Tree Genetics & Genomes (2021) 17: 46
https://doi.org/10.1007/s11295-021-01528-5

ORIGINAL ARTICLE

The haplotype-resolved reference genome of lemon (*Citrus limon* L. Burm f.)

Mario Di Guardo[1] · Marco Moretto[2] · Mirko Moser[2] · Chiara Catalano[1] · Michela Troggio[2] · Ziniu Deng[3] · Alessandro Cestaro[2] · Marco Caruso[4] · Gaetano Distefano[1] · Stefano La Malfa[1] · Luca Bianco[2] · Alessandra Gentile[1,3]

Received: 27 July 2021 / Revised: 27 October 2021 / Accepted: 31 October 2021 / Published online: 9 November 2021

Abstract

Lemon (*Citrus limon* (L.) Burm. f.) is an evergreen tree belonging to the genus *Citrus*. The fruits are particularly prized for the organoleptic and nutraceutical properties of the juice and for the quality of the essential oils in the peel.
Herein, we report, for the first time, the release of a high-quality reference genome of the two haplotypes of lemon. The sequencing has been carried out coupling Illumina short reads and Oxford Nanopore data leading to the definition of a primary and an alternative assembly characterized by a genome size of 312.8 Mb and 324.74 Mb respectively, which agree well with an estimated genome size of 312 Mb. The analysis of the transposable element (TE) allowed the identification of 2878 regions on the primary and 2897 on the alternative assembly distributed across the nine chromosomes. Furthermore, an in silico analysis of the microRNA genes was carried out using 246 mature miRNA and the respective pre-miRNA hairpin sequences of *Citrus sinensis*. Such analysis highlighted a high conservation between the two species with 233 mature miRNAs and 51 pre-miRNA stem-loops aligning with perfect match on the lemon genome.
In parallel, total RNA was extracted from fruit, flower, leaf, and root enabling the detection of 35,020 and 34,577 predicted transcripts on primary and alternative assemblies respectively. To further characterize the annotated transcripts based on their function, a gene ontology and a gene orthology analysis with other *Citrus* and *Citrus*-related species were carried out. The availability of a reference genome is an important prerequisite both for the setup of high-throughput genotyping analysis and for functional genomic approaches toward the characterization of the genetic determinism of traits of agronomic interest.

Keywords Citrus · Haplotype-resolved assembly · Gene annotation · Oxford Nanopore

Introduction

Lemon (*Citrus limon* (L.) Burm. f.) is ranked third for cultivated area in the world among the *Citrus* species and, together with lime, their global harvested area is more than one million hectares (FAOSTAT 1997). Lemon cultivation has long been restricted to the coastal areas of Spain, Greece, and Italy, while, nowadays, lemons are cultivated all over the world in areas characterized by Mediterranean-type climate (mild winters and warm and dry summers) such as Argentina, Brazil, China, Mexico, Southern California, South Africa. Lemon is grown for fresh consumption and for processing. Fruits are widely priced for their quality in terms of flavor and nutraceutical value thanks to the high content in components such as citric acid, vitamin C, flavonoids, and minerals (Sun et al. 2019; Muccilli et al. 2020) and for the essential oils in the peel (Mehl et al. 2014).

Communicated by D. Chagné

Mario Di Guardo, Marco Moretto and Mirko Moser contributed equally to this work and share first authorship.

✉ Gaetano Distefano
distefag@unict.it

✉ Luca Bianco
luca.bianco@fmach.it

[1] Department of Agriculture, Food and Environment (Di3A), University of Catania, via Valdisavoia 5, 95123 Catania, Italy

[2] Research and Innovation Centre, San Michele All' Adige, Fondazione Edmund Mach, Trento, Italy

[3] College of Horticulture and Landscape, Hunan Agricultural University, Changsha 410128, China

[4] Research Centre for Olive, Fruit and Citrus Crops, CREA, Corso Savoia 190, 95024 Acireale, Italy

Even though the center of origin of lemon is still unknown, the species likely originated in an area comprising Eastern Himalaya, Middle East, and India (Singh 1981). This hypothesis is supported by the fact that such an area is characterized by the occurrence of lemons growing in a wild state (and bearing high-quality fruit) and that most of the known *Citrus* species originated from the same area as well.

