OXFORD

Resource Article: Genomes Explored

The genome sequence of tetraploid sweet basil, *Ocimum basilicum* L., provides tools for advanced genome editing and molecular breeding

Itay Gonda [1], Adi Faigenboim[1], Chen Adler[1], Renana Milavski[1], Merrie-Jean Karp[1], Alona Shachter[1], Gil Ronen[3], Kobi Baruch[3], David Chaimovitsh[1], and Nativ Dudai[1]*

[1]Unit of Aromatic and Medicinal Plants, Newe Ya'ar Research Center, Agricultural Research Organization, Ramat Yishay, Israel, [2]Institute of Plant Sciences and Genetics in Agriculture, The Robert H. Smith Faculty of Agriculture, The Hebrew, University of Jerusalem, Rehovot, Israel, and [3]NRGene Ltd, Park HaMada, Ness Ziona, Israel

*To whom correspondence should be addressed. Tel. +972-506220010. Fax. +972-4-983-6936. Email: nativdud@agri.gov.il

Abstract

Sweet basil, *Ocimum basilicum* L., is a well-known culinary herb grown worldwide, but its uses go beyond the kitchen to traditional medicine, cosmetics and gardening. To date, the lack of an available reference genome has limited the utilization of advanced molecular breeding methods. We present a draft version of the sweet basil genome of the cultivar 'Perrie', a fresh-cut Genovese-type basil. Genome sequencing showed basil to be a tetraploid organism with a genome size of 2.13 Gbp, assembled in 12,212 scaffolds, with > 90% of the assembly being composed of 107 scaffolds. About 76% of the genome is composed of repetitive elements, with the majority being long-terminal repeats. We constructed and annotated 62,067 protein-coding genes and determined their expression in different plant tissues. We analysed the currently known phenylpropanoid volatiles biosynthesis genes. We demonstrated the necessity of the reference genome for a comprehensive understanding of this important pathway in the context of tetraploidy and gene redundancy. A complete reference genome is essential to overcome this redundancy and to avoid off-targeting when designing a CRISPR: Cas9-based genome editing research. This work bears promise for developing fast and accurate breeding tools to provide better cultivars for farmers and improved products for consumers.

Key words: sweet basil, tetraploidy, genes redundancy, *Ocimum basilicum*, phenylpropanoids

1. Introduction

The herb *Ocimum basilicum* L. (Lamiaceae) is one of ~160 different species comprising the genus *Ocimum*.[1] Known as basil, this herb is widely cultivated as an aromatic crop, but it is also grown in home gardens the world over. The diversity of *O. basilicum* genotypes—as manifested by a variety of unique aromas, leaf sizes and shapes, leaf and stem colors, inflorescence colors and structures, growth habits and seed morphologies[2]—reflects the diverse uses of this herb. It is used in traditional medicine, as an ornamental plant, and as a source of essential oils,[3] but primarily as a culinary herb, by virtue of the

distinctive aroma compounds that accumulate in specialized glandular trichomes.[4] The most widely grown of the genotypes is sweet basil, the type of basil used in the famous Italian pesto sauce. Still, despite the importance of sweet basil as an agricultural crop, genomic studies on this species are scarce.

To date, studies aimed to estimate the genome size of various genotypes of different *Ocimum* species on the basis of DNA content have used flow cytometry methods. Yet, these works have not produced consistent results. With the rough rule of thumb that 1 pg DNA = 0.978 Mbp,[5] the calculated haplotypic (1 C) genome size of 20 *O. basilicum* genotypes varies from 2.04 Gbp to 2.32 Gbp.[6] For the cultivar 'Perrie', a fresh-cut Genovese-type basil, the genome size was estimated to be 1.56 Gbp.[7] For holy basil, *O. sanctum sanctum* L. (synonym *O. tenuiflorum* L.), the genome size was estimated to be 1.39 Gbp in one study[7] and 350 Mbp in another.[6] A start was made on addressing these differences in two studies that sequenced the genome of holy basil; in these studies, Rastogi et al.[8] assembled 386 Mbp, and Upadhyay et al.[9] assembled 374 Mbp. The scaffold N50 values in these two studies were 303 kbp and 27 kbp, with a total of 9,059 and 78,224 scaffolds, respectively. Similarly, there are discrepancies in the chromosome number (2*n* value) for basil, which varies markedly among different studies and different genotypes. For *O. basilicum*, for example, the chromosome numbers were found to be 48, 52, 50, 52, 56, 72 or 74.[1,10–12] More recently, Carović-Stanko et al.[6] reported 2*n* = 48 for 20 *O. basilicum* genotypes and 2*n* = 72 for 2 *O. basilicum* var. *purpurascens* Benth genotypes. For *O. sanctum*, 2*n* has been reported to be 16, 32, 36 and 72.[11,13,14] A major advance was recently achieved with the publication of a draft genome for the home garden basil cultivar 'Genovese' by Bornowski et al.[15] They assembled 2.07 Gbp in > 17 thousand scaffolds (greater than 10 Kbp) with an N50 of ∼ 1.6 Mbp.

A shortage in advanced genetic material, together with the lack of a well-established and contiguous reference genome, may explain why basil breeding programs lag behind those of other crops, such as wheat, rice or tomato. This lag also hinders the development of advanced genome editing tools for basil, since it is challenging to overcome off-targeting and redundancy without a reference genome. In the project reported here, we followed-up on our previous effort[16] to close the-above described research and breeding gaps. To this end, the genome of the cultivar 'Perrie' was sequenced on Illumina platforms and assembled using NRGENE's DeNovoMagic™ assembler. The resulting haploid genome was found to be 2.13 Gbp in size and is currently composed of 12,212 scaffolds. The completeness of the assembly was validated using the Benchmarking Universal Single-Copy Orthologs (BUSCO) pipeline, with less than 2% of the gene-set being fragmented. To add another layer of usable information, we built—assisted by RNA-seq data—gene models for 62,067 protein-coding genes. In parallel, we conducted a homeologous-specific analysis of the biosynthesis pathway of basil aroma, which revealed a complex genomic path of both redundancy and speciation. We have thus established, and report here, a contiguous draft reference genome and annotated genes that will serve the research and agricultural communities for many years ahead.

