Genome Biology

Genome sequencing of the high oil crop sesame provides insight into oil biosynthesis

Genome Biology 2014, **15**:R39 doi:10.1186/gb-2014-15-2-r39

Linhai Wang (linhai827@163.com)
Sheng Yu (yusheng@genomics.cn)
Chaobo Tong (tongchaobo@gmail.com)
Yingzhong Zhao (zhaoyz63@163.com)
Yan Liu (zpteacher@hotmail.com)
Chi Song (songchi@genomics.cn)
Yanxin Zhang (zhyx026@yahoo.com.cn)
Xudong Zhang (zhangxudong2@genomics.cn)
Ying Wang (wangying2@genomics.cn)
Wei Hua (huawei@oilcrops.cn)
Donghua Li (ldh360681@163.com)
Dan Li (lidan4@genomics.cn)
Fang Li (lifang@genomics.cn)
Jingyin Yu (yujyinfor@gmail.com)
Chunyan Xu (xuchunyan@genomics.cn)
Xuelian Han (hanxuelian@genomics.cn)
Shunmou Huang (shunmouh@yahoo.com.cn)
Shuaishuai Tai (taishuaishuai@genomics.cn)
Junyi Wang (wangjunyi@genomics.cn)
Xun Xu (xuxun@genomics.cn)
Yingrui Li (liyingrui@genomics.cn)
Shengyi Liu (liusy@oilcrops.cn)
Rajeev K Varshney (R.K.Varshney@CGIAR.ORG)
Jun Wang (wangj@genomics.cn)
Xiurong Zhang (zhangxr@oilcrops.cn)

ISSN 1465-6906

Article type Research

Submission date 17 July 2013

Acceptance date 24 January 2014

Publication date 27 February 2014

Genome sequencing of the high oil crop sesame provides insight into oil biosynthesis

Linhai Wang[1,†]
Email: linhai827@163.com

Sheng Yu[2,†]
Email: yusheng@genomics.cn

Chaobo Tong[1,†]
Email: tongchaobo@gmail.com

Yingzhong Zhao[1,†]
Email: zhaoyz63@163.com

Yan Liu[4,†]
Email: zpteacher@hotmail.com

Chi Song[2]
Email: songchi@genomics.cn

Yanxin Zhang[1]
Email: zhyx026@yahoo.com.cn

Xudong Zhang[2]
Email: zhangxudong2@genomics.cn

Ying Wang[2]
Email: wangying2@genomics.cn

Wei Hua[1]
Email: huawei@oilcrops.cn

Donghua Li[1]
Email: ldh360681@163.com

Dan Li[2]
Email: lidan4@genomics.cn

Fang Li[2]
Email: lifang@genomics.cn

Jingyin Yu[1]
Email: yujyinfor@gmail.com

Chunyan Xu[2]
Email: xuchunyan@genomics.cn

Xuelian Han[2]
Email: hanxuelian@genomics.cn

Shunmou Huang[1]
Email: shunmouh@yahoo.com.cn

Shuaishuai Tai[2]
Email: taishuaishuai@genomics.cn

Junyi Wang[2]
Email: wangjunyi@genomics.cn

Xun Xu[2]
Email: xuxun@genomics.cn

Yingrui Li[2]
Email: liyingrui@genomics.cn

Shengyi Liu[1*]
[*] Corresponding author
Email: liusy@oilcrops.cn

Rajeev K Varshney[5,6,*]
[*] Corresponding author
Email: R.K.Varshney@CGIAR.ORG

Jun Wang[2,3,*]
[*] Corresponding author
Email: wangj@genomics.cn

Xiurong Zhang[1*]
[*] Corresponding author
Email: zhangxr@oilcrops.cn

[1] Oil Crops Research Institute of the Chinese Academy of Agricultural Sciences, Key Laboratory of Biology and Genetic Improvement of Oil Crops of the Ministry of Agriculture, Wuhan 430062, China

[2] Beijing Genomics Institute (BGI) − Shenzhen, Shenzhen, China

[3] Department of Biology, University of Copenhagen, Copenhagen, Denmark

[4] Yanzhuang oil CO., LTD, Hefei, China

[5] International Crops Research Institute for the Semi − Arid Tropics (ICRISAT), Patancheru, India

[6] CGIAR Generation Challenge Programme (GCP), c/o CIMMYT, Mexico, DF, Mexico

[†] Equal contributors.

Abstract

Background

Sesame, *Sesamum indicum* L., is considered the queen of oilseeds for its high oil content and quality, and is grown widely in tropical and subtropical areas as an important source of oil and protein. However, the molecular biology of sesame is largely unexplored.

Results

Here, we report a high-quality genome sequence of sesame assembled *de novo* with a contig N50 of 52.2 kb and a scaffold N50 of 2.1 Mb, containing an estimated 27,148 genes. The results reveal novel, independent whole genome duplication and the absence of the Toll/interleukin-1 receptor domain in resistance genes. Candidate genes and oil biosynthetic pathways contributing to high oil content were discovered by comparative genomic and transcriptomic analyses. These revealed the expansion of type 1 lipid transfer genes by tandem duplication, the contraction of lipid degradation genes, and the differential expression of essential genes in the triacylglycerol biosynthesis pathway, particularly in the early stage of seed development. Resequencing data in 29 sesame accessions from 12 countries suggested that the high genetic diversity of lipid-related genes might be associated with the wide variation in oil content. Additionally, the results shed light on the pivotal stage of seed development, oil accumulation and potential key genes for sesamin production, an important pharmacological constituent of sesame.

Conclusions

As an important species from the order Lamiales and a high oil crop, the sesame genome will facilitate future research on the evolution of eudicots, as well as the study of lipid biosynthesis and potential genetic improvement of sesame.

Background

Sesame (*Sesamum indicum*), a widely grown crop in tropical and subtropical areas, is documented as the most ancient oil crop that provides humans with their essential daily energy. Vegetable oil consumption is expected to reach almost 200 billion kilograms by 2030 [1], which will increase the demand for oil-rich crops; genetic studies to improve oil content in vegetable will help address this demand. Compared to other edible oil crops such as soybean (*Glycine max*), rapeseed (*Brassica napus*), peanut (*Arachis hypogaea*) and olive (*Olea europaea*), sesame has innate superiority for its high oil content (~55% of dry seed) [2], and thus is an attractive potential model for studying lipid biosynthesis [3].

The sesame seed had been considered the “queen of oilseeds” for its high oil content and quality [4], and has been traditionally categorised it as a health foods in China, Japan and other East Asian countries [5]. The antioxidative furofuran lignans in sesame has been analysed by pharmacologists for their potent pharmacological properties in decreasing blood lipids [6] and lowering cholesterol levels [7]. Recently, the gene encoding sesamin synthase was identified [8].

Taxonomically, sesame belongs to Lamiales, an order comprising of 23,810 flowering plants in the clade asterids [9]. Lamiales includes many other well-known or economically important species, such as olive (*Olea europaea*), leonurus (*Leonurus japonicas*), lavender (*Lavandula spica*) and basil (*Ocimum basilicum*). However, intensive genetic or genomics study are lacking in most Lamiales species. As high-throughput sequencing has become routine, several studies in sesame have been performed. A number of simple sequence repeat (SSR) markers have been developed [10-12]. The loci associated with indehiscent capsule trait, determinate growth habit and seed coat colour have been detected respectively [13-15], and the expression levels of sesame genes have been explored using Sanger and high-throughput DNA sequencing technologies [16-18]. The phylogenetic position of sesame has been determined using chloroplast genomic data, which indicated the core lineage of *Sesamum* in Lamiales [19]. These studies along with the recently published minute genome of *Utricularia gibba* and the genome survey of sesame have contributed new insight into the Lamiales [20,21].

Here, we report a high-quality draft genome of the sesame genotype ‘Zhongzhi No. 13’, an elite cultivar with high oil content (59%), which has been introduced to most major sesame planting areas in China over the last 10 years. In addition to the well-assembled genome sequences, a new high-density genetic map, 12 in-depth RNA-Seq data sets and resequencing data for 29 sesame accessions were generated for understanding and analysing genome structure, evolution and important nutritional characters, including lipid and sesamin synthesis, in the most comprehensive way. Together, these results will open a door for genetic studies in favour of a variety of purposes in, but not restricted to, sesame.

Results and discussion

De novo genome sequencing

Assembly and assessment

After reads filtering, 54.5 Gb of high-quality data from sesame cultivar “Zhongzhi No. 13” were obtained using the Illumina Hiseq2000 platform (Figure S1, Table S1 and S2 in Additional file 1; Data S1 in Additional file 2), representing approximately 152.7 fold coverage of the predicted sesame genome. SOAPdenovo [22] was used to assemble the genome, which resulted in a draft genome of 274 Mb with contig N50 and scaffold N50 of 52.2 kb and 2.1 Mb, which are ~2.7 and ~92.9-fold longer, respectively, than a previous survey of the sesame genome [21] (Table 1; Table S3 and S4 in Additional file 1). Using a newly constructed genetic map consisting of 406 markers (Data S2 in Additional file 2), we anchored 150 large scaffolds (117 oriented) into 16 pseudomolecules, which harboured 85.3% of the genome assembly and 91.7% of the predicted genes (numbered as LG1 - LG16; Table 1; Figure 1; Table S5 and Figure S3-S5 in Additional file 1). The estimate heterozygosity of the assembled sequenced line was 1.08×10^{-4}. This low heterozygosity was not unexpected because sesame is a self-pollinated plant [23], and we had performed successive selfing for five generations on the sample before sequencing to guarantee its homozygosity.

Table 1 Summary of sesame genome assembly and annotation

Assembly		**Number**	**N50 (size/number)**	**N90 (size/number)**	**Total length**
Contigs	All	26,239	52.17 kb/1,545	11.40 kb/5,534	270 Mb
Scaffolds	All	16,444	2.10 Mb/42	268.23 kb/169	274 Mb
	Anchored on chromosomes	150	−−	−−	234 Mb
	Anchored on chromosomes and oriented	117	−−	−−	207 Mb
Annotation		Number	Total length	Percentage of the assembly	
Protein coding genes	All	27,148	86.08 Mb	31.46	
Transposable elements	All	−−	78.86 Mb	28.46	
	LTR-Retroelements	−−	48.03 Mb	17.56	
	non-LTR Retrotransposons*	−−	11.70 Mb	4.28	
	DNA transposons	−−	10.88 Mb	3.98	
	Unknown	−−	14.64 Mb	5.35	
Noncoding RNAs	rRNA fragments	386	89.66 kb	< 0.04	
	tRNAs	870	65.31 kb	< 0.03	
	miRNAs	207	25.41 kb	< 0.01	
	snRNAs	268	33.93 kb	< 0.02	

*Long interspersed nuclear elements (LINEs) and short interspersed nuclear elements (SINEs).

Figure 1 Distribution of the basic genomic elements of sesame. (A) Pseudomolecules. **(B)** Gene density (mRNA); the frequency of sites within gene regions per 500 kb ranged from 0.04 to 0.61. **(C)** DNA TE (transposon element) density; the frequency of sites within DNA TE regions per 500 kb ranged from 0 to 0.22. **(D)** Retrotransposon element density; the frequency of sites per 500 kb within retrotransposon element regions ranged from 0 to 0.71. **(E)** GC content; the ratio of GC sites per 100 kb ranged from 0.32 to 0.40. Inner ribbons indicate self-collinearity of sesame, and the homologous regions of more than 1 Mb are highlighted. Circos (http://circos.ca) was used to construct the diagram.