The use of molecular markers and whole genome sequencing (WGS) approaches provided fundamental insights to decipher the phylogeny of the genus *Citrus*. Cultivated citrus species are interspecific hybrid and/or admixture of three founder species (i.e., accessions without interspecific admixture): citron (*Citrus medica* L.), pummelo (*Citrus maxima* (Burm.) Merr.), and mandarin (*Citrus reticulata* Blanco) (Wu et al. 2014). Lemon is a sour orange × citron hybrid, with the female parent being a F_1 hybrid between pummelo and a pure mandarin. Analysis of cpDNA highlighted that lemon shares the chloroplast of pummelo through sour oranges (Nicolosi et al. 2000; Gulsen and Roose 2001; Carbonell-Caballero et al. 2015). WGS approaches confirmed the hybrid origin of lemon and sour orange, both characterized by high heterozygosity coupled with low interspecific diversity (Wu et al. 2018). The same study further elucidated the contribution of the three founders on the genetic makeup of lemon with citron, mandarin, and pummelo characterizing the 50%, 19%, and 31% of the lemon genome (Wu et al. 2018).

Even though several *Citrus* species were sequenced and assembled in the last years, like in the case of Clementine (Wu et al. 2014), pummelo, citron (Wang et al. 2017) and mandarin (Wang et al. 2018), the reference genome of lemon is not yet publicly available. In this paper we present the de novo sequencing of the cv 'Femminello Siracusano' lemon. The group of lemons named 'Femminello' represents the most cultivated, Italian lemon variety (Barry et al. 2020). 'Femminello' cultivars usually show a marked ever-blooming and ever-bearing habit, which is particularly priced for the out-of-season production of *verdelli* lemons during the summer months (Barry et al. 2020). Among the 'Femminello' group, 'Femminello Siracusano' is one of the most widely cultivated thanks to the quality of the fruits coupled with the high yield.

The de novo assembled contigs obtained from the combination of Illumina short reads and Oxford Nanopore long reads resulted were then arranged in nine pseudomolecules that were built by synteny with the *C. maxima* genome (Wang et al. 2017), the only chromosome-scale reference genome among the ancestors of lemon. The assembly was complemented by a de novo annotation of the genes and of the non-coding regions (long terminal repeats and sRNA) of the genome.

Materials and methods

DNA extraction and genome sequencing

Young leaves of the *Citrus limon* cultivar 'Femminello Siracusano' were sampled at the experimental farm of the University of Catania (Italy) in 2020. Total genomic DNA was isolated using the NucleoSpin Plant II Midi kit (Macherey Nagel, Germany). The same plant was also employed to collect leaf, flower, fruit, and root tissues for RNA-seq analysis. Roots were sampled from self-rooted cuttings grown in the MS medium. Total RNA was extracted using the RNeasy Plant Power kit (Qiagen, Hilden, Germany) following the manufacturer protocol and tested for quality. As for the fruit, the total RNA was extracted separately from flavedo, albedo, and pulp and then combined for mRNA sequencing. Library preparation for Illumina sequencing was performed using standard Illumina protocols and Illumina paired-ends adapters to generate reads of 2 × 150 nucleotides.

Long reads sequencing was performed using a modified version of the Oxford Nanopore (ONT) Ligation sequencing kit (SQK-LSK109). Briefly, 2 µg of extracted DNA was diluted to a final volume of 47 µl using nuclease free water. The DNA repair and end-prep step mix was then prepared as described in the 1D Lambda Control Experiment protocol (ONT) with an incubation time of 30' at 20 °C followed by 10' at 65 °C. Then, 50 µl of AMPure XP beads was added to the 60-µl reaction and incubated 40' at RT under gentle mixing. The washing steps were performed as described in the ONT protocol, but the elution step was extended to 20'. Similarly, the ligation was performed as described, but the incubation extended to 20'. Subsequently, 40 µl of AMPure XP beads was added, and, again, the mix incubated under gentle mixing for 20'. The remaining steps were performed as described in the ONT protocol with the final elution step extended to 20'. The preparation of the sample for the loading on the flow cell (R10.3 and R9.4.3) was performed as described by ONT. Basecalling was performed with Guppy (ONT trademark) software version 4.5.3.