2. Materials and methods

2.1. Plant material and high molecular weight genomic DNA extraction

For genome sequencing and construction, multiple clones of a single sweet basil plant of the cultivar 'Perrie' were grown in the dark to generate etiolated leaf tissue. For high molecular weight DNA extraction, 150 mg of the etiolated leaves were harvested, frozen in liquid nitrogen, and ground with mortar and pestle, according to Healey et al.[17] with some modifications. The fine powder was mixed with 1.5 ml of extraction buffer (100 mM Tris pH 7.5, 1.5 M NaCl, 20 mM EDTA, 2% w/v CTAB, 2% w/v PVP-40, and 0.3% v/v *β*-mercaptoethanol) and incubated at 60°C for 1 h with occasional shaking. The sample was centrifuged for 5 min at 5000 *g* at 4°C, and the pellet was discarded. To the supernatant was added an equal volume of isoamyl alcohol: chloroform (1:24 v/v), followed by gentle mixing for 5 min and centrifugation for 10 min at 5000 *g* at 4°C. The pellet was discarded, RNAse was added to the supernatant at a final concentration of 20 μg × ml⁻¹, and the sample was incubated for 15 min at 37°C with occasional shaking. Then, an equal volume of isoamyl alcohol: chloroform (1:24 v/v) was added, and the sample was gently mixed for 5 min, followed by 10 min of centrifugation at 5000 *g* at 4°C. The supernatant was transferred to a clean tube, a half volume of 5 M NaCl was added, and the sample was mixed gently. Then, 3 volumes of ice-cold 95% (v/v) ethanol were added, and the sample was gently mixed for 2 min, followed by incubation at −20°C for 1 h. The sample was centrifuged for 10 min at 5000 *g* at 4°C, the supernatant was discarded, and the pellet was washed with 3 ml of ice-cold 70% (v/v) ethanol. Finally, the sample was centrifuged for 10 min at 5000 *g* at 4°C, the ethanol was discarded, and the tube was left open for the ethanol residues to evaporate. The precipitated DNA was resuspended in 75 μl of TE buffer at 37°C. DNA quality and quantity were evaluated with a NanoDrop™ spectrophotometer and a Qubit fluorometer. The size of the extracted DNA was estimated by overnight running on a 0.5% agarose gel.

2.2. Preparation and sequencing of genomic libraries

Five size-selected genomic DNA libraries ranging from 470 bp to 10 Kb were constructed from the extracted basil gDNA. Two shotgun paired-end (PE) libraries were constructed using DNA template fragments of a selected size of ∼470 bp, without PCR amplification (PCR-free). This size was chosen to yield overlapping sequences of the paired reads from the Hiseq2500 v2 Rapid run mode as 2 × 265 bp, thereby allowing 'stitching' of the read-pairs up to 520 bp long reads. A PCR-free genomic library of 700-bp DNA fragments was prepared using the TruSeq DNA Sample Preparation Kit version 2 according to the manufacturer's instructions (Illumina, San Diego, CA, https://www.illumina.com/). Three independent MP libraries with insert sizes of 2–5 Kbp, 5–7 Kbp and 8–10 Kbp were constructed with the Illumina Nextera MP Sample Preparation Kit (Illumina). The 700-bp shotgun libraries and all three MP libraries were sequenced on an Illumina NovaSeq6000 instrument, using an S4 flow cell, as 2 × 150 bp reads. DNA fragments longer than 50 Kbp were used to construct a GemCode library using a Chromium instrument (10× Genomics, Pleasanton, CA https://www.10xgenomics.com/). This library was sequenced on a NovaSeq6000 instrument, using an S4 flow cell, as 2 × 150 bp reads. Construction and sequencing of PE, mate-pair (MP) and Chromium libraries were conducted at the Roy J. Carver Biotechnology Center, University of Illinois at Urbana-Champaign.

2.3. Genome assembly

The sweet basil genome was assembled with DeNovoMAGIC software (NRGene Ltd., Ness Ziona, Israel). This assembler is based on a De Bruijn-graph, designed to optimize raw data usage to resolve the complexity arising from genome polyploidy, heterozygosity and

repetitiveness. This was done by using accurate-read-based traveling in the graph, which iteratively connects consecutive phased sequences (contigs) over local repeats to generate long-phased scaffolds[18–22].

2.4. Genome annotations and construction of gene models

To assist gene annotations, we conducted TruSeq RNA-sequencing (RNA-seq) of leaves (seventh and eighth leaf pairs, three biological samples), flowers (three biological samples), stems (one sample) and roots (one sample). Total RNA was extracted according to Gonda et al.[23] PE RNA libraries were constructed and sequenced by Marcrogen (Seoul, South Korea, https://www.macrogen.com/en/main/index.php) using the TruSeq Stranded mRNA LT Sample Prep Kit (Illumina). The libraries were sequenced on a NovaSeq6000 S4 flowcell as 2×100 bp reads. To establish basil gene models, we combined data from the following platforms: Trinity,[24] SNAP,[25] AUGUSTUS[26] and MAKER.[27] The genome annotation process comprised the following five steps: (i) RepeatModeler[28] was used to build a repeat database. (ii) RNA reads were *de novo* assembled using Trinity software to generate genome-independent gene predictions. (iii) MAKER was run using the *de novo* Trinity transcriptome, the protein sequences database of *Salvia splendens* (extracted from the NCBI database, CA_004379255.1_SspV1_protein.faa) as a Lamiaceae species reference, with the RepeatMasker option using the RepeatModeler output. (iv) Thereafter, several MAKER iterative rounds were performed. SNAP was used to generate *ab initio* gene prediction models, and AUGUSTUS was trained with the previous MAKER runs to generate a species model. (v) Then, a final MAKER run was performed, followed by Blast2GO[29] and UNIPROT analysis to create a set of reliable protein-coding genes.

2.5. Gene expression analysis

Raw RNA-seq reads were filtered and cleaned using Trimmomatic[30] to remove adapters and the FASTX-Toolkit for (i) trimming read-end nucleotides with quality scores <30 using fastq_quality_trimmer and (ii) removing reads with <70% base pairs with quality scores ≤30 using fastq_quality_filter. Reads were mapped with STAR[31] against the *de novo* sweet basil genome scaffolds. Transcript quantification was performed using Cufflinks (v2.2) software[32] combined with gene MAKER-generated annotations. Differential expression analysis was completed using the R package edgeR.[33]

2.6. Data availability

All sequence data from the project were deposited in NCBI under BioProject ID PRJNA660922. Genomic and RNA-seq raw reads were deposited in the NCBI short-reads archive (SRA) under accessions SRR12568995[34] and SRR12569347,[35] respectively. The genome assembly and gene models were deposited in the CoGe database under genome ID 59011.[36]

3. Results and discussion

3.1. Genome assembly

Altogether, DNA libraries generated about 257 Gbp of short-reads data. Based on the previous estimation of the genome size of the 'Perrie' cultivar as 1.56 Gbp,[7] the data generated represent $165\times$ genome coverage. The short reads were initially used to build 128,921 contigs with an N50 value of ∼45 Kbp (Table 1). Then, the contigs

Table 1. Statistics summary for the contigs and scaffolds

	Contigs	Scaffolds
Total sequences[a]	128,921	12,212
Assembly size (bp)	2,105,853,635	2,133,958,912
Gap size (bp)	—	27,347,277
Gap %	—	1.28
N50 (bp)	45,710	19,298,043
N50 #sequences[b]	12,028	33
N90 (bp)	8,583	5,853,927
N90 #sequences[c]	53,359	107

[a]Number of sequences in the assembly.
[b]Number of sequences composing 50% of the assembly size.
[c]Number of sequences composing 90% of the assembly size.