The assembly covered 77.4 - 81.3% of the genome size according to the estimations derived from 17-mer depth distribution (357 Mb) (Figure S2 in Additional file 1) and flow cytometry (337 Mb) (Figure S3 in Additional file 1). The integrity of gene space in the genome assembly was demonstrated by the successful mapping of 99.3% of 3,328 expressed sequence tags (ESTs) [16] retrieved from GenBank, and 98.5% of 86,222 unigenes that were assembled *de novo* from previously reported RNA-Seq data [17] (Table S6 in Additional file 1). In addition, the large-scale assembly accuracy was assessed using five fosmid clones (33.5-38.6 kb) that were sequenced thoroughly using the Sanger sequencing technology (see Supplementary Note in Additional file 1), whereby 99.6% of the clone sequences in average were identical to the assembly (Table S7 and Figure S7 in Additional file 1). Hence, a high-quality assembly of sesame is provided here, rendering it a valuable source for studying genome structure and evolution.

Genome annotation

We predicted 27,148 protein-coding genes with an average transcript size of 3,171 bp by *ab initio* and homology-based analyses (Table S8-S10 in Additional file 1), together with RNA-Seq reads-assisted annotation. Of those, 23,635 (87.1%) were supported by unigenes or protein similarity, with only 12.9% derived solely from *ab initio* gene predictions (Table S9 in Additional file 1). With regard to non-coding genes, we identified 207 miRNAs, 870 tRNAs, 268 snRNAs, and 386 rRNA fragments from the assembly, respectively (Table S11 in Additional file 1).

Transposable elements play an important role in and are the major components of plants. A comprehensive annotation revealed that the repeat elements in sesame are lower (28.5% of assembly) than in grapevine (52.2%) [24], tomato (63.2%) [25] and potato (54.5%) [26] (Table 1; Table S12 and S13 in Additional file 1). As observed in other sequenced genomes, long terminal repeats (LTRs) in sesame occupy the majority (51.1%) of repeat sequences. When the number of full-length LTRs (*Copia* and *Gypsy*) was plotted against their insertion time inferred from the intra-sequence divergence of LTR regions (Figure S8 in Additional file 1), the resulting age distributions exhibited typical ‘L’ shapes with the accumulation of many recent LTRs and many fewer old LTRs. These age distributions reflected a steady-state stochastic birth/death model for the dynamics of LTR accumulation and activity [27].

For the two major members of LTRs, the proportion of *Copia* (7.3% of the genome) in sesame is comparable to that in grapevine, tomato and potato, but the percentage of *Gypsy* (6.6%) is extremely underrepresented (Table S13 in Additional file 1). Unlike that observed in tomato and potato, the distribution of the divergence rate of *Gypsy* is very smooth and low in sesame, suggesting that it had not experienced explosive accumulation or activity (Figure S9 in Additional file 1), and may be associated with the low proportion of *Gypsy*, which in turn relates to the lower repeat element ratio and smaller genome size of sesame.

Evolution in the sesame genome

Gamma and a recent whole genome duplication event in sesame

Sesame belongs to the asterids clade of eudicots [28]. Taxonomically, is mostly related to *Utricularia gibba* [20], tomato and potato, the other whole genome sequences available and published thus far in this clade [25,26]. Based on the shared single-copy genes from 11 sequenced species (Table S14 and Figure S10 in Additional file 1), sesame was estimated to have diverged from *U. gibba* ~98 million years ago (68.6 - 145.2 MYA), and from the tomato–potato lineage ~125 MYA (89.8 - 185.8 MYA) (Figure S11 in Additional file 1).

We identified 11,934 shared dicots-monocots, 14,158 shared asterids-rosids (two clades of dicots) and 20,563 shared asterids lineage (sesame, *U. gibba*, tomato and potato) gene clusters (Figure 2a), representing their ancestral gene families. Moreover, we identified 450 gene families containing 2,638 genes, plus 3,972 single-copy genes that were specific to sesame (Table S14 in Additional file 1; Figure S10 in Additional file 1; Data S3 and S4 in Additional file 2). Many of these genes were or encoded P450 genes, zinc finger proteins, transposases, transcription factors and disease resistance genes, suggesting their possible roles in species differentiation and adaptability in the sesame lineage or Lamiales.

Figure 2 Genome evolution of sesame. (a) Left: shared and unique gene families in each lineage; Right: distribution of the asterids-specific gene families in sesame, *Utricularia gibba*, tomato and potato. Gene numbers in related gene families are listed in brackets. **(b)** The relationships between grapevine and two subgenomes of sesame. These regions were listed in Table S15 in Additional file 1. For each segments in the grapevine chromosome, two obvious duplicated collinear segments from sesame were aligned. According to the retained gene ratio in duplicated segments, the high (red) and low-fractionated segments (blue) are represented. **(c)** Polyploidisation events in the asterids I lineage. Age was estimated according to *K*s distribution. γ: gamma triplication event in eudicot ancestor; T: triplication event in the tomato-potato lineage; D: recent duplication event in sesame; Dmr: the most recent duplication event in *U. gibba*; Sin: *Sesamum indicum*; Sly: *Solanum lycopersicum*; Stu:*Solanum tuberosum*; Vvi: *Vitis vinifera*; Ugi: *Utricularia gibba*. See Supplementary Note 4 in Additional file 1 for more details.

Based on synonymous substitution rates (K_s) and synteny relationships, we uncovered the dicot-specific gamma (γ) event and a novel independent whole genome duplication (WGD) event in the lineage leading to sesame (Figure S12 and S13 in Additional file 1). The recent WGD event in the sesame lineage was further corroborated in that the single grapevine region always aligned with two sesame segments (Figure 2b; Table S15 and Figure S14 in Additional file 1). We tentatively partitioned the recent sesame lineage-specific WGD genomic regions into two nonoverlapping “subgenomes”. The two subgenomes of the WGD corresponded to ~61 Mb (7,781 genes) and ~74 Mb (7,975 genes) regions, respectively, constituting ~50% of the current sesame genome assembly. Among all referred grape genomic loci, 79.1% were found to have undergone substantial gene loss, with a copy retained in only one of duplicated syntenic regions (subgenomes) of the sesame genome (Supplementary Note, Table S16 and S17 in Additional file 1; Data S5 and S6 in Additional file 2), following the WGD that occurred in the sesame-lineage. To estimate the age of recent WGD events in the sesame genome, we extracted 1,239 duplicated sesame genes (both retained in two subgenomes; Data S7 in Additional file 2) and calculated their K_s values. We observed nearly parallel peaks and identical ranges in the K_s distributions of these duplicated gene pairs and those from the triplication event in tomato (Figure S15 in Additional file 1), which have been dated to ~71 (±19) MYA [25]. Therefore, the recent WGD event of sesame should have occurred independently in the parallel period of the triplication event in the tomato-potato lineage, but older than the most recent duplication in *U. gibba* after their divergence (Figure 2c). Genes retained in duplicate are not evenly distributed among different functional categories [29], and an obvious bias was observed for the genes corresponding to transport, regulation, signal transduction and metabolism in sesame (Table S18 in Additional file 1). These over-retained genes may function necessarily in increasing the complexity of the regulatory networks accountable for the interaction between genotype and environment for the species after WGD.

Absence of N-terminal Toll/interleukin-1 receptor (TIR) nucleotide binding sites encoding resistance genes in sesame

Genes encoding nucleotide-binding sites (NBSs) are the largest class of plant disease resistance genes. Based on whether they contain a TIR domain [30], NBS resistance genes can be further categorised into two subclasses (TIR and non-TIR). We identified a total of 171 genes with an NBS domain in sesame (Table S19 in Additional file 1), and 65.2% were organised in tandem arrays (Figure S16 in Additional file 1). Intriguingly, all of the NBS-encoding resistance genes belonged to non-TIR type. The absence of TIR domain-containing resistance genes has been reported generally in monocots [31], but is rare in eudicots although it has been detected in sugar beet (*Beta vulgaris*) by polymerase chain reaction (PCR) [32]. Analysis of homologous genes from 11 species including 8 eudicots and 3 monocots has also shown that sesame and monocots are absent from the OrthoMCL clusters of genes with the TIR domain (Data S8 in Additional file 2; Figure S17 in Additional file 1). The absence of the NBS genes with the TIR domain in sesame was further exanimated by searching the gene-masked assembly and the unmapped reads; neither showed a TIR domain in sesame. Hence, the unambiguous absence of TIR domain-containing resistance genes at the whole genome scale in sesame, a species of eudicots, provides a new paradigm for the study of the evolution of resistance genes [33]. However, the mechanism that induced such loss, and whether it is common in the order Lamiales require further elucidation.

Quality characters of sesame

Molecular foundation for the high oil content of sesame

Elucidation of the sesame genome allowed the unprecedented opportunity to study oil biosynthesis for its high oil content. By searching the lipid-related gene database consisting of 222 functional families of *Arabidopsis thaliana*, we found that sesame has unexpectedly low gene copy numbers (708) compared to *A. thaliana* (736), soybean (1,298), grapevine (732), tomato (902) and rice (805) (Data S9 in Additional file 2). For the two edible oil crops, the conspicuous discrepancy between sesame and soybean in their oil content (~55% vs. ~20%, respectively, of dry seed) and their predicted lipid gene number (708 vs. 1,298, respectively) implies that different properties or mechanisms of oil biosynthesis exist in the two distantly related oil crops.

In contrast, soybean contains more copies than sesame in ~94.1% of the 222 lipid-related gene families. We found that the families encoding lipid transfer protein type 1 (LTP1), midchain alkane hydroxylase, FAD4-like desaturase (FAD4-like), and alcohol-forming fatty acyl-CoA reductase (AlcFAR) have been expanded by tandem duplication in sesame (Figure S18 - S20 in Additional file 1). Among these families, LTP1 is the largest containing 34 genes with 29 clustering into four tandem arrays (Figure 3a). The high sequence similarity among the genes in tandem 3 and 4 suggested that each might have experienced a recent expansion. Of these LTP1 genes, more than 90% were expressed in a set of 12 sesame seed transcriptomes (FPKM > 1), confirming their functional activities in lipid biosynthesis (Figure 3b). Expansion and retention of these genes may reflect the selection for genomic variation corresponding to the production of high oil content during domestication because the enhancement of the LTP1 family may benefit oil accumulation by strengthening the transport of fatty acids, acyl-CoAs, and other lipid molecules [34]. In addition, the two cytosolic lipoxygenase (LOX) and lipid acyl hydrolase-like (LAH) families related to the degradation of lipids [35,36] are both contracted in sesame (8 LOXs and 18 LAHs) when compared to soybean (45 LOXs and 42 LAHs) (Figure S21 and S22 in Additional file 1). Based on these data, we speculate that the expansion of some lipid gene families, especially the type 1 lipid transport genes, and the contraction of lipid degradation-related families may lead to higher oil content in sesame than soybean.