Kmer analysis and genome size estimation

The Illumina reads were analyzed to estimate the genome size and the level of heterozygosity. In particular, the kmer spectrum with kmer size of 21 was computed from the reads with Jellyfish v.2.3.0 (commands 'count' and 'histo') and then fed to the online tool genomescope (http://qb.cshl.edu/

genomescope/; Vurture et al. 2017) with maximum kmer coverage set to 6000.

Genome assembly and curation

The genome sequence assembly was initially performed using MaSuRCA v.3.4.1 (Zimin et al. 2017) as it is a hybrid approach that directly combines the benefits of long reads with the accuracy of short reads. Several assemblies were performed by changing the LHE_COVERAGE parameter (i.e., the integer threshold value X to select the longest reads granting an X-fold coverage of the genome) from 15 up to 40. Another assembly was then performed by using FALCON build 180,808 (Chin et al. 2016) with standard parameters. Due to the computational resources required by falcon, only one assembly with recommended parameters was attempted.

Eventually, the falcon assembly was identified as the 'best' assembly based on the N50 of the contigs, assembly size, and BUSCO v.4.0.2 (Simão et al. 2015) results and therefore carried on to the next steps. The falcon assembled contigs were then polished with Polca (Zimin and Salzberg 2020) bundled with MaSuRCA v.3.4.1, Racon v1.4.17, and Pilon v.1.23 (Walker et al. 2014) to take full advantage of both types of reads also in this case. For all these software, we used the recommended parameters. The full analysis pipeline is reported in Fig. 1.

We used the information about the available parentals to reach a more accurate separation of the two haplotypes. In particular, the assembled and polished contigs have been aligned with BLAST + (version 2.11; Camacho et al. 2009) against the sequences of pummelo and mandarin as representatives of the maternal haplotype (H1) and to citron as the representative of the paternal haplotype (H2). The alignments were filtered with a minimum similarity threshold of 95% and subsequently parsed to assign each contig to the most similar haplotype. The sequences of the two haplotypes were then de-duplicated to correct potential positioning errors by using Purge Haplotigs v.1.1.1 (Roach et al. 2018). The final set of contigs of the primary haplotype was built as the primary sequence of H1 plus the alternative sequence of H2. Similarly, the primary sequence of H2 and the alternative sequence of H1 were used for the alternative haplotype.

The pseudo-chromosomes were finally built by RaGOO v.1.11 (Alonge et al. 2019). The software was fed with pummelo as the guiding sequence because, among the currently available genomes, it is the closest to *C. limon* that is organized into chromosomes. All the other parameters were set as default. The completeness of the gene space was then assessed with BUSCO v.4.0.2 (Simão et al. 2015) by using