were assembled by the DeNovoMagic assembler, and the final assembly consisted of 12,212 scaffolds with a total genome size of ∼2.13 Gbp, and less than 1.3% gaps (Table 1). The genome size is in good agreement with that of Bornowski et al.[15] who found a genome size of 2.07 Gbp. The N50 scaffold size was approximately 19 Mbp, and the N90 scaffold size was approximately 5.8 Mbp. A small number of scaffolds (107) comprised more than 90% of the assembled genome (N90 #scaffold; Table 1). Since $2n = 48$ for *O. basilicum*,[6] the N90 #scaffold value suggests that the size of some of the scaffolds is almost that of a complete chromosome, thereby indicating the high quality of the assembly. Bornowski et al.[15] reported an N50 of ∼ 1.5 Mbp with > 17,000 scaffolds greater than 10 Kbp. The apparent differences may be attributed to all/any of the three following reasons: (i) in this study, we used the highly homozygous cultivar, 'Perrie'; (ii) the coverage of $121\times$ used in this study is almost twice the $65\times$ used by Bornowski et al.[15]; and (iii) the assembler used by us has previously been shown to produce highly contiguous genomes of multiple polyploid organisms.[20,37,38] Overall, the draft genome presented here provides a more contiguous version of the sweet basil genome for the use of the scientific community at large. Sequences that could not be placed within the assembly (1.1 million) have a total size of ∼0.34 Gbp and an N50 value of 384 bp (Supplementary Table S1). Screening of the genome for the prevalence of repetitive elements across the scaffolds by using RepeatModeler[28] showed that the basil genome consists of 76% repetitive elements. Long-terminal repeats (LTRs) were the most commonly found types (48% of the genome), and of those, copia-like and gypsy were the most prevalent (Table 2). These values are higher than those reported by Bornowski et al.[15] who found 67% repetitive elements and 37% LTR elements.

3.2. Busco validation

For an independent evaluation of the quality of the assembly, we performed an analysis with the BUSCO pipeline, which is comprised of 1,440 single-copy ortholog genes.[39] Of this set, 93% of the genes were found to be complete, 5.6% were missing and only 1.5% were found to be fragmented (Table 3). It was also found that 80% of the complete BUSCO genes were in a multi-copy state. Of them, 86% were duplicated (Supplementary Table S2). This result indicated that sweet basil is a tetraploid organism, a finding that supports previous estimations.[6,15,40] We further used the BUSCO data to pair homeologous scaffolds that share multiple common genes (Supplementary Table S3), indicating their relatedness to different subgenomes. If not taken into consideration, the redundancy that can arise from tetraploidy can be a major obstacle in designing a genome editing study.

Table 2. Repetitive elements in the basil genome

Repeat type	Counts	Accumulative size (bp)	% of repeats	% of genome
LTR/Copia	376,822	595,520,121	37	28
Unknown	1,001,003	422,700,179	26	20
LTR/Gypsy	260,590	391,107,314	24	18
Simple_repeat	391,284	53,286,505	3	2
DNA/hAT-Ac	103,500	29,188,640	2	1
LTR/ERV1	27,690	22,745,607	1	1
DNA/CMC-EnSpm	47,333	20,879,441	1	1
DNA/MuLE-MuDR	22,466	18,242,016	1	1
LINE/L1	17,940	13,290,589	1	1
RC/Helitron	19,201	8,176,042	1	<0.5
Other non-LTR	152,209	33,025,638	2	2
Other LTR	17,226	12,162,726	1	1
Total LTR	682,328	1,021,535,768	63	48
Total	2,437,264	1,620,324,818	100	76

Table 3. BUSCO statistics summary

Complete BUSCO genes	1339 (93.0%)
Complete BUSCO genes–single copy	267 (18.5%)
Complete BUSCO genes–duplicated	1072 (74.4%)
Fragmented BUSCO genes	21 (1.5%)
Missing BUSCO genes	80 (5.6%)
Total BUSCO genes searched	1440

The report of Navet and Tian[41] indicated that the lack of a reference genome prevented them from evaluating off-targeting when designing gRNAs for CRISPR: Cas9 genome editing study. With the draft sweet basil genome in hand, we will be able to design guide RNAs that will knock out both homeologous genes, as well as to screen for genome-wide off-targets. The reference genome and gene models developed here will also enable us to evade exon–exon junctions when designing guide RNAs, especially if there are large introns in the target gene that harden the sequencing of the relevant genomic region.

3.3. Genome annotations and gene models construction

A total of 157.9 million RNA-seq read-pairs, accounting for about 32 Gbp, were generated (Supplementary Table S4). After the final MAKER[27] run (see Materials and Methods), 74,150 constructed genes were evident. We then used Blast2GO[29] and UNIPROT to account for the protein-coding genes. After proteins smaller than 50 amino acids had been filtered out, it was determined that the basil genome contains 62,067 protein-coding genes (Supplementary File S1). Of them, 10,665 genes have alternative models. These numbers are in a good agreement with the numbers of genes and high-confidence genes found by Bornowski et al.[15] Here, we used ∼ 10 times more RNA-seq data from all plant parts, which plausibly produced more complete and less fragmented gene models. Indeed, the analysis of the BUSCO gene-set in our genome was more complete, with less fragmented and missing genes.[15,16] We also found higher percentage of duplicated genes in the set (74% vs. 56%), suggesting a better capture of homeologous genes. Finally, of the annotated genes, 55,484 proteins were assigned to at least one GO process, function or cellular component (Supplementary Fig. S1).

3.4. Phylogenetic analysis

We used the known systematics of sweet basil within the plant kingdom and within the Lamiaceae family to further validate the data. To this end, we used O. basilicum's closest relative for which the genome has been sequenced, i.e., holy basil, O. sanctum,[8] and two other members of the Lamiaceae family with available published genomes, i.e. the teak tree, Tectona grandis, and scarlet sage, S. splendens.[42,43] We also included in the phylogenetic tree two species belonging to the order Limiales, namely, snapdragon, Antirrhinum majus,[44] and the Asiatic witchweed, Striga asiatica.[45] Another 12 eudicots from nine families and two monocots were included in the analysis to provide a broader phylogenetic view. The determination of phylogenetic relationships between gene sequences was performed with the OrthoFinder interface.[46] Ocimum basilicum fell into the Lamiaceae family within the Lamiales order (Fig. 1). The tree gives genomic certainty to the place of sweet basil within the plant kingdom and within the Lamiaceae. The tree confirmed that the closest species to O. basilicum is O. sanctum. To assess the similarity of the O. basilicum genes to those of O. sanctum, a sequence similarity analysis was performed using the BLAST (Basic Local Alignment Search Tool) algorithm with an e-value cut-off of 10^{-5}. The mean similarity between the proteins of sweet basil and those of holy basil was 72% amino acid identity, with a median of 80%. Another close species to sweet basil in the Ocimum genus is Ocimum americanum, whose genome is yet to be published. Various ploidy levels have been reported for that species,[47] but successful crosses of O. basilicum and O. americanum for breeding purposes have been obtained.[48] Although there is no genome available, there are RNA-seq data for O. americanum in the NCBI database.[49] We used STAR to align the raw reads to the sweet basil genome and found a 78% overall read mapping rate and a 59% concordant pair alignment rate. The above-described validations, taken together, indicate that we are publishing a robust genome and annotations that will serve the scientific community for many purposes, such as breeding programs with inter- and intra-specific crosses, quantitative trait locus (QTL) mapping, development of molecular markers, gene expression experiments and phylogenetic analyses.