Figure 3 Diagrams represent gene expansion of LTP1 in sesame. (a) Maximum–likelihood tree showing expansions of LTP1 in sesame. Red dots, sesame; green dots, *Arabidopsis*; yellow dots, soybean; blue dots, grapevine. **(b)** Expression patterns and tandem arrays of LTP1 in sesame. Twelve transcriptomes corresponding to the seeds of 10, 20, 25 and 30 DPA of three accessions (from left to right): ZZM4728 (high oil content, 59% of seed), ZZM3495 (low oil content, 51% of seed) and ZZM2161 (low oil content, 48% of seed).

Differentially expressed lipid genes in seed development

To investigate the potential mechanism underlying the variation in oil content in sesame, we evaluated two accessions representing low oil content materials (48% and 51%) in germplasm, along with the high oil content cultivar "Zhongzhi No. 13" (59%), each in four developmental stages [10, 20, 25, and 30 days post anthesis, (DPA)] for RNA-Seq analysis (Table S1 in Additional file 1). Clustering of the expression profiles of the 416 sesame genes that were predicted to be orthologous to the lipid-related genes of *A. thaliana* (Supplementary Note in Additional file 1), clearly distinguished the 10 DPA from other stages and also the high and low oil sesame accessions (Figure 4a), suggesting that the determination of different oil content by lipid-related genes begins in the early stages of seed development.

Figure 4 Expression profiles of the lipid-related genes in sesame seed. Hierarchical clustering of the sesame seed samples based on the exp. **(a)** Hierarchical clustering of the sesame seed samples based on the expression levels (RPKM) of 416 predicted lipid-related genes. **(b)** Differently expressed genes (DEGs) of 10 DPA seeds in the downstream part of the TAG synthesis pathway. DEGs between the high and low lipid content accessions in this pathway are marked in red. Coloured squares indicate the expression level (RPKM) of the 10 DEGs encoding phosphatidylcholine: diacylglycerol cholinephosphotransferase (PDCT), phospholipid: diacylglycerol acyltransferase (PDAT), oil-body proteins (steroleosin, caleosin and oleosin) and transcription factors (ABI3, ABI4 and WRI1) in the three accessions (from left to right): ZZM4728 (high oil content, 59% of seed) and ZZM3495 (low oil content, 51% of seed), ZZM2161 (low oil content, 48% of seed). The correlation (r) of expression patterns between transcription factors and genes were calculated using Pearson's correlation coefficients (PCCs) based on all 12 transcriptomes. Dashed arrows indicate potential positive regulation.

After checking all oil-related pathways, we found that the genes expressed differentially at 10 DPA were enriched significantly in the downstream of triacylglycerol (TAG) biosynthesis pathway, including genes encoding phosphatidylcholine: diacylglycerol cholinephosphotransferase (PDCT) [37], phospholipid: diacylglycerol acyltransferase (PDAT) [38], oil-body proteins (steroleosin, caleosin and oleosin), and transcription factors (ABI3, ABI4 and WRI1) (Data S10). Until recently, the last step in TAG biosynthesis was assumed to be uniquely catalyzed by acylCoA: diacylglycerol acyltransferase (DGAT). However, some plants (e.g. sunflower, castor bean and *Crepis palaestina*) and yeast were found to have an acyl-CoA-independent mechanism for TAG synthesis, which uses phospholipids as acyl donors and diacylglycerol (DAG) as the acceptor [39,40]. In the present study, PDAT was expressed 2- to 3.5-fold higher in the one high-oil accession than the two low-oil accessions at 10 DPA (Figure 4b), whereas DGAT showed no significant differences in expression. These results are in accordance with the results in yeast, which revealed that overexpression of PDAT can increase TAG by 2-fold in the early logarithmic phase [40,41]. Collectively, these data strongly suggested that the expression of PDAT in collaboration with other genes plays a pivotal role in shaping the oil accumulation in the early stage of sesame seed development.

Population variation in sesame lipid-related genes

To screen the sequence variation in lipid-related genes, 29 sesame accessions from 12 countries with oil content variation ranging from 48.6% to 59.8% were selected for genome resequencing. More than 120 Gb clean data corresponding to 13-fold genomic coverage for each accession were generated, resulting in the identification of 2,348,008 single nucleotide polymorphisms (SNPs) (Data S11 in Additional file 2). From these SNPs, population diversity (π) and Watterson's estimator of segregating sites (θw) were estimated to be 0.0025 and 0.0032, respectively, in the population (Figure 5a; Figure S23 in Additional file 1). This genetic diversity is lower than that in rice [42], but higher than that in chickpea (*Cicer arietinum*) [43], watermelon (*Citrullus lanatus*) [44] and soybean [45] (Table S20 in Additional file 1). Lipid-related genes in sesame showed a wide variation with π values ranging from 0 to 0.0099, and were similar to the average of other genes (0.0021 vs. 0.0020). In addition, lipid-related genes with high π values (top 10%) were significantly ($P < 0.0001$) enriched in the two biological processes of lipid transport and lipid localisation. For example, most genes in the LTP1 family exhibited high diversity (Figure 5b). Specifically, the "youngest" tandem 4 nearly occupied all of the highest π values in lipid genes. Furthermore, using read depth of coverage, we found about a quarter of the 708 predicted lipid-related genes in sesame had copy number variations (CNVs) (Data S12 in Additional file 2) that were enriched significantly ($P < 0.01$) in the biological process of lipid transport and the molecular function of lipid binding. For the LTP1 genes, 3 of tandem 1, 3 of tandem 2, 4 of tandem 3 and 8 of tandem 4 were observed to have CNVs (2–4 copies in max) among the population (Data S13 in Additional file 2). The abundant variation in LTP1 might be associated with the intra-species differences in oil contents.

Figure 5 Schematics of the pairwise nucleotide diversity (π, red) and total polymorphism (θw, blue) of sesame. Schematics of the pairwise nucleotide diversity (π, red) and total polymorphism (θw, blue) of sesame. **(a)** Distributions of π (red) and θw (blue) of the sesame genome and positions of lipid-related genes (part, see Figure S23 in Additional file 1). The two lines of bars below the axis of π or θw indicate the positions of the lipid-related genes in sesame. Blue bars, lipid-related genes except for LTP1; red bars, LTP1 genes. **(b)** Scatterplots of π and θw values of LTP1 genes (red triangle), non-LTP1 lipid-related genes (blue triangle) and the other genes (green circles).

We also found that the oleate desaturase (FAD2, $\pi = 0.0016$) and linoleate desaturase (FAD8, $\pi = 0.0018$) genes usually keep low diversity, which may partially explain the low variation (40 ± 5% of oil) in oleic and linoleic acid content in sesame accessions [2]. For the oil-body protein families, higher variation was detected in the caleosin family (with an average π value of 0.0047) than in oleosin (0.0025) and steroleosin (0.0024). In contrast, transcription regulatory families containing WRI1, ABI3 and ABI4 exhibited relatively low π values (< 0.0008), indicating a conserved function for these transcription factors that participate in seed development and oil biosynthesis.

Genes for sesamin biosynthesis

Sesamin is an oil-soluble furofuran lignan typically present in sesame seed. It can increase oil stability and has been reported to be positively correlated with the oil content [46]. Sesamin biosynthesis involves two key genes encoding dirigent protein (DIR) and piperitol/sesamin synthase (PSS), respectively [47] (Figure S24a in Additional file 1). By conducting a BLAST search against the 11 species (Table S14 in Additional file 1), we found that the DIR homologues are present in sesame (SIN_1015471) and tomato, but the PSSs are only detected in sesame (SIN_1025734), indicating the genetic foundation for the sesame-specific product.

The gene expression level usually correlates with its product [48]. In the four stages of three sesame accessions with different sesamin content (Table S1 in Additional file 1), DIR (SIN_1015471) expression was generally highest in the three 10 DPA seed samples, which decreased in the following stages. In contrast, PSS showed different patterns among the three accessions (Figure S24b in Additional file 1), with the highest expression levels detected at 20 DPA in the low sesamin accessions as reported previously [49] but at higher levels than that detected at 10 DPA in the high sesamin accession. The higher PSS level in the early seed developmental stage in the high sesamin accession might be associated with its dual catalytic property, thereby producing more piperitol for subsequent sesamin biosynthesis. Although we expected to find PSS sequence variations between the high and low sesamin accessions, we found that it was especially conserved when we checked the 29 resequenced accessions (the three accessions used for RNA-Seq were also included). Thus, we speculated that other genes might regulate the different expressions of PSS; therefore, we selected a list of co-expressed genes of PSS for further study (Data S14 in Additional file 2). These genes are mainly involved in metabolic processes and catalytic activity (Figure S25 in Additional file 1).

Conclusions

The genome of sesame assembled *de novo* offers a new whole genome sequence in the order Lamiales that follows the typical minute genome *U. gibba*. This information provides an important resource for genetic and evolutionary studies. The evolutionary scenario outlined for sesame clearly revealed a more recent WGD event at ~71(±19) MYA, which occurred after the split from tomato and potato and presents a new resource for studying the intricate paleopolyploidisation processes in plants. The evolution of genes in Lamiales or asterids may be more complicated, considering the complete loss of the TIR-type NBS-encoding resistance genes in sesame, which undoubtedly presents a new paradigm in elucidating the fate of resistance genes along with their interactions with diseases. Moreover, determining whether similar mechanisms exist that induce or offset the loss of the TIR domain in both eudicots and monocots will be of great interest.

Although many studies have focused on the mechanisms of lipid biosynthesis and accumulation and the detection of lipid-related genes in different species, the genes involved in the dozens of complex oil biosynthesis pathways require further elucidation. For primary edible oil crops such as rapeseed, peanut and soybean, it is inextricable to encounter intertwined polyploidy or large genome sizes with many lipid-relates genes, which makes studies on oil biosynthesis much more daunting. In contrast, the higher oil content and fewer lipid-related genes in the small and diploid genome of sesame make it an invaluable potential model plant for studying oil biosynthesis. The *de novo* assembled genome, a set of 12 transcriptomes and 29 resequenced accessions provide a large resource for exploring the mechanisms underlying different oil contents between sesame and soybean, as well as among sesame accessions. The extensive expansion and high diversity of LTP1 genes, and key genes differentially expressed in the downstream of TAG biosynthesis pathway should aid future genetic studies in sesame. Undoubtedly, the advancement of study on sesame will improve the quantity and quality of the edible oil crops to fight food and nutrition crises.

Materials and methods

DNA and RNA Isolation

The sample used for whole genome *de novo* sequencing was “Zhongzhi No. 13”, an elite sesame cultivar that has been introduced to most of the major sesame planting areas over the last 10 years. Genomic DNA was extracted from leaves with a standard CTAB extraction method [50]. The materials used for RNA-Seq to analyse lipid and sesamin synthesis were three sesame accessions with different lipid and sesamin content (Table S1 in Additional file 1). The seeds of 10, 20, 25 and 30 DPA were sampled for RNA-Seq. The procedure described by Wei *et al*. [17] was used for RNA extraction and sequencing (or see Additional file 1).