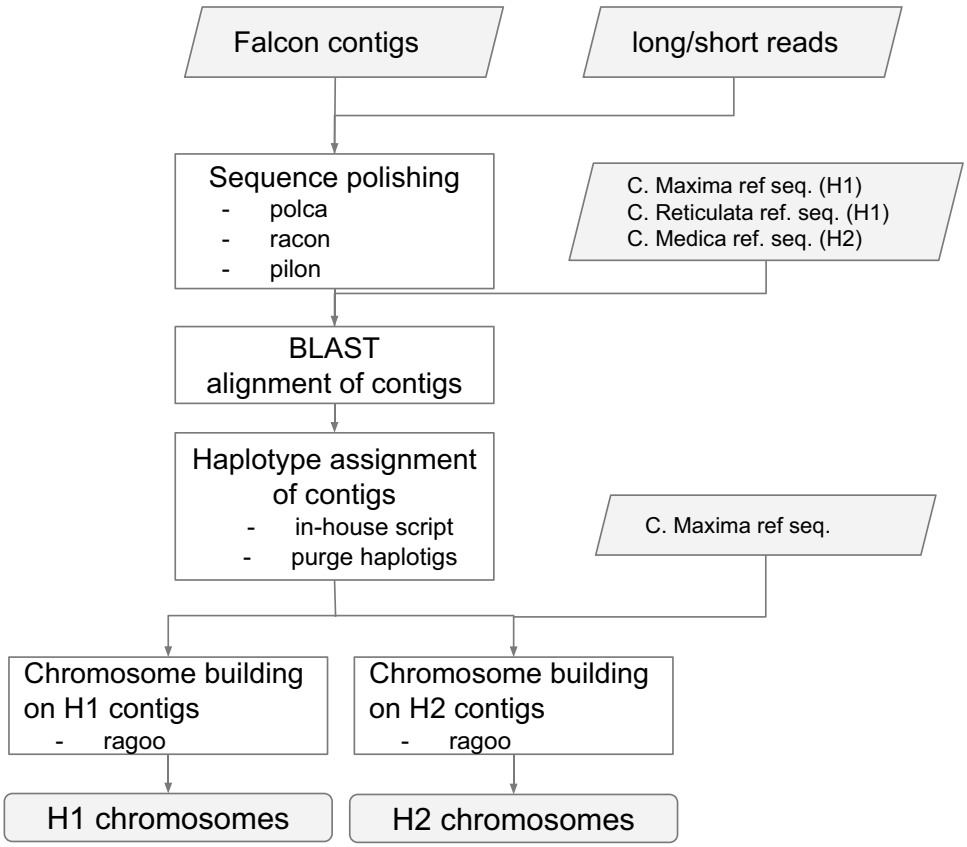

Fig. 1 Haplotype assembly workflow. Inputs are reported as shaded parallelograms, outputs as shaded rounded rectangles. Briefly, de novo assembled contigs were polished with long and short reads, then each polished sequence was aligned on the reference genomes of *C. maxima*, *C. reticulata*, and *C. medica*, and the alignments were then parsed to assign contigs to the H1 and H2 haplotypes. Purge Haplotigs software was then applied on these two sets to correct possible errors to this method. The two sets of contigs were then arranged into H1 and H2 chromosomes by using RaGOO and *C. maxima* as the guiding sequence

the 'embryophyta' lineage (-l parameter) featuring 1614 target genes.

A further analysis of the de novo assembled sequence was performed with the kmer-based software Merqury (Rhie et al. 2020) with default parameters.

Finally, the synteny with other published citrus genomes (i.e., *C. maxima*, *C. medica*, and *C. reticulata*) was visually inspected with the standalone version of D-Genies v.1.2.0 (Cabanettes and Klopp 2018).

miRNA and transposable elements (TEs) prediction and annotation

The sequences of known mature miRNA and the respective pre-miRNA hairpin sequences of sweet orange (*Citrus sinensis* (L.) Osbeck) were retrieved from the miRBase repository (version 22.1). The mature sequences were aligned against the primary and alternative assemblies using bowtie (version 1.2.2; Langmead et al. 2009) with the option end-to-end (-v) with 0 mismatches. The pre-miRNA sequences were blastn against the primary and alternative assemblies. Filters were adopted to assign miRNA loci with high confidence on the genome. Only miRNAs characterized by one locus on each genome assembly were considered for further filtering. The regions comprising the mature miRNA overlapping to the alignment of the associated pre-miRNA, with a p-identity $\leq 95\%$ and a ratio between the alignment length and the pre-miRNA length ≥ 0.95, were retrieved and the coordinates reported for each assembly. In case that 5p and 3p mature miRNA were available, then both had to be covered by the pre-miRNA sequence alignment. miRNA that did not fulfill the above-mentioned conditions were classified as low confident loci. A manual inspection was performed on this set to select those miRNAs that could be clearly recognized but that did not pass the pident or the ratio cutoffs. All the others were reported in a separate list.

The analysis to identify genomic regions associated with transposable elements (TEs) was performed using the package EDTA following the author's description (Ou et al. 2019) without providing a curated TE library and with default parameters but with –anno 1 and –evaluate 1.