3.5. Gene expression analysis using RNA sequencing

Altogether, 87% of the read-pairs were uniquely mapped to the genome, and 9% were mapped to multiple loci (Table 4). The relatively
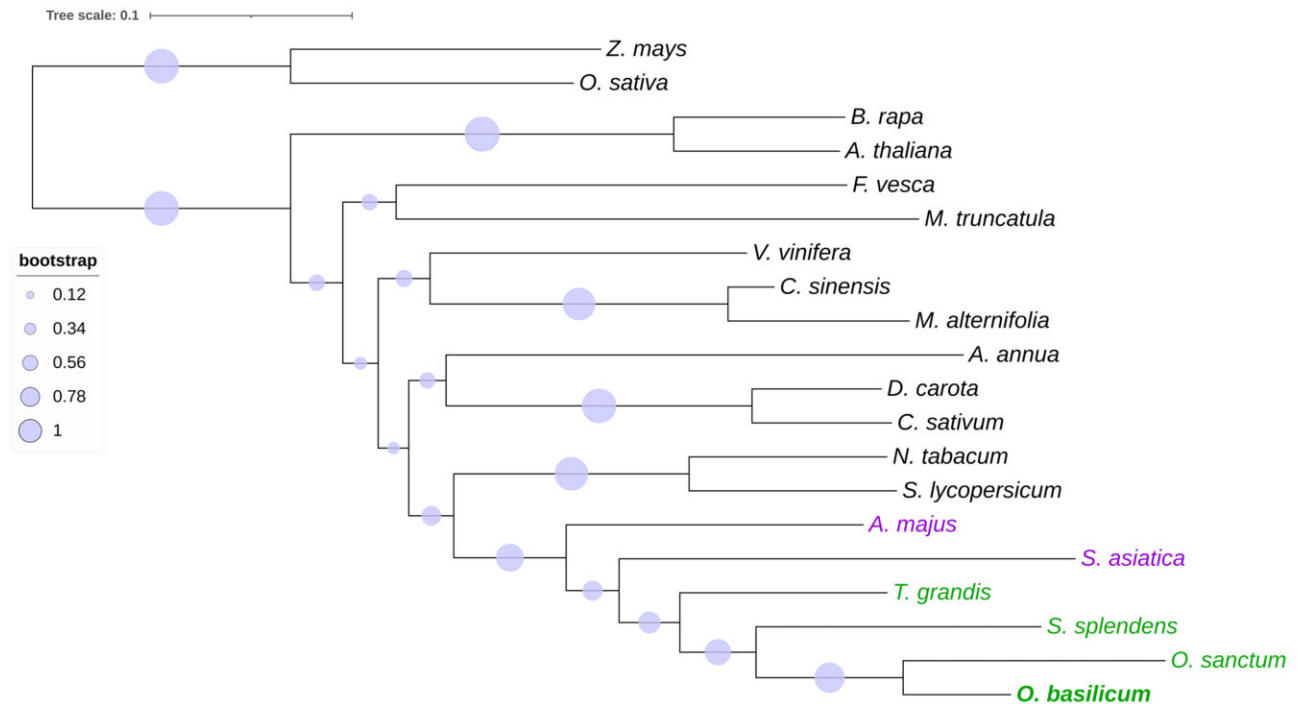
Figure 1. Phylogenetic analyses of the sweet basil genome. A phylogenetic tree depicting the similarity of *O. basilicum* genes with those of other plant species. Members of the Lamiaceae family are shown in green; other members of the Lamiales order are shown in purple. The tree was constructed with OrthoFinder software.[46] The full species names appear in Supplementary Table S5.

Table 4. RNA-sequencing mapping statistics

Tissue	Input read-pairs[a]	% of uniquely mapped reads[b]	% of reads mapped to multiple loci[b]	Overall mapping rate (%)
Leaves	61,790,493	86	10	96
Flowers	59,970,402	87	8	95
Stem	20,034,193	88	8	96
Roots	16,200,077	89	9	98
Total	157,995,165	87	9	96

[a]Numbers indicate read-pairs of the PE libraries. The total number of reads was double.

[b]Successful mapping was when both pair's reads were mapped.

high rate of mapping to multiple loci is probably due to basil tetraploidy, with fragments of homeologous genes being highly similar to each other. The overall mapping rate of all RNA-seq reads was 96% (Table 4), indicating very good continuity of the assembly. This high mapping rate in comparison to the 83% of read-pairs mapped by Bornowski et al.[15] to their genome is probably a result of the higher contiguity and homozygosity of our genome. Next, we evaluated gene expression levels in the different tissues and looked for differentially expressed genes. Overall, 1,029 to 8,832 genes were differentially expressed between at least two different tissues (*p*-adjust < 0.001, fold change > 4) (Table 5).

3.6. Phenylpropanoid volatiles biosynthesis pathway in tetraploid sweet basil

Eugenol, accumulated by Genovese-type basils such as 'Perrie', and methyl chavicol, accumulated by Thai-type basils such as 'Cardinal', determine the aroma, and hence the chemotype, of the cultivar. The biosynthesis pathway of these volatiles is partially known (Fig. 2), and four genes and four aroma genes have previously been

Table 5. Differentially expressed genes in the different basil tissues

Tissue	Leaves	Stem	Roots
Flowers	8,832	5,473	7,472
Leaves		6,990	7,267
Stem			1,029