Whole genome shotgun sequencing and assembly

We carried out whole genome shotgun sequencing with the Illumina Hiseq 2000 platform. Eight paired-end sequencing libraries with insert sizes of ~180 bp, 500 bp, 800 bp, 2 kb, 5 kb, 10 kb and 20 kb were constructed, which generated a total data amount of 99.54 Gb. To reduce the effect of sequencing error on assembly, we took a series of stringent filtering steps on reads generation (see Supplementary Note in Additional file 1). After the above quality-control and filtering steps, 54.46 Gb of clean data, ~150-fold coverage of the predicted genome size was remained (Table S2 in Additional file 1; Data S1 in Additional file 2). The quality and quantity of the filtered data were checked by the distributions of the clean reads from every library (Figure S1 in Additional file 1). For all of the 37.63 Gb of clean data from short-insert size libraries, a custom programme based on the *k*-mer frequency methodology was used to trim reads and correct bases [25]. Next, all of the remaining data were used for *de novo* genome assembly.

We carried out the whole genome assembly using SOAPdenovo [29,51].

Contig construction: First, we split the reads from the short-insert size libraries into *k*-mers (k = 71) and constructed a *de Bruijn* graph. We then simplified the graph referring to the parameters, and lastly connected the *k*-mer path to produce the contig file.

Scaffold construction: All usable reads were realigned onto the contig sequences; the amount of shared paired end relationships between each pair of contigs and the rate of consistent and conflicting paired-ends, were calculated to construct the scaffolds step by step, from short-insertion-size paired–ends to long-insertion-size paired–ends, and finally, we obtained scaffolds.

Gap filling: We used the tool GapCloser [52] to close the gaps inside the constructed scaffolds, which were mainly composed of repeats masked before scaffold construction. We used the paired-end information to retrieve the read pairs that had one end mapped to the unique contig and the other located in the gap region. Then, we preformed local assembly for these collected reads. Finally, about 274 Mb of sesame genome was assembled, 98.8% of which is non-gapped sequences (see Additional file 1).

Estimation of genome size by flow cytometry

Flow cytometry (FCM) was used to determine the DNA content of sesame [53]. Sesame samples and reference material were analysed on an EPICS Elite ESP cytometer (Beckman-Coulter, Hialeah, FLA) with an air-cooled argon laser (Uniphase) at 488 nm using 20 mW. Salmon erythrocytes (2.16 pg/1C) were used as internal biological reference materials. Nuclear DNA content (in pg) of sesame samples was estimated according to the following equation: 1C nuclear DNA content = (1C reference in pg × peak means of sesame)/(peak mean of reference). The number of base pairs per haploid genome was calculated based on the equivalent of 1 pg DNA = 978 Mb [54]. As a result, the C-value of sesame was estimated to be 0.34 pg/1C, and its genome size was estimated to be ~337 Mb (Figure S3 in Additional file 1).

Anchoring of genome assembly to sesame genetic map

We used a combination method of specific length amplified fragment (SLAF) sequencing and experiment marker analysis to construct a new genetic map using 107 F_2 lines derived from the Zhongzhi No.13/ZZM2289 population. In total, 2,719 SNPs, 97 insertions & deletions (InDel) and 2,282 SSR markers were developed and screened against the population. After filtering the markers with low PCR quality, those having no polymorphism and those showing significantly distorted segregation in the population, the retained 45 InDel, 259 SNP and 124 SSR markers were used to construct the genetic map using Joinmap3 software. Finally, we successfully constructed a genetic map that spans 1790.08 cM and has 406 markers including 39 InDel, 251 SNP and 116 SSR markers. Based on the genetic map, 150 large scaffolds were anchored onto 16 pseudomolecules (see details in the Supplementary Note 3 in Additional file 1).

Gene structure prediction and function annotation

To predict genes in the assembled genome, we used both homology-based and *de novo* methods. For the homology-based prediction, *A. thaliana*, grapevine, castor, and potato proteins were mapped onto the assembled genome using Genewise [55] to define gene models. For *de novo* prediction, Augustus [56] and GlimmerHMM were employed using appropriate parameters. Data from these complementary analyses were merged to produce a nonredundant reference gene set using GLEAN [57]. In addition, RNA-Seq data of multi − tissues (young roots, leaves, flowers, developing seeds, and shoot tips) from our previous study [17] were also incorporated to aid in gene annotation. RNA-Seq data were mapped to the assembled genome using Tophat [58], and transcriptome-based gene structures were obtained by cufflinks [59]. Then, we compared this gene set with the previous one to get the final nonredundant gene set of sesame (Table S8 – S10 in Additional file 1). The noncoding gene predications and gene function annotations were conducted as described in the Supplementary Note 3 and Table S11 in Additional file 1.

Repeat annotation

We identified repeat content in the sesame genome using a combination of *de novo* and homology-based approaches (see Supplementary Note, Table S12 and S13 in Additional file 1). Full-length LTR retrotransposons were identified by LTR_STRUC [60] and classified as *Gypsy*, *Copia* and other types of transposons using the programme RepeatClassifer implemented in the RepeatModeler package [61]. Then the insertion time of LTR retrotransposons was dated according to the method described by JessyLabbé [62] (see Supplementary Note and Figure S8 in Additional file 1).

Evolution analysis

Gene clustering was conducted with OrthoMCL [63] by setting the main inflation value 1.5 and other default parameters. PHYML [64] was selected to reconstruct the phylogenetic tree based on the HKY85 model [65]. The programme MCMCTree of the PAML package [66] was used to estimate species divergence time. Mcscan [67] was used to construct the chromosome collinearity. Detailed descriptions about the identification of recent WGD events and two subgenomes are provided in the Supplementary Note 4 in Additional file 1.

Analysis of resistance genes in sesame

HMMER V3.0 [68] was used to screen predicted sesame proteome against the raw hidden Markov model (HMM) corresponding to the Pfam NBS (NB-ARC), and further build sesame–specific NBS HMM for screening. The TIR and LRR domains were identified using local Pfam_Scan (−E 0.01 --domE 0.01). MARCOIL [69] with a threshold probability of 90 and the programme paircoil2 [70] with a P-score cutoff of 0.025 were used as the settings for the CC motif identification.

The absence of the NBS gene with a TIR domain in the sesame genome was further validated by checking the gene-masked assembly and the unassembled reads. For the masked assembly, we found 9 NB-ARC fragments (> 300 bp), but no TIR hit was obtained. Among all the unmapped reads, only 19 showed homology to TIR domain, but all the reads together covered less than half of the TIR region. Considering the above results, the NBS genes with a TIR domain were absent from sesame (for detailed methods see Supplementary Note 5 in Additional file 1).

Analysis of important characteristics in the genome

The homologous lipid gene in sesame and other crops were identified by blastp (1e-5, identity >30%) based on the database of acyl-Lipid metabolism in *A. thaliana* [71] (for detailed methods see the Supplementary Note 7 in Additional file 1).

Genome resequencing and SNP calling

For each accession, a paired-end sequencing library with insert sizes of 500 bp was constructed and then sequenced on the Hiseq 2000 platform. The raw reads were then subjected to a series of stringent filtering steps that had been used in *de novo* genome assembly (see Supplementary Note 1.3 in Additional file 1). Finally, we generated more than 120 Gb clean data totally with each sample at over 13-fold sequence depth (Data S11 in Additional file 2). The clean reads were mapped to the assembled sesame genome using BWA software [72]. After mapping, SNPs were identified with read mapping quality ≥20 on the basis of the mpileup files generated by SAMtools [73] (Data S7 in Additional file 2). The SNPs extracted by the above process were first filtered by the sequencing depth: ≥30 and ≤581 using the vcfutils programme in SAMtools; then the raw SNP sites were further filtered using the following criteria: copy number ≤2 and a minimum of 5 bp apart with the exception of minor allele frequencies (MAFs ≥ 0.05), whereby SNPs were retained when the distance between SNPs was less than 5 bp. Diversity parameters π and θ_w were measured using a window of 10 kb with a step of 1 kb [42,44].

Detection of CNVs was performed as described by Zhang *et al.* and Jiao *et al.* [74,75] (see Supplementary Note 8 in Additional file 1).

Data access

Genomic data generated by the whole project are available at NCBI the whole project under the accession number APMJ00000000 [76]. WGS raw reads are deposited under the SRA study: SRA122008 [77]. The raw RNA-Seq data are deposited under the SRA study: SRA122023 [78]. Genome assembly, annotation and RNA-Seq data, are also available at [79].

Abbreviations

AlcFAR, alcohol–forming fatty acyl-CoA reductase; CNV, copy number variation; DAG, diacylglycerol; DGAT, diacylglycerol acyltransferase; DIR, dirigent protein; DPA, days post anthesis; ESTs, expressed sequence tags; FAD4-like, FAD4-like desaturase; InDel, insertions & deletions; *Ks*, synonymous substitution rate; LAH, lipid acylhydrolase-like; LG, linkage group; LINEs, long interspersed nuclear elements; LOX, lipoxygenase; LTP, lipid transfer protein; LTRs, long terminal repeats; MAF, minor allele frequency; MYA, million years ago; NBS, nucleotide-binding site; PCC, pearson's correlation coefficients; PCR, polymerase chain reaction; PDAT, phospholipid: diacylglycerol acyltransferase; PDCT, phosphatidylcholine: diacylglycerol cholinephosphotransferase; PSS, piperitol/sesamin synthase; SLAF, specific-length amplified fragment; SNP, single nucleotide polymorphisms; SNPs, single nucleotide polymorphisms; SSR, simple sequence repeat; TAG, triacylglycerol; TF, transcription factors; TIR, Toll⁄interleukin-1 receptor; WGD, whole genome duplication.

Competing interests

The authors declare that they have no competing interests.

Authors' contributions

XRZ, JW and SYL contributed to the design of the research. LHW, SY, CS, RKV and CBT participated in the genome analysis and wrote the manuscript. YZZ, YL and WH participated in the study design and constructed the genetic map. LHW, SY, SYL and RKV participated in co-ordination and finalisation of the manuscript. CS, XDZ, YW, DL, FL and CYX participated in the genome and transcriptome analysis. XLH and SST participated in the resequencing and analysis. YXZ and DHL prepared materials and performed the experiments. JYY performed the database construction. JYW, XX, YRL and SMH participated in the statistical analysis. All authors read and approved the final manuscript.

Acknowledgements

The studies conducted by the Oil Crops Research Institute of the Chinese Academy of Agricultural Sciences, were supported by the National Basic Research Program of China (973 Program) (no. 2011CB109304 and no. 2011CB109305), the National Natural Science Foundation of China (no. 31271766, no. 31101182 and no. 31201242), Central Public–interest Scientific Institution Basal Research Fund, and China Agriculture Research System (no. CARS − 15).

References

1. Troncoso-Ponce MA, Kilaru A, Cao X, Durrett TP, Fan J, Jensen JK, Thrower NA, Pauly M, Wilkerson C, Ohlrogge JB: **Comparative deep transcriptional profiling of four developing oilseeds.** *Plant J* 2011, **68:**1014–1027.

2. Wei W, Zhang Y, Lv H, Li D, Wang L, Zhang X: **Association analysis for quality traits in a diverse panel of Chinese sesame (*Sesamum indicum* L.) germplasm.** *J Integr Plant Biol* 2013, **58:**745–758.

3. Ke T, Mao H, Hui FL, Dong CH, Chai GH, Liu SY: **Bioinformatics analysis and functional annotation of complete expressed sequence tag collection for oil crops.** *China J Bioinformatics* 2010, **8:**165–170.

4. Johnson LA, Suleiman TM, Lusas EW: **Sesame protein: a review and prospectus.** *J Am Oil Chem Soc* 1979, **56:**463–468.