Gene prediction and annotation

Gene prediction was performed using AUGUSTUS (Stanke et al. 2008) on both primary and alternative assemblies trained with assembled transcripts from Bridger (version 2014–12-01; Chang et al. 2015). RNA-seq reads obtained from 4 different tissues: leaf, fruit, flower, and root. Reads were filtered by quality, trimmed using Trimmomatic (version 0.39; Bolger et al. 2014), and independently assembled using Bridger. The four assembled datasets were then clustered together with CD-HIT (version 4.8.1; Li and Godzik 2006) using a threshold of 99% identity. GeneMarkS-T (GeneMark version 3.20;GMST; Tang et al. 2015) was then employed to retain transcripts with high coding sequence potential. The resulting transcripts were translated into the corresponding amino acid sequences and aligned using BLAST + to the Uniprot Uniref100 Viridiplantae dataset (The UniProt Consortium 2021; Bateman et al. 2021). Only the transcripts with alignment similarity greater than 80% and alignment length greater than 90% (compared to the shortest sequence) were kept. Those high-quality sets of assembled transcripts were then aligned on the Primary assembly using GenomeThreader (GTH version 1.7.1; Gremme et al. 2005), and a GeneBank file comprising 1 k upstream and downstream of genomic sequence was created for training AUGUSTUS. RNA-seq alignments using GSNAP (version 2020–12-16; Wu et al. 2016) were used to create exon and intron hint files to be included in the final prediction of AUGUSTUS as specified in the AUGUSTUS online documentation. Only predictions starting with ATG, without a stop-codon within the sequence and with biological evidence from such alignments, were considered and annotated using eggNOG (Huerta-Cepas et al. 2018), InterproScan (version 5.48.83; Jones et al. 2014), and BLAST + alignment on the Uniprot Uniref100 Viridiplantae dataset.

The complete gene prediction pipeline has been created using Singularity (version 2.6) and Nextflow (version 20.10; Di Tommaso et al. 2017). All the scripts used to produce these results are available on GitHub (https://github.com/marcomoretto/legend_transcriptome). The phylogenetic tree was created using Dendroscope v3.7.5 (Huson and Scornavacca 2012) starting from Orthofinder v2.5.4 multiple alignment results. Orthofinder was fed with amino acid sequences from all Citrus-related species as well as *C. limon.* using all default parameters except for the -M msa parameter used to obtain the multiple sequence alignment output file. The multiple sequence alignment file was then used with IQTree v1.6.12 (Nguyen et al. 2015) with ModelFinder parameters in order to determine the best-fit model. Branch support was then assessed with ultrafast bootstrap approximation using the following parameters -m JTT + F + R2 -alrt 1000 -bb 1000. Finally, the resulting tree file was fed to Dendroscope to create the phylogenetic tree.

Results and discussion

Lemon DNA sequencing

The genomic DNA of the lemon cv. 'Femminello Siracusano' was extracted from young leaves and sequenced coupling Illumina paired end and ONT sequencing platforms. A total of 40.2 Gb Illumina short reads were generated

Table 1 Illumina and Oxford Nanopore Technology (ONT) sequencing statistics

	Number of reads (M)	N50 (Kbp)	Number of bases (Gbp)
DNA (Illumina)	267	-	40.2
DNA (ONT)	5.04	13.7	42.2
	Number of reads (M)	Trimmed reads (M)	Number of bases (Gbp)
RNA (Flower)	23.42	23.25	3.51
RNA (Fruit)	22.41	22.28	3.36
RNA (Leaf)	20.03	19.84	3.00
RNA (Root)	23.18	23.01	3.48

(Table 1) and employed first to estimate the genome size and heterozygosity rate and then to perform hybrid assembly and polishing of the assembled sequence. As for the long reads sequencing, a total of 42.2 Gb reads were generated with the ONT platform with a N50 equal to 13.7, min and max read length equal to 0.1 and 139.9 Kbp, and a median length of the reads of 4.4 Kbp (Table 1, Supplementary Fig. 1).

Kmer analysis and genome size estimation

The haploid genome size was estimated to be about 312 Mb with a heterozygosity level of 3.56% and about 182 Mb of non-repetitive sequence (the 21-mer spectrum is reported in Supplementary Fig. 2).