characterized.[50–53] Here, we analysed their prevalence in the two subgenomes and estimated the redundancy effect that could influence genome editing attempts. We used BLASTN to assign the known genes to their genomic location and to their Gene ID. The coding genes of the four enzymes, *p*-coumaroyl shikimate 3′-hydroxylase (CS3′H),[50] eugenol synthase (EGS),[51] coniferyl alcohol acetyltransferase (CAAT)[52] and eugenol O-methyl-transferase (EOMT),[53] were characterized for eugenol/methyl eugenol biosynthesis.
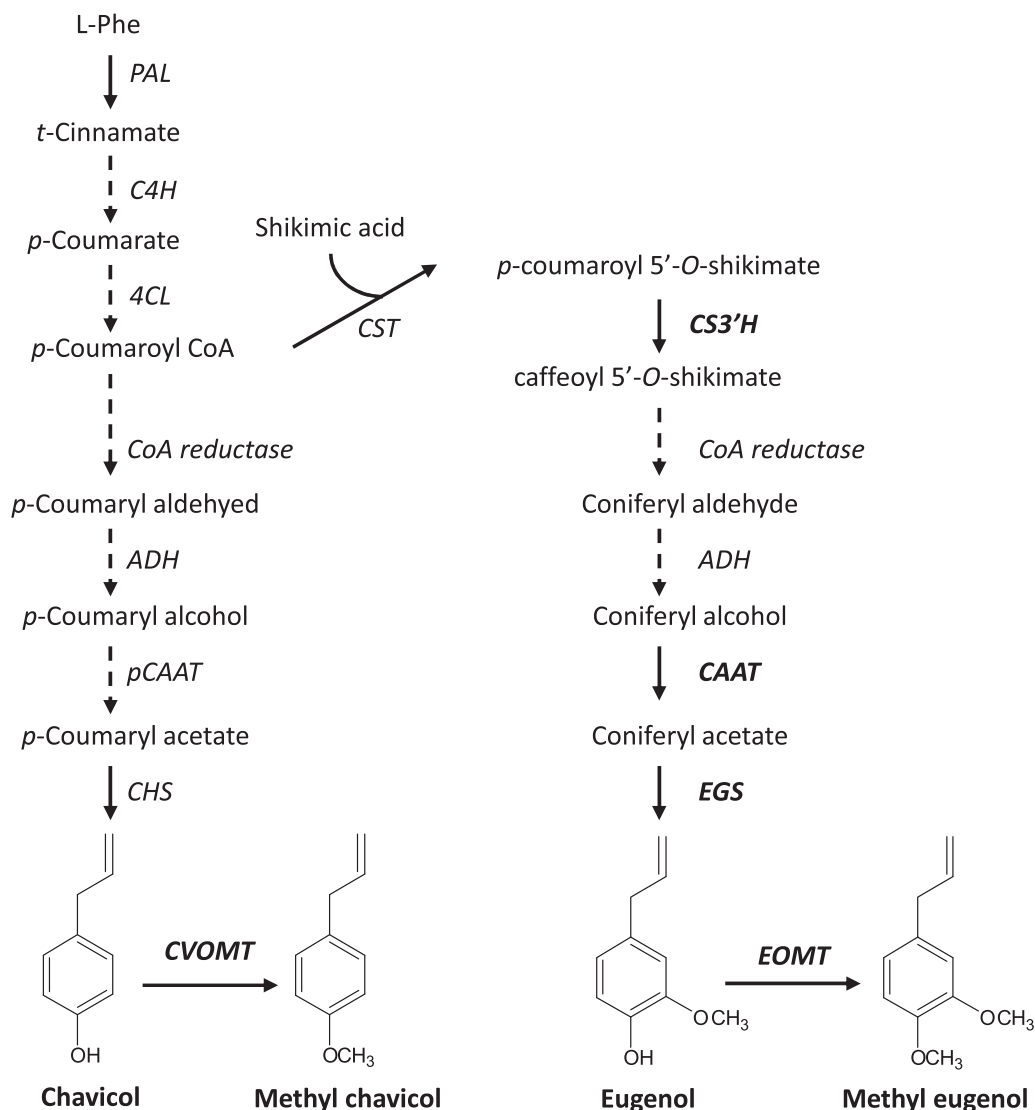
Figure 2. Biosynthetic pathways of phenylpropanoid volatiles in basil. Solid arrows represent reactions that have been demonstrated in basil. Bold enzyme names represent enzymes whose encoding genes have been characterized from basil. Bold compound names represent volatile compounds found in basil essential oil. PAL, phenylalanine ammonia lyase; C4H, *t*-cinnamate 4-hydroxylase; 4CL, *p*-coumarate CoA ligase; CST, *p*-coumaroyl-CoA: shikimic acid *p*-coumaroyl transferase; CS3′H, *p*-coumaroyl shikimate 3′-hydroxylase; EGS, eugenol synthase; ADH, alcohol dehydrogenase; CAAT, coniferyl alcohol acetyltransferases; pCAAT, *p*-coumaryl alcohol acetyltransferases; CHS, chavicol synthase; EOMT, eugenol *O*-methyltransferase; CVOMT, chavicol *O*-methyltransferase; CoA, coenzyme A.

3.6.1. *p-Coumaroyl shikimate 3′-hydroxylase*

Gang et al.[50] characterized two genes encoding CS3′H enzymes, namely, *ObCS3′H1* and *ObCS3′H2* (GenBank accession numbers AY082611 and AY082612, respectively), which are 98.2% homologous. BLASTN analysis indicated two scaffolds, 33 and 243, harboring one *CS3′H* gene, each with three exons (Supplementary Fig. S2A). These two scaffolds are considered homeologous since they share six BUSCO genes (Supplementary Table S3). Homology analysis of these genes, XLOC_034008 (in scaffold 33) and XLOC_024688 (in scaffold 234), with the previously characterized genes *ObCS3′H1* and *ObCS3′H2*, showed that the gene in scaffold 33 is *ObCS3′H1* (99.5% homology) and the gene in scaffold 234 is *ObCS3′H2* (99.1% homology) (Supplementary Fig. S2B). The minor differences are probably due to the different cultivars used for the genome construction and the original *ObCS3′H* study.[50] The encoded

proteins were found to be highly identical, with only minor amino acid substitutions (Supplementary Fig. S3). The sequenced genome and the RNA-seq performed in this study enabled a gene-specific expression analysis showing that both genes were expressed in all sampled tissues, namely, leaves, flowers, stems and roots (Fig. 3A). To check for expression bias, we performed a paired Student's *t*-test across all eight samples and found a bias toward *ObCS3′H2* ($\alpha < 0.01$); nonetheless, it seems that both homeologous genes might contribute to the CS3′H enzymatic activity in the phenylpropanoid biosynthesis pathway (Fig. 3A).

3.6.2. *Eugenol synthase*

Only one EGS encoding gene, *ObEGS1* (GenBank accession number DQ372812), was characterized by Koeduka et al.[51] However, as for *ObCS3′H*, BLASTN analysis indicated two *EGS* genes, one in
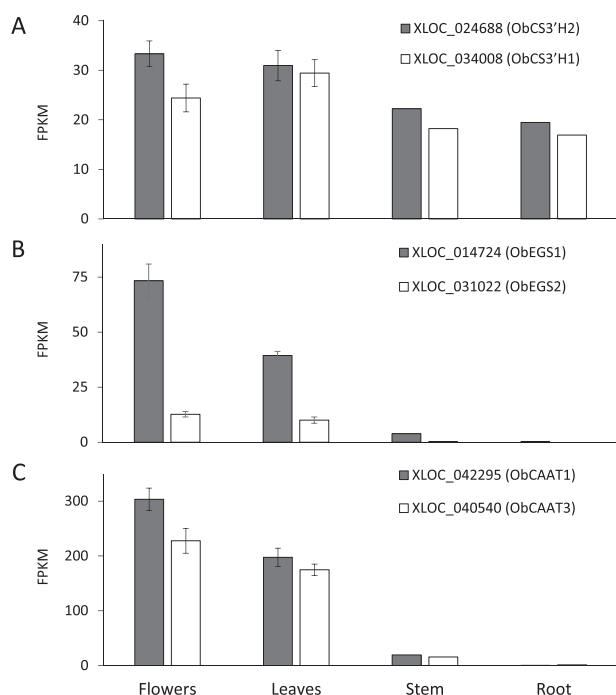
Figure 3. Expression of phenylpropanoid volatiles biosynthetic genes in sweet basil. The normalized gene expression of: (A) *p-coumaroyl shikimate 3′-hydroxylases* (CS3′H); (B) *eugenol synthases* (EGS); (C) *coniferyl alcohol acetyltransferases* (CAAT). Gene expression levels were determined based on RNA-seq data obtained [fragments per kilobase of transcript per million mapped reads (fpkm)]. Values are means of three biological repeats ± SE (leaves and flowers), or of one biological repeat (stem and roots). *ObCS3′H1/2* are the genes characterized by Gang et al.[50]; *ObEGS1* is the gene characterized by Koeduka et al.[51]; and *ObCAAT1* is the gene sequenced and silenced by Dhar et al.[52]