5. Miyake Y, Fukumoto S, Okada M, Sakaida K, Nakamura Y, Osawa T: **Antioxidative catechol lignans converted from sesamin and sesaminol triglucoside by culturing with Aspergillus.** *J Agric Food Chem* 2005, **53:**22–27.

6. Hirata F, Fujita K, Ishikura Y, Hosoda K, Ishikawa T, Nakamura H: **Hypocholesterolemic effect of sesame lignan in humans.** *Atherosclerosis* 1996, **122:**135–136.

7. Tsai CM, Chen PR, Chien KL, Su TC, Chang CJ, Liu TL, Cheng HC: **Dietary sesame reduces serum cholesterol and enhances antioxidant capacity in hypercholesterolemia.** *Nutr Res* 2005, **25:**559–567.

8. Noguchi A, Fukui Y, Iuchi-Okada A, Kakutani S, Satake H, Iwashita T, Nakao M, Umezawa T, Ono E: **Sequential glucosylation of a furofuran lignan, (+)-sesaminol, by Sesamum indicum UGT71A9 and UGT94D1 glucosyltransferases.** *Plant J* 2008, **54:**415–427.

9. **Angiosperm Phylogeny Website.** [http://www.mobot.org/MOBOT/research/APweb/].

10. Wang L, Zhang Y, Qi X, Gao Y, Zhang X: **Development and characterization of 59 polymorphic cDNA-SSR markers for the edible oil crop *Sesamum indicum* (Pedaliaceae).** *Am J Bot* 2012, **99:**e394–e398.

11. Zhang H, Wei L, Miao H, Zhang T, Wang C: **Development and validation of genic-SSR markers in sesame by RNA-seq.** *BMC Genomics* 2012, **13:**316.

12. Spandana B, Reddy VP, Prasanna GJ, Anuradha G, Sivaramakrishnan S: **Development and characterization of microsatellite markers (SSR) in *Sesamum* (*Sesamum indicum* L.) species.** *Appl Biochem Biotechnol* 2012, **168:**1594–1607.

13. Uzun B, Lee D, Donini P, Cagirgan MI: **Identification of a molecular marker linked to the closed capsule mutant trait in sesame using AFLP.** *Plant Breed* 2003, **122:**95–97.

14. Uzun B, Cagirgan MI: **Identification of molecular markers linked to determinate growth habit in sesame.** *Euphytica* 2009, **166:**379–384.

15. Zhang H, Miao H, Wei L, Li C, Zhao R, Wang C: **Genetic analysis and QTL mapping of seed coat color in sesame (*Sesamum indicum* L.).** *PLoS One* 2013, **8:**e63898.

16. Suh MC, Kim MJ, Hur CG, Bae JM, Park YI, Chung CH, Kang CW, Ohlrogge JB: **Comparative analysis of expressed sequence tags from *Sesamum indicum* and *Arabidopsis thaliana* developing seeds.** *Plant Mol Biol* 2003, **52:**1107–1123.

17. Wei W, Qi X, Wang L, Zhang Y, Hua W, Li D, Lv H, Zhang X: **Characterization of the sesame (*Sesamum indicum* L.) global transcriptome using Illumina paired-end sequencing and development of EST-SSR markers.** *BMC Genomics* 2011, **12:**451.

18. Wang L, Zhang Y, Qi X, Li D, Wei W, Zhang X: **Global gene expression responses to waterlogging in roots of sesame (*Sesamum indicum* L.).** *Acta Physiol Plant* 2012, **34:**2241–2249.

19. Yi DK, Kim KJ: **Complete chloroplast genome sequences of important oilseed crop *Sesamum indicum* L.** *PLoS One* 2012, **7:**e35872.

20. Ibarra-Laclette E, Lyons E, Hernandez-Guzman G, Perez-Torres CA, Carretero-Paulet L, Chang TH, Lan T, Welch AJ, Juarez MJ, Simpson J, Fernández-Cortés A, Arteaga-Vázquez M, Góngora-Castillo E, Acevedo-Hernández G, Schuster SC, Himmelbauer H, Minoche AE, Xu S, Lynch M, Oropeza-Aburto A, Cervantes-Pérez SA, de Jesús Ortega-Estrada M, Cervantes-Luevano JI, Michael TP, Mockler T, Bryant D, Herrera-Estrella A, Albert VA, Herrera-Estrella L: **Architecture and evolution of a minute plant genome.** *Nature* 2013, **498:**94–98.

21. Zhang H, Miao H, Wang L, Qu L, Liu H, Wang Q, Yue M: **Genome sequencing of the important oilseed crop *Sesamum indicum* L.** *Genome Biol* 2013, **14:**401.

22. Li Y, Hu Y, Bolund L, Wang J: **State of the art de novo assembly of human genomes from massively parallel sequencing data.** *Hum Genomics* 2010, **4:**271–277.

23. Carlsson AS, Pham TD, Bui TM, Werlemark G, Bui TC, Merker A: **A study of genetic diversity of sesame (*Sesamum indicum* L.) in Vietnam and Cambodia estimated by RAPD markers.** *Genet Resour Crop Evol* 2009, **56:**679–690.

24. Jaillon O, Aury JM, Noel B, Policriti A, Clepet C, Casagrande A, Choisne N, Aubourg S, Vitulo N, Jubin C, Vezzi A, Legeai F, Hugueney P, Dasilva C, Horner D, Mica E, Jublot D, Poulain J, Bruyère C, Billault A, Segurens B, Gouyvenoux M, Ugarte E, Cattonaro F, Anthouard V, Vico V, Del Fabbro C, Alaux M, Di Gaspero G, Dumas V, *et al*: **The grapevine genome sequence suggests ancestral hexaploidization in major angiosperm phyla.** *Nature* 2007, **449:**463–467.

25. The Tomato Genome Consortium: **The tomato genome sequence provides insights into fleshy fruit evolution.** *Nature* 2012, **485:**635–641.

26. Xu X, Pan S, Cheng S, Zhang B, Mu D, Ni P, Zhang G, Yang S, Li R, Wang J, Orjeda G, Guzman F, Torres M, Lozano R, Ponce O, Martinez D, De la Cruz G, Chakrabarti SK, Patil VU, Skryabin KG, Kuznetsov BB, Ravin NV, Kolganova TV, Beletsky AV, Mardanov AV, Di Genova A, Bolser DM, Martin DM, Li G, Yang Y, *et al*: **Genome sequence and analysis of the tuber crop potato.** *Nature* 2011, **475:**189–195.

27. Du J, Tian Z, Hans CS, Laten HM, Cannon SB, Jackson SA, Shoemaker RC, Ma J: **Evolutionary conservation, diversity and specificity of LTR-retrotransposons in flowering plants: insights from genome-wide analysis and multi-specific comparison.** *The Plant Journal* 2010, **63:**584–598.

28. The Angiosperm Phylogeny G: **An update of the Angiosperm Phylogeny Group classification for the orders and families of flowering plants: APG III.** *Bot J Linn Soc* 2009, **161:**105–121.

29. Wang X, Wang H, Wang J, Sun R, Wu J, Liu S, Bai Y, Mun JH, Bancroft I, Cheng F, Huang S, Li X, Hua W, Wang J, Wang X, Freeling M, Pires JC, Paterson AH, Chalhoub B, Wang B, Hayward A, Sharpe AG, Park BS, Weisshaar B, Liu B, Li B, Liu B, Tong C, Song C, Duran C, *et al*: **The genome of the mesopolyploid crop species *Brassica rapa*.** *Nat Genet* 2011, **43:**1035–1039.

30. Meyers BC, Dickerman AW, Michelmore RW, Sivaramakrishnan S, Sobral BW, Young ND: **Plant disease resistance genes encode members of an ancient and diverse protein family within the nucleotide-binding superfamily.** *Plant J* 1999, **20:**317–332.

31. Tarr DE, Alexander HM: **TIR-NBS-LRR genes are rare in monocots: evidence from diverse monocot orders.** *BMC Res Notes* 2009, **2:**197.

32. Tian Y, Fan L, Thurau T, Jung C, Cai D: **The absence of TIR-type resistance gene analogues in the sugar beet (*Beta vulgaris* L.) genome.** *J Mol Evol* 2004, **58:**40–53.

33. Yue JX, Meyers BC, Chen JQ, Tian D, Yang S: **Tracing the origin and evolutionary history of plant nucleotide-binding site-leucine-rich repeat (NBS-LRR) genes.** *New Phytol* 2012, **193:**1049–1063.

34. Kader JC: **Lipid-transfer proteins in plants.** *Annu Rev Plant Physiol Plant Mol Biol* 1996, **47:**627–654.

35. Thompson JE, Froese CD, Madey E, Smith MD, Hong Y: **Lipid metabolism during plant senescence.** *Prog Lipid Res* 1998, **37:**119–141.

36. Feussner I, Kuhn H, Wasternack C: **Lipoxygenase-dependent degradation of storage lipids.** *Trends Plant Sci* 2001, **6:**268–273.

37. 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 U S A* 2009, **106:**18837–18842.

38. Mhaske V, Beldjilali K, Ohlrogge J, Pollard M: **Isolation and characterization of an *Arabidopsis thaliana* knockout line for phospholipid: diacylglycerol transacylase gene (At5g13640).** *Plant Physiol Biochem* 2005, **43:**413–417.

39. Cases S, Stone SJ, Zhou P, Yen E, Tow B, Lardizabal KD, Voelker T, Farese RV Jr: **Cloning of DGAT2, a second mammalian diacylglycerol acyltransferase, and related family members.** *J Biol Chem* 2001, **276:**38870–38876.

40. Dahlqvist A, Stahl U, Lenman M, Banas A, Lee M, Sandager L, Ronne H, Stymne S: **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.

41. Oelkers P, Cromley D, Padamsee M, Billheimer JT, Sturley SL: **The DGA1 gene determines a second triglyceride synthetic pathway in yeast.** *J Biol Chem* 2002, **277:**8877–8881.

42. Xu X, Liu X, Ge S, Jensen JD, Hu F, Li X, Dong Y, Gutenkunst RN, Fang L, Huang L, Li J, He W, Zhang G, Zheng X, Zhang F, Li Y, Yu C, Kristiansen K, Zhang X, Wang J, Wright M, McCouch S, Nielsen R, Wang J, Wang W: **Resequencing 50 accessions of cultivated and wild rice yields markers for identifying agronomically important genes.** *Nat Biotechnol* 2012, **30:**105–111.

43. Varshney RK, Song C, Saxena RK, Azam S, Yu S, Sharpe AG, Cannon S, Baek J, Rosen BD, Tar'an B, Millan T, Zhang X, Ramsay LD, Iwata A, Wang Y, Nelson W, Farmer AD, Gaur PM, Soderlund C, Penmetsa RV, Xu C, Bharti AK, He W, Winter P, Zhao S, Hane JK, Carrasquilla-Garcia N, Condie JA, Upadhyaya HD, Luo MC, *et al*: **Draft genome sequence of chickpea (*Cicer arietinum*) provides a resource for trait improvement.** *Nat Biotechnol* 2013, **31:**240–246.