Genome assembly and curation

The genome assembly of Illumina and ONT reads was carried out testing different assembly tools. A preliminary test was carried out employing the MaSuRCA genome assembler (Zimin et al. 2013). Sequences were assembled in 1600 contigs with contig N50 and N80 equal to 2,548,919 (with 63 contigs) and 628,658 (with 232 contigs) respectively. The size of the assembled sequence was equal to 668 Mb, far larger than the genome size of the known ancestors of lemon ranging from 302 Mb (pummelo) to 405 Mb (citron) and to the estimated haploid genome size of 'Femminello Siracusano' of 312 Mb. The genome assembly and annotation completeness were assessed using the BUSCO software (Simão et al. 2015) with the 'embryophyta' lineage. Results confirmed both the high quality of the sequenced data and the genome inflation since the complete sequence of 99% of the genes tested was retrieved, but only a limited fraction, 15.4%, was detected in single copy. The relevant incidence of duplicated genes is largely a consequence of the high heterozygosity of the genome hampering an efficient detection of the homolog haplotypes. This issue, if not adequately tackled, often results in the duplication of the region.

In light of this, ONT and Illumina reads were assembled using FALCON, while allelic contigs were reassigned using the BLAST + alignment against the parental genomes in combination with the Purge Haplotigs pipeline (Fig. 1, Roach et al. 2018). This approach led to the definition of a primary and an alternative haplotype: the first showed a genome size of 312.7986 Mb divided into 811 scaffolds with a N50 of 27.1 Mb. The alternative haplotype was instead characterized by a genome size of 324.74 Mb (divided into 799 scaffolds) with a N50 of 28.4 Mb. The genome assembly allowed the identification of nine pseudomolecules in each of the two assemblies (Fig. 2A). The shortest and longest chromosomes were Chr. 8 (18.72 Mb and 20.09 Mb in the primary and the alternative assembly respectively) and Chr. 2 (51.41 Mb and 49.87 Mb) (Fig. 2A, Table 2). The size of unanchored sequences was 39.31 Mb (divided into 802 contigs) and 38.81 Mb (in 790 contigs) in the two haplotypes (Supplementary Fig. 3). The correlation between the chromosome length in lemon and pummelo was highly consistent (corr = 0.97 and 0.98 for the primary and alternative assemblies respectively), with chromosomes eight and two being the longest and shortest chromosomes in *C. maxima* as well (Chr. 8 = 20.992 Mb; Chr. 2 = 52.984 Mb). Similarly, the collinearity between the two assembled haplotypes was also very high as shown in Supplementary Fig. 3C).

The kmer spectra of the assembled primary and alternative sequences were compared with the spectra of the Illumina reads by using Merqury. This analysis confirmed the haploid nature of the two sets of sequences and showed that they are a good representation of the two haplotypes (Supplementary Fig. 1). Based on the estimated genome size (312 Mb), it is likely that the alternative haplotypes might still contain a small portion of contigs that should be present in the primary sequence but that it was not possible to place correctly with the available information. We estimate this portion of contigs to be around 5–6 Mb, which corresponds to less than 2% of the total sequence.

The BUSCO analysis showed a significant number of complete sequences (95.9% and 94.8% of the 1614 'embryophyta' genes in primary and alternative haplotypes