scaffold 162 and the other in scaffold 3082, each of which has five exons (Supplementary Fig. S4A). These two scaffolds share only three BUSCO genes (Supplementary Table S3), but since scaffold 3082 is very small (1 Mbp), those scaffolds were considered as homeologous. Multiple sequence alignment of these two genes, XLOC_014724 (in scaffold 162) and XLOC_031022 (in scaffold 3082), with the characterized gene *ObEGS1*, showed that the gene in scaffold 3082 is truncated, while the gene in scaffold 162 is 99.8% homologous with *ObEGS1* (Supplementary Fig. S4B). Amino acid alignment of the encoded proteins revealed that *ObEGS1* and *EGS_162* are 100% identical, but *EGS_3082*, termed *ObEGS2*, is truncated (Supplementary Fig. S5), as a result of a frameshift due to a 2-bp insertion (Supplementary Fig. S4B). We utilized the sequenced genome and the RNA-seq data to perform gene-specific expression analysis and found that both *ObEGS1* and the truncated *ObEGS2* are expressed mainly in the leaves and the flowers (Fig. 3B). A paired Student's *t*-test across all eight samples showed a bias in expression toward *ObEGS1* ($\alpha < 0.01$). The fact that the truncated *ObEGS2* was expressed raises the question of whether this protein encodes an active enzyme. The question of whether both homeologous genes contribute to eugenol biosynthesis thus remains open for further research and discussion.
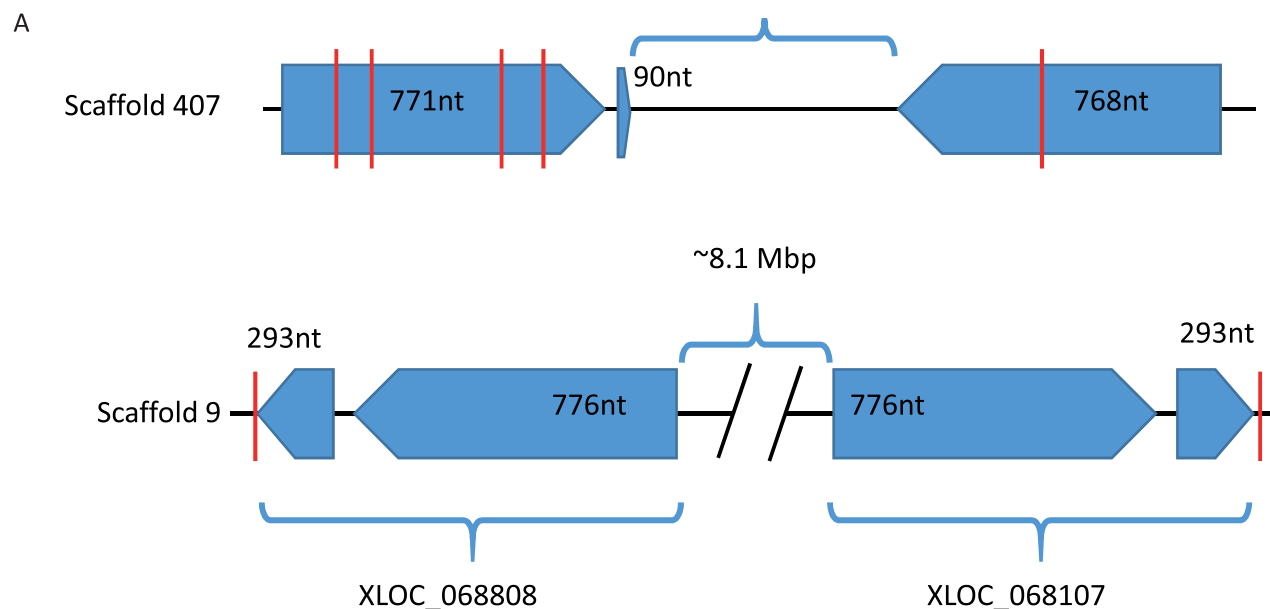
3.6.3. Coniferyl alcohol acetyltransferase

We used our sequenced genome and expression data to explain the results of a recent RNAi-based functional characterization of sweet

basil CAAT, which showed only an approximately 50% reduction both in gene expression and in eugenol accumulation.[52] We thus investigated whether the knocked-down gene, *ObCAAT1*, might be redundant in the sweet basil genome, causing residual enzymatic activity. We found two genes homologous to *ObCAAT1*, XLOC_042295 located on scaffold 4172, which is 100% identical to *ObCAAT1*, and XLOC_040540 (termed *ObCAAT3*) located on scaffold 407, which is 97% identical to *ObCAAT1* (Supplementary Fig. S6). These scaffolds are plausibly homeologous, since they share 12 BUSCO genes (Supplementary Table S3). The amino acid sequences of the two proteins suggest that they are functional, since they do not have a premature stop codon and since the critical DFGWG and HxxxD motifs[54] are conserved in both proteins (Supplementary Fig. S7). Both genes were expressed in high levels in leaf and flower tissues (Fig. 3C). Although a paired *t*-test showed bias toward *ObCAAT1* expression ($\alpha < 0.05$), the expression of *ObCAAT3* supported its role in coniferyl acetate biosynthesis. We speculate that the RNAi fragment designed by Dhar et al.[52] was specific to *ObCAAT1* but did not alter the expression of *ObCAAT3*, thereby causing only a partial reduction in eugenol biosynthesis. In contrast, the primers designed for qRT-PCR may not distinguish between the two transcripts. These findings demonstrate how the genome and annotations published here will help researchers to overcome similar difficulties in future studies.

3.6.4. Eugenol O-methyl-transferase

Two basil *O*-methyl transferase (OMT) genes have been characterized, EOMT (GenBank accession number AF435008) and chavicol *O*-methyl transferase (CVOMT; GenBank accession number AF435007). Both these genes encode proteins that are 90% identical.[53] A single amino acid, serine 261 or phenylalanine 260, determines the enzyme's role as EOMT or CVOMT, respectively. The genomic context of these two similar genes is unknown, and whether they are two alleles of the same gene, homologous genes or homeologous genes is yet to be discovered. Since 'Perrie' is a eugenol-type basil, we used *ObEOMT* as a query for BLASTN analysis against the genome. Four putative locations showed more than 90% homology with the *ObEOMT* gene, two on scaffold 9 and two on scaffold 407. These scaffolds are not considered homeologous, as they do not share any BUSCO gene. On scaffold 407, the putative genes are located 956 bp apart and in opposite directions (Fig. 4A). However, they are both truncated to a size of ∼750 bp, while *ObEOMT* is 1,074 nucleotides in size. Moreover, there are several stop codons in both genes (Fig. 4A), indicating inactive proteins, if translated. On scaffold 9, the two genes (XLOC_068107 and XLOC_068808), composed of two exons each, are located in opposite directions ∼8.1 Mbp apart (Fig. 4A). They are 98% homologous, but only ∼92% homologous to *ObEOMT* and ∼96% to *ObCVOMT*. On the amino acid level, they are 97.5% identical, ∼90% identical to ObEOMT and ∼96.5% identical to ObCVOMT. Although more similar to ObCVOMT, the catalytic serine 261 residue of ObEOMT is conserved in both proteins (Fig. 4B). This serine residue, which determines the nature of the OMT enzyme to catalyze methyl eugenol but not methyl chavicol biosynthesis, is coherent with the eugenol chemotype of 'Perrie' cultivar. Given the truncated OMT genes on scaffold 407, we conclude that only a single subgenome contributes to methyl eugenol accumulation in sweet basil. Similar findings were obtained for EGS, but whether the active ObEOMT and ObEGS are positioned on the same subgenome is yet to be determined. Moreover, since the 'Perrie' genome does not harbor an OMT