44. Guo S, Zhang J, Sun H, Salse J, Lucas WJ, Zhang H, Zheng Y, Mao L, Ren Y, Wang Z, Min J, Guo X, Murat F, Ham BK, Zhang Z, Gao S, Huang M, Xu Y, Zhong S, Bombarely A, Mueller LA, Zhao H, He H, Zhang Y, Zhang Z, Huang S, Tan T, Pang E, Lin K, Hu Q, *et al*: **The draft genome of watermelon (*Citrullus lanatus*) and resequencing of 20 diverse accessions.** *Nat Genet* 2012, **45:**51–58.

45. Lam HM, Xu X, Liu X, Chen W, Yang G, Wong FL, Li MW, He W, Qin N, Wang B, Min J, Guo X, Murat F, Ham BK, Zhang Z, Gao S, Huang M, Xu Y, Zhong S, Bombarely A, Mueller LA, Zhao H, He H, Zhang Y, Zhang Z, Huang S, Tan T, Pang E, Lin K, Hu Q: **Resequencing of 31 wild and cultivated soybean genomes identifies patterns of genetic diversity and selection.** *Nat Genet* 2010, **42:**1053–1059.

46. Zhang X, Li P, Wang X, Wang X: **Studies on relationship among lignans, oil and protein content in sesame seed.** *Chin J Oil Crop Sci* 2005, **27:**88–90.

47. Kim HJ, Ono E, Morimoto K, Yamagaki T, Okazawa A, Kobayashi A, Satake H: **Metabolic engineering of lignan biosynthesis in Forsythia cell culture.** *Plant Cell Physiol* 2009, **50:**2200–2209.

48. Greenbaum D, Colangelo C, Williams K, Gerstein M: **Comparing protein abundance and mRNA expression levels on a genomic scale.** *Genome Biol* 2003, **4:**117.

49. Ono E, Nakai M, Fukui Y, Tomimori N, Fukuchi-Mizutani M, Saito M, Satake H, Tanaka T, Katsuta M, Umezawa T, Tanaka Y: **Formation of two methylenedioxy bridges by a *Sesamum* CYP81Q protein yielding a furofuran lignan, (+)-sesamin.** *Proc Natl Acad Sci U S A* 2006, **103:**10116–10121.

50. Doyle JJ, Doyle JL: **Isolation of plant DNA from fresh tissue.** *Focus* 1990, **12:**13–15.

51. Li R, Zhu H, Ruan J, Qian W, Fang X, Shi Z, Li Y, Li S, Shan G, Kristiansen K, Yang H, Wang J: **De novo assembly of human genomes with massively parallel short read sequencing.** *Genome Res* 2010, **20:**265–272.

52. **Large-size Genome De-novo Assembler Website.** [http://sourceforge.net/projects/soapdenovo2/files/GapCloser/].

53. Dolezel J, Greilhuber J, Suda J: **Estimation of nuclear DNA content in plants using flow cytometry.** *Nat Protocols* 2007, **2:**2233–2244.

54. Dolezel J, Bartos J, Voglmayr H, Greilhuber J: **Nuclear DNA content and genome size of trout and human.** *Cytometry A* 2003, **51:**127–128. author reply 129.

55. Birney E, Durbin R: **Using GeneWise in the Drosophila annotation experiment.** *Genome Res* 2000, **10:**547–548.

56. Stanke M, Keller O, Gunduz I, Hayes A, Waack S, Morgenstern B: **AUGUSTUS: ab initio prediction of alternative transcripts.** *Nucleic Acids Res* 2006, **34:**W435–W439.

57. **GLEAN Website.** [http://glean-gene.sourceforge.net/].

58. Trapnell C, Pachter L, Salzberg SL: **TopHat: discovering splice junctions with RNA-Seq.** *Bioinformatics* 2009, **25:**1105–1111.

59. **Cufflinks Website.** [http://cufflinks.cbcb.umd.edu/].

60. McCarthy EM, McDonald JF: **LTR_STRUC: a novel search and identification program for LTR retrotransposons.** *Bioinformatics* 2003, **19:**362–367.

61. **RepeatMasker Website.** [http://www.repeatmasker.org].

62. Labbe J, Murat C, Morin E, Tuskan GA, Le Tacon F, Martin F: **Characterization of transposable elements in the ectomycorrhizal fungus Laccaria bicolor.** *PLoS One* 2012, **7:**e40197.

63. Li L, Stoeckert CJ Jr, Roos DS: **OrthoMCL: identification of ortholog groups for eukaryotic genomes.** *Genome Res* 2003, **13:**2178–2189.

64. Guindon S, Dufayard JF, Lefort V, Anisimova M, Hordijk W, Gascuel O: **New algorithms and methods to estimate maximum-likelihood phylogenies: assessing the performance of PhyML 3.0.** *Syst Biol* 2010, **59:**307–321.

65. Hasegawa M, Kishino H, Yano T: **Dating of the human-ape splitting by a molecular clock of mitochondrial DNA.** *J Mol Evol* 1985, **22:**160–174.

66. Yang Z: **PAML 4: phylogenetic analysis by maximum likelihood.** *Mol Biol Evol* 2007, **24:**1586–1591.

67. **MCscan Website.** [http://chibba.agtec.uga.edu/duplication/mcscan].

68. Finn RD, Clements J, Eddy SR: **HMMER web server: interactive sequence similarity searching.** *Nucleic Acids Res* 2011, **Web Server Issue 39:**W29–W37 [http://hmmer.janelia.org/software].

69. Delorenzi M, Speed T: **An HMM model for coiled-coil domains and a comparison with PSSM-based predictions.** *Bioinformatics* 2002, **18:**617–625.

70. McDonnell AV, Jiang T, Keating AE, Berger B: **Paircoil2: improved prediction of coiled coils from sequence.** *Bioinformatics* 2006, **22:**356–358.

71. **Arabidopsis Acyl-lipid Metabolism Website.** [http://aralip.plantbiology.msu.edu].

72. Li H, Durbin R: **Fast and accurate short read alignment with Burrows-Wheeler transform.** *Bioinformatics* 2009, **25:**1754–1760.

73. Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, Marth G, Abecasis G, Durbin R: **The Sequence Alignment/Map format and SAMtools.** *Bioinformatics* 2009, **25:**2078–2079.

74. Zheng LY, Guo XS, He B, Sun LJ, Peng Y, Dong SS, Liu TF, Jiang S, Ramachandran S, Liu CM, Jing HC: **Genome-wide patterns of genetic variation in sweet and grain sorghum (*Sorghum bicolor*).** *Genome Biol* 2011, **12:**R114.

75. Jiao Y, Zhao H, Ren L, Song W, Zeng B, Guo J, Wang B, Liu Z, Chen J, Li W, Zhang M, Xie S, Lai J: **Genome-wide genetic changes during modern breeding of maize.** *Nat Genet* 2012, **44:**812–815.

76. **Whole Genome Shotgun Sequencing Project of *Sesamum indicum*.** [http://www.ncbi.nlm.nih.gov/nuccore/APMJ00000000].

77. **The Sequence Read Archive (SRA) of the Genome of *Sesamum indicum*.** [http://www.ncbi.nlm.nih.gov/sra/?term=SRA122008].

78. **The Sequence Read Archive (SRA) of the Transcriptomes of *Sesamum indicum*.** [http://www.ncbi.nlm.nih.gov/sra/?term=SRA122023].

79. **Sinbase: A Comprehensive *Sesamum indicum* Genomics Database.** [http://ocri-genomics.org/Sinbase].

Additional files

Additional_file_1 as DOCX

Additional file 1. Consists of the Supplementary Note, Table S1 to S20, and Figure S1 to S25.

Additional_file_2 as XLSX

Additional file 2. Consists of the supplementary Data S1 to S14.

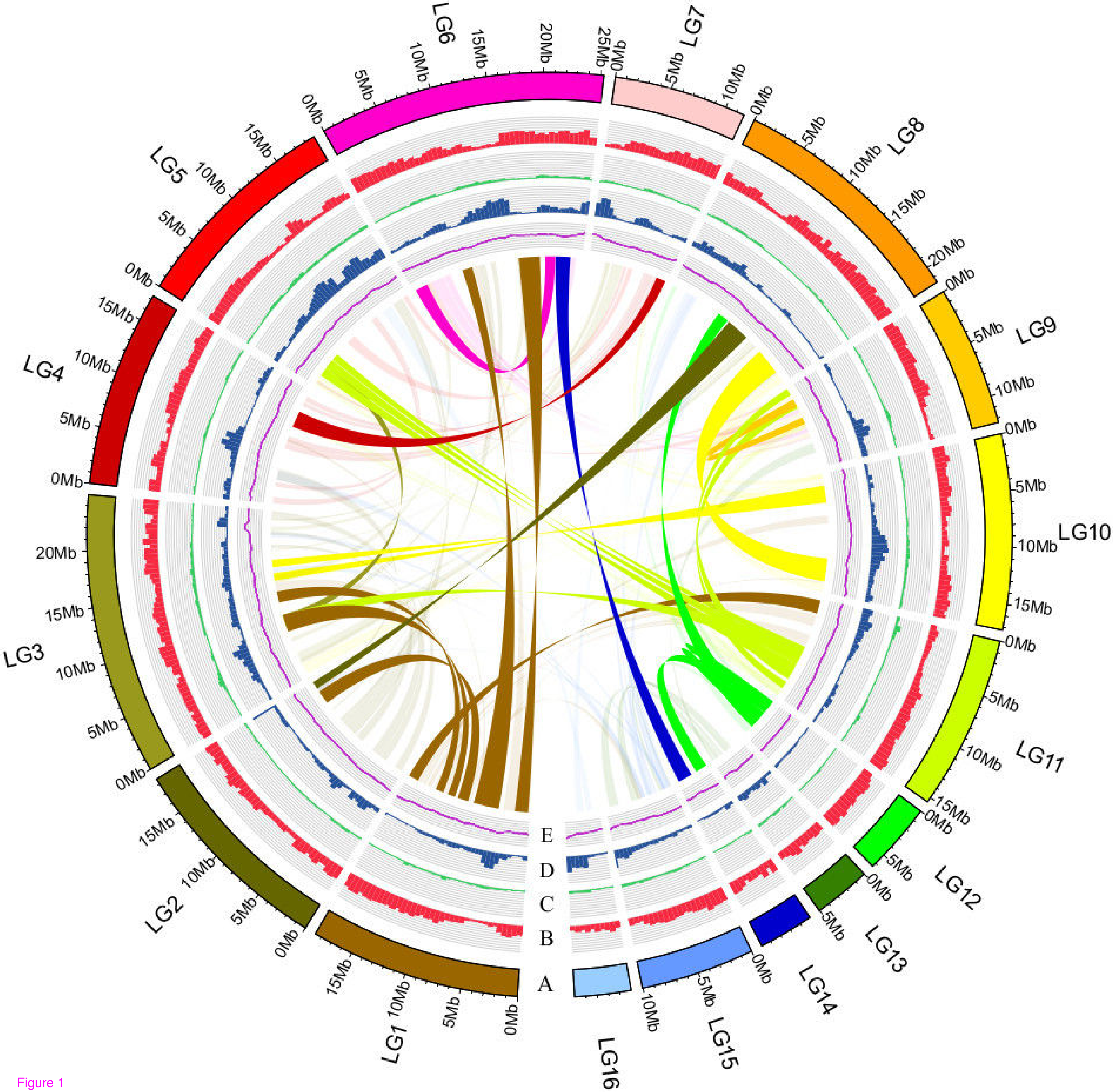
LG1
0Mb
5Mb
10Mb
15Mb
LG2
0Mb
5Mb
10Mb
15Mb
LG3
0Mb
5Mb
10Mb
15Mb
20Mb
LG4
0Mb
5Mb
10Mb
15Mb
LG5
0Mb
5Mb
10Mb
15Mb
LG6
0Mb
5Mb
10Mb
15Mb
20Mb
25Mb
LG7
0Mb
5Mb
10Mb
LG8
0Mb
5Mb
10Mb
15Mb
20Mb
LG9
0Mb
5Mb
10Mb
LG10
0Mb
5Mb
10Mb
15Mb
LG11
0Mb
5Mb
10Mb
15Mb
LG12
0Mb
5Mb
LG13
0Mb
5Mb
LG14
0Mb
LG15
5Mb
10Mb
LG16
A
B
C
D
E