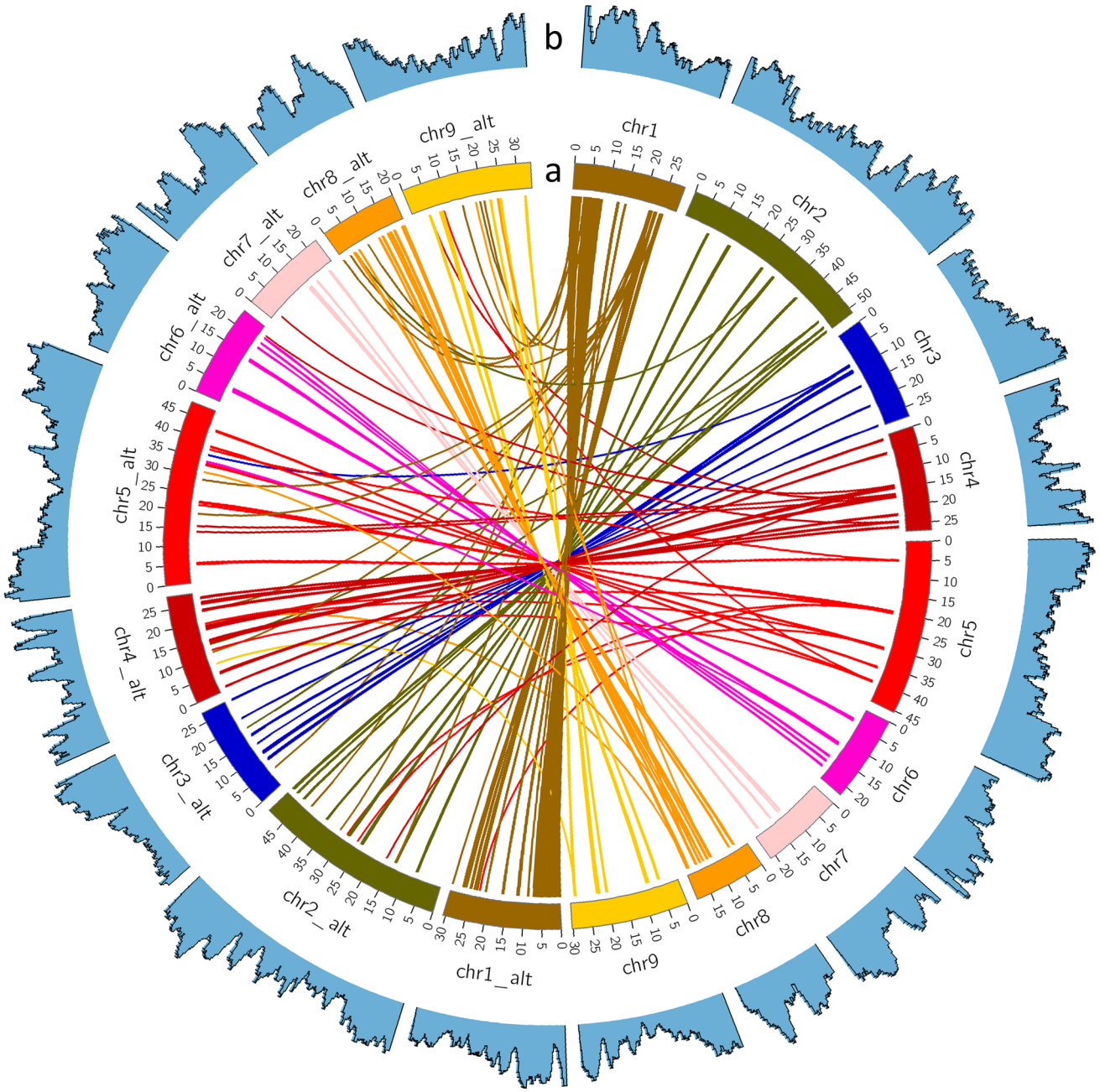

Fig. 2 **(A)** Primary (derived from sour orange) and alternative (derived from citron) assemblies of the lemon genome. Chromosomes are represented as colored blocks; the two haplotypes of the same chromosome are represented by the same color. The position of genes showing a BLAST + sequence identity of at least 96% and an alignment length of at least 60% compared to the smallest sequence are linked by straight lines. **(B)** Gene density distribution (blue histogram) across the primary and alternative assembly: the number of genes is calculated on a sliding window of 1 Mb with a 200 kb step

respectively) with a slightly increased number of duplicated sequences in the alternative haplotype (8.8% and 10.0% for primary and alternative haplotypes respectively).

miRNA and TE prediction and annotation

The annotation of microRNA genes on the genome was carried out performing an in silico analysis using the mature miRNA and the respective pre-miRNA hairpin sequences of *C. sinensis* present in the miRBase repository. The dataset

Table 2 Summary statistics of the primary and the alternative genome assembly. For each chromosome, the length in Mb is reported, as well as the number of transcripts, intact transposable elements (TEs), and the miRNA that were found with high confidence level detected