Figure 4. Genomic and sequence analysis of sweet basil *O-methyl transferase* (*OMT*) genes. A. The genomic locations of sweet basil OMT gene, as evident from tblastn analysis with the *ObEOMT* gene. B. Multiple sequence alignment of ObEOMT, XLOC_068107 and XLOC_068808. Alignment was carried out with ClustalOmega (https://www.ebi.ac.uk/Tools/msa/clustalo/) and visualized with BoxShade3.21 (https://embnet.vital-it.ch/software/BOX_form.html). Black shaded amino acids are identical in all three proteins, and gray shaded amino acids represent similar amino acids. Yellow shaded serine is the residue dictating the formation of methyl eugenol rather than methyl chavicol. Color shaded amino acids are conserved in the OMT family. ObEOMT is the protein characterized by Gang et al.[53]

coding gene with catalytic phenylalanine to promote methyl chavicol production, it is reasonable to conclude that *ObEOMT* and *ObCVOMT* are two alleles of the same gene(s). In our dataset, extracted from the seventh and eighth leaf-pair positions, there was no expression of *ObEOMT* genes, a finding that is in keeping with our prior observation that the 'Perrie' cultivar accumulates methyl eugenol exclusively in the first and second leaf-pair positions.[55]

4. Concluding remarks

The draft genome of sweet basil published here illustrates the importance of a contiguous reference genome for this tetraploid plant. The high contiguity of the genome assembled here enabled us to take the initial steps toward a complete homeologous analysis. The dissection of the phenylpropanoid volatiles biosynthetic pathway into homeologous scaffolds provides essential information for future studies

focusing on pathway discoveries, genome editing and molecular breeding. Such an analysis has been shown to be essential for overcoming the polyploids inherent redundancy, which is an obstacle to achieve successful gene(s) knockout in transgenic and gene-edited plants.

Supplementary data

Acknowledgments

Accession numbers

All sequence data from the project were deposited in NCBI under BioProject ID PRJNA660922. Genomic and RNA-seq raw reads were deposited in the NCBI short-reads archive (SRA) under accessions SRR12568995 and SRR12569347, respectively. The genome assembly and gene models were deposited in the CoGe database under genome ID 59011.

Conflict of interest

None declared.

Author contributions

IG, AF, KB and ND wrote the paper. RM and IG were responsible for growing the plants. MJK and RM extracted the high-molecular-weight DNA. CA extracted the RNA. GR and KB assembled the genome. AF created the gene models, performed all RNA-seq alignments and analyses, and constructed the phylogenetic tree. DC, AS and ND stabilized the 'Perrie' cultivar over the years. IG and ND conceived and designed the project.

References

1. Paton, A., Harley, R.M. and Harley, M.M. 1999, Ocimum: an overview of classification and relationships. In: Hiltunen, R. and Holm, Y. (eds), Basil, Harwood Academic, Amsterdam, The Netherlands, pp. 1–32.

2. UPOV 2003, Basil. In: *Plants*. International Union for the Protection of New Varieties of Plants, Geneva. https://www.yumpu.com/en/document/read/21898520/basil-international-union-for-the-protection-of-new-varieties-of-plants

3. Dudai, N. and Belanger, F.C. 2016, Aroma as a factor in the breeding process of fresh herbs – the case of basil. In: Dudai, N. and Belanger, F.C. (eds), *Biotechnology in Flavor Production*, John Wiley & Sons Ltd, Chichester, West Sussex, UK, pp. 32–61.

4. Gang, D.R., Wang, J., Dudareva, N., et al. 2001, An investigation of the storage and biosynthesis of phenylpropenes in sweet basil, *Plant Physiol.*, **125**, 539–55.

5. Dolezel, J., Bartos, J., Voglmayr, H. and Greilhuber, J. 2003, Nuclear DNA content and genome size of trout and human, *Cytometry A.*, **51**, 127–8.

6. Carović-Stanko, K., Liber, Z., Besendorfer, V., et al. 2010, Genetic relations among basil taxa (Ocimum L.) based on molecular markers, nuclear DNA content, and chromosome number, *Plant Syst. Evol.*, **285**, 13–22.

7. Koroch, A.R., Wang, W., Michael, T.P., Dudai, N., Simon, J.E. and Belanger, F.C. 2010, Estimation of nuclear DNA content of cultivated Ocimum species by using flow cytometry, *Isr. J. Plant Sci.*, **58**, 183–9.

8. Rastogi, S., Kalra, A., Gupta, V., et al. 2015, Unravelling the genome of Holy basil: an "incomparable" "elixir of life" of traditional Indian medicine, *BMC Genomics.*, **16**, 413.

9. Upadhyay, A.K., Chacko, A.R., Gandhimathi, A., et al. 2015, Genome sequencing of herb Tulsi (Ocimum tenuiflorum) unravels key genes behind its strong medicinal properties, *BMC Plant Biol.*, **15**, 212.

10. Pushpangadan, P. and Sobti, S.N. 1982, Cytogenetical studies in the genus Ocimum, *Cytologia.*, **47**, 575–83.

11. Paton, A. and Putievsky, E. 1996, Taxonomic problems and cytotaxonomic relationships between and within varieties of *Ocimum basilicum* and related species (Labiatae), *Kew Bull.*, **51**, 509–24.

12. Mukherjee, M., Datta, A.K. and Maiti, G.G. 2005, Chromosome number variation in *Ocimum basilicum* L, *Cytologia.*, **70**, 455–8.

13. Mukherjee, M. and Datta, A.K. 2005, Secondary chromosome associations in *Ocimum basilicum* L. and Ocimum tenuiflorum L, *Cytologia.*, **70**, 149–52.

14. Rastogi, S., Meena, S., Bhattacharya, A., et al. 2014, De novo sequencing and comparative analysis of holy and sweet basil transcriptomes, *BMC Genomics.*, **15**, 588.

15. Bornowski, N., Hamilton, J.P., Liao, P., Wood, J.C., Dudareva, N. and Buell, C.R. 2020, Genome sequencing of four culinary herbs reveals terpenoid genes underlying chemodiversity in the Nepetoideae, *DNA Res.*, **27**, doi: 10.1093/dnares/dsaa1016.

16. Dudai, N., Carp, M.-J., Milavski, R., et al. 2018, High-quality assembly of sweet basil genome. *bioRxiv*, 476044.

17. Healey, A., Furtado, A., Cooper, T. and Henry, R.J. 2014, Protocol: a simple method for extracting next-generation sequencing quality genomic DNA from recalcitrant plant species, *Plant Methods.*, **10**, 21.

18. Lu, F., Romay, M.C., Glaubitz, J.C., et al. 2015, High-resolution genetic mapping of maize pan-genome sequence anchors, *Nat. Commun.*, **6**, 6914.

19. Hirsch, C.N., Hirsch, C.D., Brohammer, A.B., et al. 2016, Draft assembly of elite inbred line PH207 provides insights into genomic and transcriptome diversity in maize, *Plant Cell.*, **28**, 2700–14.

20. Avni, R., Nave, M., Barad, O., et al. 2017, Wild emmer genome architecture and diversity elucidate wheat evolution and domestication, *Science.*, **357**, 93–7.

21. Luo, M.C., Gu, Y.Q., Puiu, D., et al. 2017, Genome sequence of the progenitor of the wheat D genome Aegilops tauschii, *Nature.*, **551**, 498–502.