Figure 1

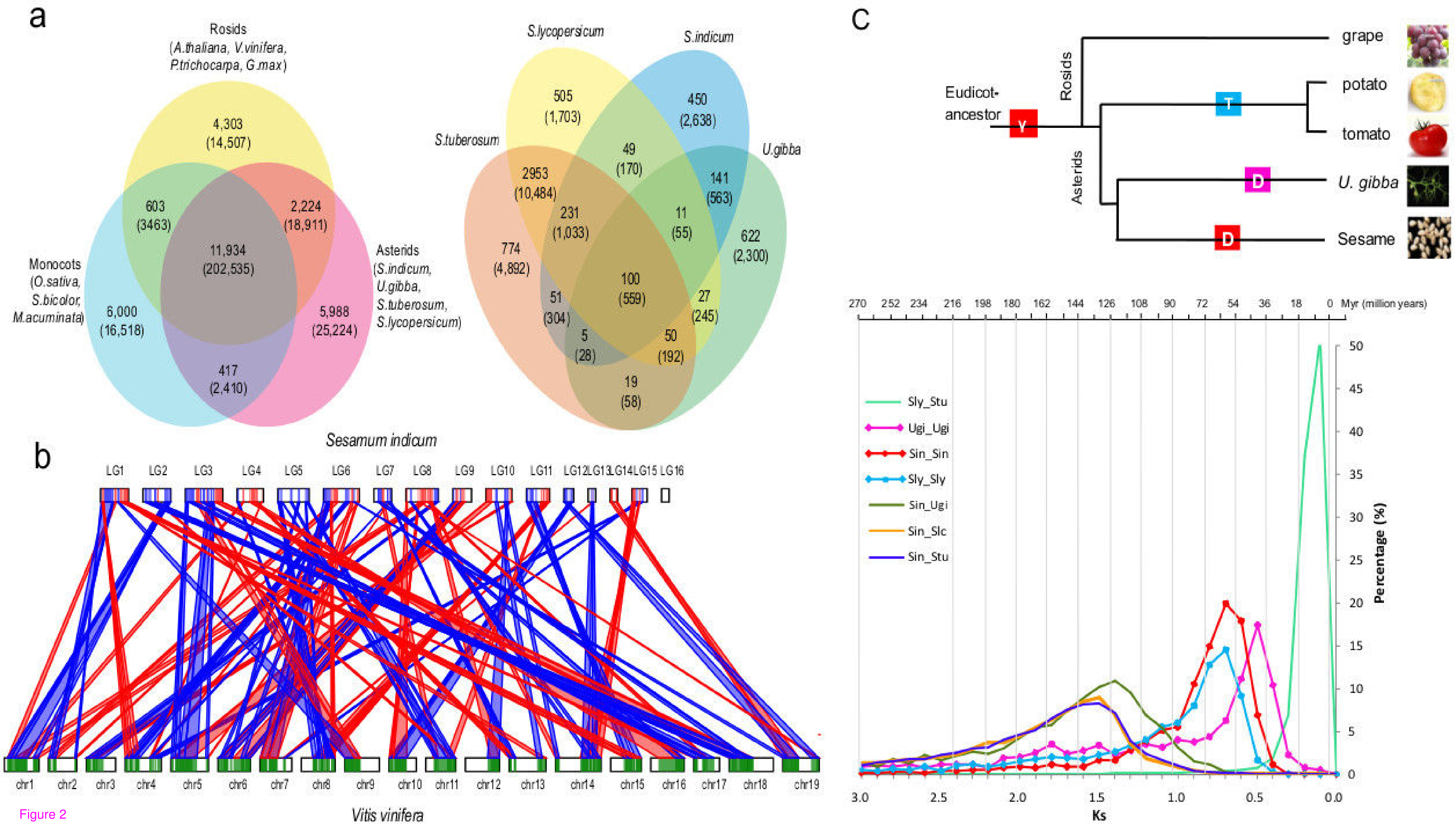
a
Rosids
(A.thaliana, V.vinifera, P.trichocarpa, G.max)
Monocots
(O.sativa, S.bicolor, M.acuminata)
Asterids
(S.indicum, U.gibba, S.tuberosum, S.lycopersicum)
4,303
(14,507)
603
(3463)
2,224
(18,911)
11,934
(202,535)
6,000
(16,518)
5,988
(25,224)
417
(2,410)
S.lycopersicum
S.indicum
S.tuberosum
U.gibba
505
(1,703)
450
(2,638)
49
(170)
2953
(10,484)
141
(563)
231
(1,033)
11
(55)
774
(4,892)
622
(2,300)
100
(559)
51
(304)
27
(245)
5
(28)
50
(192)
19
(58)
b
Sesamum indicum
LG1 LG2 LG3 LG4 LG5 LG6 LG7 LG8 LG9 LG10 LG11 LG12 LG13 LG14 LG15 LG16
chr1 chr2 chr3 chr4 chr5 chr6 chr7 chr8 chr9 chr10 chr11 chr12 chr13 chr14 chr15 chr16 chr17 chr18 chr19
Vitis vinifera
c
Eudicot-ancestor
γ
Rosids
Asterids
T
D
D
grape
potato
tomato
U. gibba
Sesame
270 252 234 216 198 180 162 144 126 108 90 72 54 36 18 0 Myr (million years)
Sly_Stu
Ugi_Ugi
Sin_Sin
Sly_Sly
Sin_Ugi
Sin_Slc
Sin_Stu
Percentage (%)
0 5 10 15 20 25 30 35 40 45 50
3.0 2.5 2.0 1.5 1.0 0.5 0.0
Ks

Figure 2

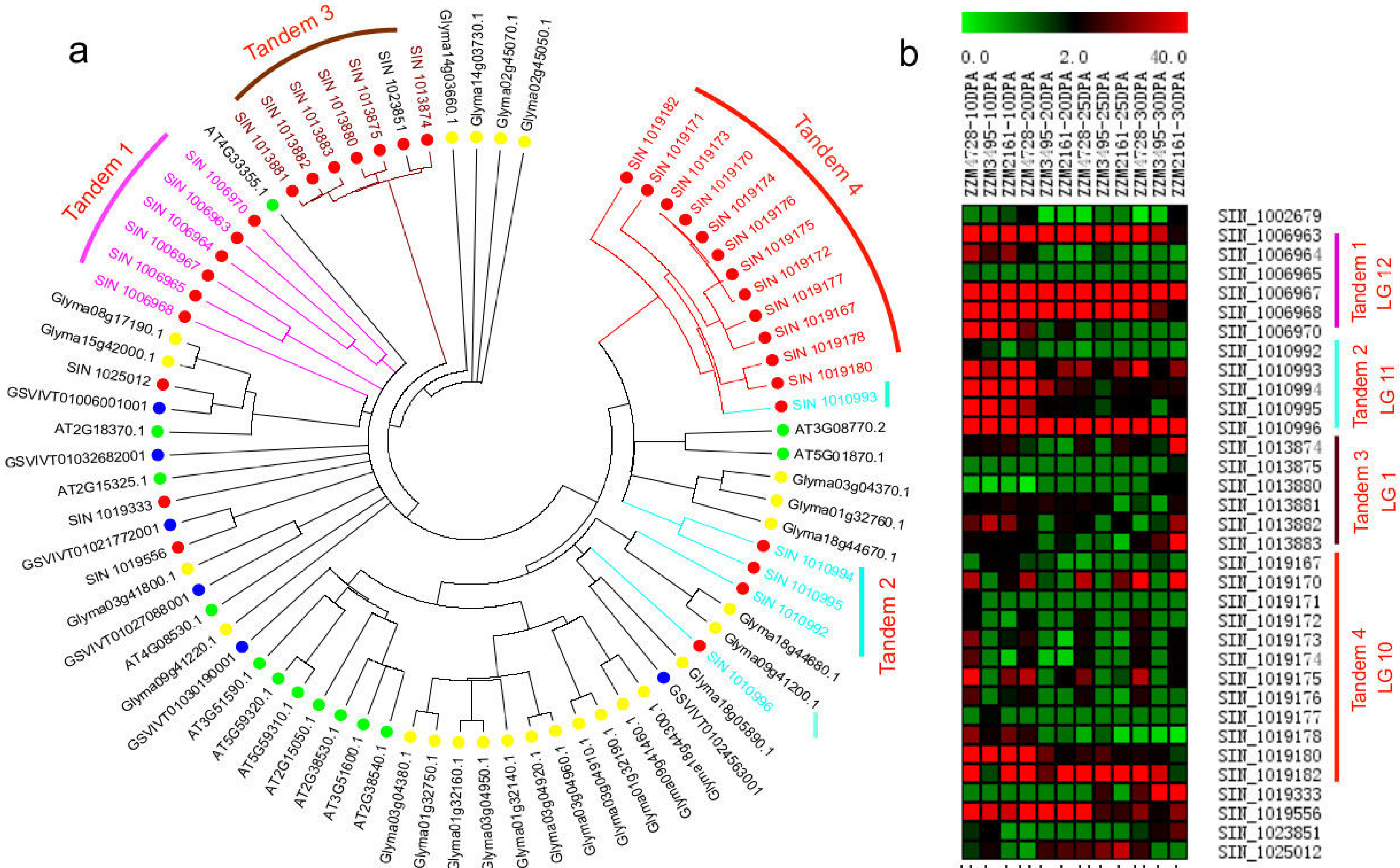
a
b
Tandem 1
Tandem 2
Tandem 3
Tandem 4
SIN 1006970
SIN 1006963
SIN 1006964
SIN 1006967
SIN 1006965
SIN 1006968
AT4G33355.1
SIN 1013881
SIN 1013882
SIN 1013883
SIN 1013880
SIN 1013875
SIN 1023851
SIN 1013874
Glyma14g03660.1
Glyma14g03730.1
Glyma02g45070.1
Glyma02g45050.1
SIN 1019182
SIN 1019171
SIN 1019173
SIN 1019170
SIN 1019174
SIN 1019176
SIN 1019175
SIN 1019172
SIN 1019177
SIN 1019167
SIN 1019178
SIN 1019180
SIN 1010993
AT3G08770.2
AT5G01870.1
Glyma03g04370.1
Glyma01g32760.1
Glyma18g44670.1
SIN 1010994
SIN 1010995
SIN 1010992
Glyma18g44680.1
Glyma09g41200.1
SIN 1010996
Glyma18g05890.1
GSVIVT01024563001
Glyma18g44300.1
Glyma09g41460.1
Glyma01g32190.1
Glyma03g04910.1
Glyma03g04960.1
Glyma03g04920.1
Glyma01g32140.1
Glyma03g04950.1
Glyma01g32160.1
Glyma01g32750.1
Glyma03g04380.1
AT2G38540.1
AT3G51600.1
AT2G38530.1
AT2G15050.1
AT5G59310.1
AT5G59320.1
AT3G51590.1
GSVIVT01030190001
Glyma09g41220.1
AT4G08530.1
GSVIVT01027088001
Glyma03g41800.1
SIN 1019556
GSVIVT01021772001
SIN 1019333
AT2G15325.1
GSVIVT01032682001
AT2G18370.1
GSVIVT01006001001
SIN 1025012
Glyma15g42000.1
Glyma08g17190.1
0.0
2.0
40.0
ZZM4728-10DPA
ZZM3495-10DPA
ZZM2161-10DPA
ZZM4728-20DPA
ZZM3495-20DPA
ZZM2161-20DPA
ZZM4728-25DPA
ZZM3495-25DPA
ZZM2161-25DPA
ZZM4728-30DPA
ZZM3495-30DPA
ZZM2161-30DPA
SIN_1002679
SIN_1006963
SIN_1006964
SIN_1006965
SIN_1006967
SIN_1006968
SIN_1006970
SIN_1010992
SIN_1010993
SIN_1010994
SIN_1010995
SIN_1010996
SIN_1013874
SIN_1013875
SIN_1013880
SIN_1013881
SIN_1013882
SIN_1013883
SIN_1019167
SIN_1019170
SIN_1019171
SIN_1019172
SIN_1019173
SIN_1019174
SIN_1019175
SIN_1019176
SIN_1019177
SIN_1019178
SIN_1019180
SIN_1019182
SIN_1019333
SIN_1019556
SIN_1023851
SIN_1025012
Tandem 1 LG 12
Tandem 2 LG 11
Tandem 3 LG 1
Tandem 4 LG 10
H L L H L L H L L H L L

