *cis*-prenyltransferase and is necessary for natural rubber biosynthesis. *J Biol Chem* 2015; **290**: 1898–914.
16. Simao FA, Waterhouse RM, and Ioannidis P *et al.* BUSCO: assessing genome assembly and annotation completeness with single-copy orthologs. *Bioinformatics* 2015; **31**: 3210–12.
17. Zhang Z, Berman P, and Miller W. Alignments without low-scoring regions. *J Comput Biol* 1998; **5**: 197–210.
18. Charlesworth D, and Willis JH. The genetics of inbreeding depression. *Nat Rev Genet* 2009; **10**: 783–96.
19. Scaglione D, Sebastian R, and Alberto A *et al.* The genome sequence of the outbreeding globe artichoke constructed *de novo* incorporating a phase-aware low-pass sequencing strategy of F1 progeny. *Sci Rep* 2016; **6**: 19427.

20. Peng Y, Lai Z, and Lane T *et al*. De novo genome assembly of the economically important weed horseweed using integrated data from multiple sequencing platforms. *Plant Physiol* 2014; **166**: 1241–54.
21. Xu H, Song J, and Luo H *et al*. Analysis of the genome sequence of the medicinal plant *Salvia miltiorrhiza*. *Mol Plant* 2016; **9**: 949–52.
22. Tang C, Yand M, and Fang Y *et al*. The rubber tree genome reveals new insights into rubber production and species adaptation. *Nat Plants* 2016; **2**: 16073.
23. Xu Z, and Wang H. LTR FINDER: an efficient tool for the prediction of full-length LTR retrotransposons. *Nucleic Acids Res* 2007; **35**: W265–8.
24. Barreda VD, Palazzesil L, and Tellería MC *et al*. Eocene patagonia fossils of the daisy family. *Science* 2010; **329**: 1621.
25. Stuessy T. The rise of sunflowers. *Science* 2010; **329**: 1605–6.
26. Cornish K. The separate roles of plant *cis* and *trans* prenyl transferases in *cis*-1,4-polyisoprene biosynthesis. *Eur J Biochem* 1993; **218**: 267–71.
27. Sando T, Takeno S, and Watanabe N *et al*. Cloning and characterization of the 2- *C*-methyl- D-erythritol 4-phosphate (MEP) pathway genes of a natural-rubber producing plant, *Hevea brasiliensis*. *Biosci Biotechnol Biochem* 2014; **72**: 2903–17.
28. Sando T, Takaoka C, and Mukai Y *et al*. Cloning and characterization of Mevalonate pathway genes in a natural rubber producing plant, *Hevea brasiliensis*. *Biosci Biotechnol Biochem* 2014; **72**: 2049–60.
29. Kharel Y, and Koyama T. Molecular analysis of *cis*-prenyl chain elongating enzymes. *Nat Prod Rep* 2003; **20**: 111–8.
30. Takahashi S, and Koyama T. Structure and function of *cis*-prenyl chain elongating enzymes. *Chem Rec* 2006; **6**: 194–205.
31. Brasher MI, Surmacz L, and Leong B *et al*. A two-component enzyme complex is required for dolichol biosynthesis in tomato. *Plant J* 2015; **82**: 903–14.
32. Horn PJ, James CN, and Gidda SK *et al*. Identification of a new class of lipid droplet-associated proteins in plants. *Plant Physiol* 2013; **162**: 1926–36.
33. The International Barley Genome Sequencing Consortium. A physical, genetic and functional sequence assembly of the barley genome. *Nature* 2012; **491**: 711–6.
34. Rahman AY, Usharraj AO, and Misra BB *et al*. Draft genome sequence of the rubber tree *Hevea brasiliensis*. *BMC Genomics* 2013; **14**: 75.
35. Badouin H, Gouzy J, and Grassa CJ X *et al*. The sunflower genome provides insights into oil metabolism, flowering and Asterid evolution. *Nature* 2017; **546**: 148–52.
36. Jiao Y, Peluso P, and Shi J *et al*. Improved maize reference genome with single-molecule technologies. *Nature* 2017; **546**: 524–7.
37. Zhang Y, Iaffaldano BJ, and Zhuang X *et al*. Chloroplast genome resources and molecular markers differentiate rubber dandelion species from weedy relatives. *BMC Plant Biol* 2017; **17**: 34.
38. Chakrabarty R, Qu Y, and Ro D-K. Silencing the lettuce homologs of small rubber particle protein does not influence natural rubber biosynthesis in lettuce (*Lactuca sativa*). *Phytochem* 2015; **113**: 121–9.