22. Zhao, G., Zou, C., Li, K., et al. 2017, The Aegilops tauschii genome reveals multiple impacts of transposons, *Nat. Plants.*, **3**, 946–55.

23. Gonda, I., Ashrafi, H., Lyon, D.A., et al. 2019, Sequencing-based bin map construction of a tomato mapping population, facilitating high-resolution quantitative trait loci detection, *Plant Genome.*, **12**, 180010.

24. Grabherr, M.G., Haas, B.J., Yassour, M., et al. 2011, Full-length transcriptome assembly from RNA-Seq data without a reference genome, *Nat. Biotechnol.*, **29**, 644–52.

25. Korf, I. 2004, Gene finding in novel genomes, *BMC Bioinformatics.*, **5**, 59.

26. Stanke, M., Steinkamp, R., Waack, S. and Morgenstern, B. 2004, AUGUSTUS: a web server for gene finding in eukaryotes, *Nucleic Acids Res.*, **32**, W309–12.

27. Cantarel, B.L., Korf, I., Robb, S.M.C., et al. 2007, MAKER: an easy-to-use annotation pipeline designed for emerging model organism genomes, *Genome Res.*, **18**, 188–96.

28. Smit, A., Hubley, R. and Green, P. 2013-2015, RepeatMasker Open-4.0. http://www.repeatmasker.org.

29. Conesa, A., Götz, S., García-Gómez, J.M., Terol, J., Talón, M. and Robles, M. 2005, Blast2GO: a universal tool for annotation, visualization and analysis in functional genomics research, *Bioinformatics.*, **21**, 3674–6.

30. Bolger, A.M., Lohse, M. and Usadel, B. 2014, Trimmomatic: a flexible trimmer for Illumina sequence data, *Bioinformatics.*, **30**, 2114–20.

31. Dobin, A., Davis, C.A., Schlesinger, F., et al. 2013, STAR: ultrafast universal RNA-seq aligner, *Bioinformatics.*, **29**, 15–21.

32. Roberts, A., Trapnell, C., Donaghey, J., Rinn, J.L. and Pachter, L. 2011, Improving RNA-Seq expression estimates by correcting for fragment bias, *Genome Biol.*, **12**, R22.

33. Robinson, M.D., McCarthy, D.J. and Smyth, G.K. 2010, edgeR: a Bioconductor package for differential expression analysis of digital gene expression data, *Bioinformatics.*, **26**, 139–40.

34. NCBI Sequence Read Archive. 2020, Sweet basil WGS. https://www.ncbi.nlm.nih.gov/sra/?term=SRR12568995.

35. NCBI Sequence Read Archive. 2020, Sweet basil RNAseq. https://www.ncbi.nlm.nih.gov/sra/?term=SRR12569347.

36. CoGe: Comparative Genomics. 2020, Sweet basil genome. https://genomevolution.org/coge/GenomeInfo.pl?gid=59011.

37. Colle, M., Leisner, C.P., Wai, C.M., et al. 2019, Haplotype-phased genome and evolution of phytonutrient pathways of tetraploid blueberry, *Gigascience.*, **8**, giz012.

38. Chen, X., Lu, Q., Liu, H., et al. 2019, Sequencing of cultivated peanut, arachis hypogaea, yields insights into genome evolution and oil improvement, *Mol. Plant.*, **12**, 920–34.

39. Simao, F.A., Waterhouse, R.M., Ioannidis, P., Kriventseva, E.V. and Zdobnov, E.M. 2015, BUSCO: assessing genome assembly and annotation completeness with single-copy orthologs, *Bioinformatics.*, **31**, 3210–2.

40. Pushpangadan, P. and Bradu, B.L. 1995, Basil. In: Chadha, K. L. and Gupta, R. (eds), *Advances in Horticulture. Medicinal and Aromatic Plants*, Malhotra Publishing House, New Delhi.

41. Navet, N. and Tian, M. 2020, Efficient targeted mutagenesis in allotetraploid sweet basil by CRISPR/Cas9, *Plant Direct.*, **4**, e00233.

42. Zhao, D., Hamilton, J.P., Bhat, W.W., et al. 2019, A chromosomal-scale genome assembly of Tectona grandis reveals the importance of tandem gene duplication and enables discovery of genes in natural product biosynthetic pathways, *Gigascience.*, **8**, giz005.

43. Dong, A.X., Xin, H.B., Li, Z.J., et al. 2018, High-quality assembly of the reference genome for scarlet sage, *Salvia splendens*, an economically important ornamental plant. *Gigascience*, **7**, giy068.

44. Li, M., Zhang, D., Gao, Q., et al. 2019, Genome structure and evolution of Antirrhinum majus L, *Nat. Plants.*, **5**, 174–83.

45. Yoshida, S., Kim, S., Wafula, E.K., et al. 2019, Genome sequence of Striga asiatica provides insight into the evolution of plant parasitism, *Curr. Biol.*, **29**, 3041–3052.e4.

46. Emms, D.M. and Kelly, S. 2015, OrthoFinder: solving fundamental biases in whole genome comparisons dramatically improves orthogroup inference accuracy, *Genome Biol.*, **16**, 157.

47. Rewers, M. and Jedrzejczyk, I. 2016, Genetic characterization of Ocimum genus using flow cytometry and inter-simple sequence repeat markers, *Ind. Crops Prod.*, **91**, 142–51.

48. Ben-Naim, Y., Falach, L. and Cohen, Y. 2018, Transfer of downy mildew resistance from wild basil (Ocimum americanum) to sweet basil (O. basilicum), *Phytopathology.*, **108**, 114–23.

49. Zhan, X., Yang, L., Wang, D., Zhu, J.K. and Lang, Z. 2016, De novo assembly and analysis of the transcriptome of Ocimum americanum var. pilosum under cold stress, *BMC Genomics.*, **17**, 209.

50. Gang, D.R., Beuerle, T., Ullmann, P., Werck-Reichhart, D. and Pichersky, E. 2002, Differential production of meta hydroxylated phenylpropanoids in sweet basil peltate glandular trichomes and leaves is controlled by the activities of specific acyltransferases and hydroxylases, *Plant Physiol.*, **130**, 1536–44.

51. Koeduka, T., Fridman, E., Gang, D.R., et al. 2006, Eugenol and isoeugenol, characteristic aromatic constituents of spices, are biosynthesized via reduction of a coniferyl alcohol ester, *Proc. Natl. Acad. Sci. U S A.*, **103**, 10128–33.

52. Dhar, N., Sarangapani, S., Reddy, V.A., et al. 2020, Characterization of a sweet basil acyltransferase involved in eugenol biosynthesis, *J. Exp. Bot.*, **71**, 3638–52.

53. Gang, D.R., Lavid, N., Zubieta, C., et al. 2002, Characterization of phenylpropene O-methyltransferases from sweet basil, *Plant Cell.*, **14**, 505–19.

54. D'Auria, J.C. 2006, Acyltransferases in plants: a good time to be BAHD, *Curr. Opin. Plant Biol.*, **9**, 331–40.

55. Fischer, R., Nitzan, N., Chaimovitsh, D., Rubin, B. and Dudai, N. 2011, Variation in essential oil composition within individual leaves of sweet basil (*Ocimum basilicum* L.) is more affected by leaf position than by leaf age, *J. Agric. Food Chem.*, **59**, 4913–22.