Figure 3

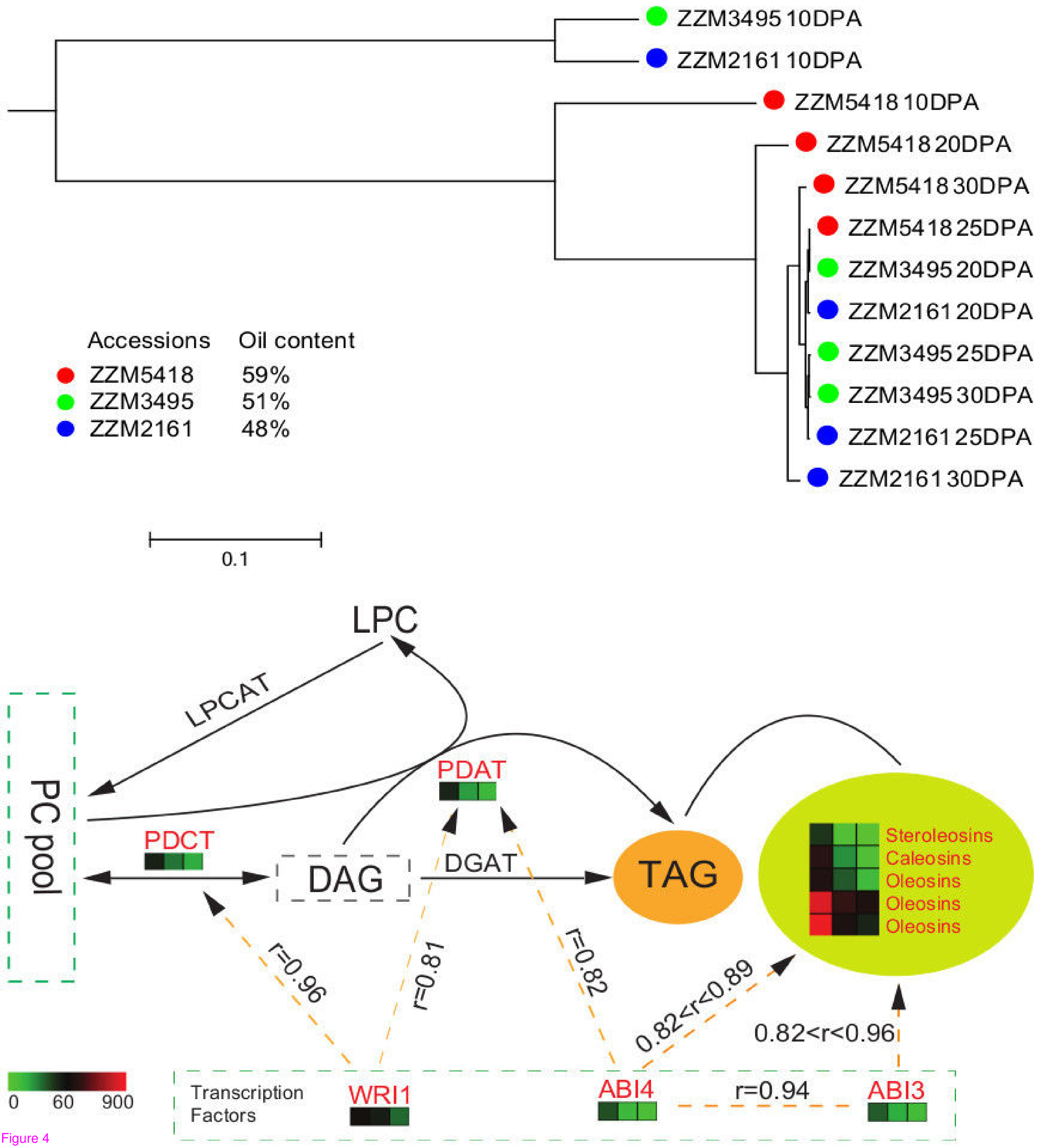
ZZM3495 10DPA
ZZM2161 10DPA
ZZM5418 10DPA
ZZM5418 20DPA
ZZM5418 30DPA
ZZM5418 25DPA
ZZM3495 20DPA
ZZM2161 20DPA
ZZM3495 25DPA
ZZM3495 30DPA
ZZM2161 25DPA
ZZM2161 30DPA
Accessions
Oil content
ZZM5418 59%
ZZM3495 51%
ZZM2161 48%
0.1
LPC
LPCAT
PC pool
PDAT
PDCT
DAG
DGAT
TAG
Steroleosins
Caleosins
Oleosins
Oleosins
Oleosins
r=0.96
r=0.81
r=0.82
0.82<r<0.89
0.82<r<0.96
r=0.94
0
60
900
Transcription
Factors
WRI1
ABI4
ABI3

Figure 4

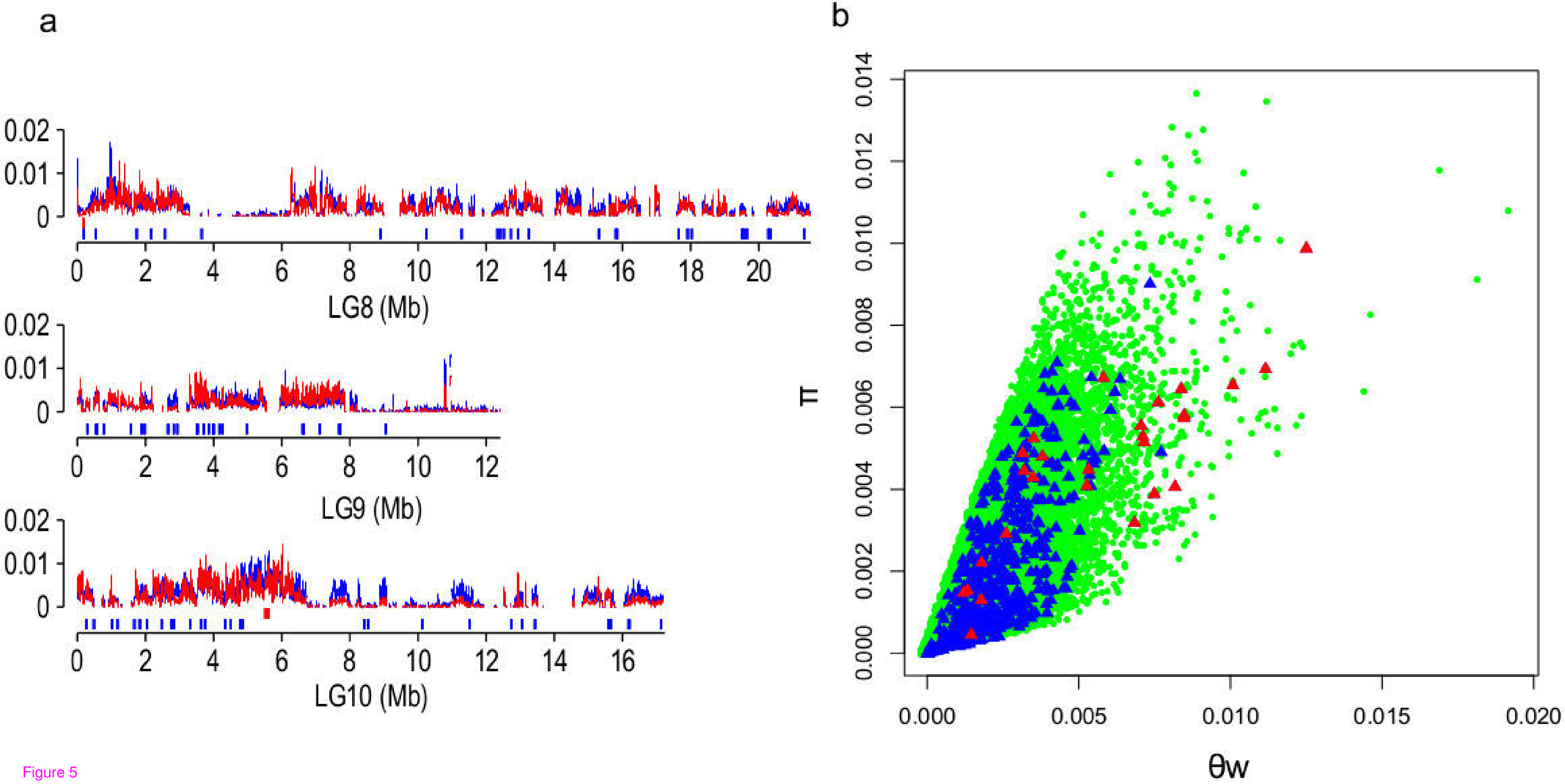
a
0.02
0.01
0
0 2 4 6 8 10 12 14 16 18 20
LG8 (Mb)
0.02
0.01
0
0 2 4 6 8 10 12
LG9 (Mb)
0.02
0.01
0
0 2 4 6 8 10 12 14 16
LG10 (Mb)
b
π
0.000 0.002 0.004 0.006 0.008 0.010 0.012 0.014
θw
0.000 0.005 0.010 0.015 0.020

Figure 5

Additional files provided with this submission:

Additional file 1: 3079078931033946_add1.docx, 4366K
http://genomebiology.com/imedia/5341480571204318/supp1.docx
Additional file 2: 3079078931033946_add2.xlsx, 1408K
http://genomebiology.com/imedia/2060045172120431/supp2.xlsx