Chromosome	Length (Mb)		Transcripts		TEs		Mature MiRNA	
	Primary	Alternative	Primary	Alternative	Primary	Alternative	Primary	Alternative
chr1	29.11929	30.652384	3352	3333	342	327	5	5
chr2	51.414874	49.879738	5611	5413	546	506	51	29
chr3	27.512417	27.616602	3221	3077	293	299	9	6
chr4	27.165673	28.424913	3122	3080	291	275	11	10
chr5	45.399278	48.255654	5633	5547	436	434	14	14
chr6	22.460738	24.280274	2808	2882	212	259	10	10
chr7	21.365374	22.905193	2981	2968	211	213	12	11
chr8	18.720588	20.093211	2449	2422	188	226	7	6
chr9	30.330071	33.825048	3099	3085	359	358	8	7

was selected being the most complete in terms of the number of miRNA annotated for a *Citrus* species and considering the relatively low genetic distance between our genotype and sweet orange. The alignment of the 246 mature miRNA sequences showed that the miRNAs are well conserved between *C. sinensis* and *C. limon* with 233 aligning with perfect match on the genome, whereas 13 showed differences at nucleotide level. The alignments of the pre-miRNA sequences produced hits with perfect identity for 54 out of 151 pre-miRNA, 28 with identity between 99% and < 100%, 27 with identity between 98% and < 99%, 11 with identity between 97% and < 98%, 8 with identity between 96% and < 97%, and 4 with identity between 95% and < 96%. Among the remaining pre-miRNA, 10 were characterized by an identity between 90% and < 95% and the remaining by an identity < 90% with several mismatches and gaps or only with partial alignments. However, most of the pre-miRNA (115 out of 151 unique pre-miRNA in the miRBase dataset corresponding to 76%) were positioned univocally (Table 2, Supplementary Table 1) on the chromosomes intersecting the alignment information associated with the mature miRNA alignments. In addition, 22 loci characterized by the miRNA hairpin/mature association were assigned with lower confidence (Supplementary Table 2). A minor number of miRNA hairpin/mature pairs did not pass the filters applied in the analysis, and for these either no information could be retrieved due to the poor alignment quality (csi-miR3950, csi-miR3952, csi-miR3954) or the annotation was not confident enough to univocally assign the position on the genome (csi-miR169a, csi-miR395a, miR395c, csi-miR396b, csi-miR399c, csi-miR399a, csi-miR3948). The results indicate that the miRNAs are well conserved in *C. limon* compared to other members of the *Citrus* genus. Although this was expected for the mature miRNAs, it is noteworthy that almost one third of them presented a 100% identity with those of *C. sinensis*, while 131 pre-miRNA showed a range of identity higher or equal to 95%.

The analysis on the transposable elements identified regions where the entire structural elements characterizing the TEs could be annotated (intact TEs) as well as regions containing only part of these elements (fragmented TEs). Among the intact TEs, 933 LTRs, 412 MITEs, 1479 TIRs, and 54 Helitrons were identified in the primary assembly, whereas in the alternative assembly 1236 LTRs, 373 MITEs, 1237 TIRs, and 51 Helitrons were annotated (Supplementary Table 3). The distributions of the intact TEs on the chromosomes varied from 188 on Chr. 8 to 546 on Chr. 2 for the primary assembly and from 213 on Chr. 7 to 506 on Chr. 2 for the alternative assembly (Table 2). The total size of the genomic regions associated with intact transposable elements reached 9,277,703 bp (3.98% of the genome) and 10,892,621 bp (3.81% of the genome) in the primary and alternative assembly, respectively. The portion of the genome considering also all the regions associated with fragmented transposable elements and repeat regions amounted to 90,448,816 bp (33.07% of the genome) and 106,134,826 bp (37.12% of the genome) for the primary and alternative assembly, respectively (see supplementary gff3 files for the coordinates on the genome).

Gene prediction and annotation

Raw paired-ends RNA-seq reads coming from four tissues, 20,026,773 from leaf, 23,182,013 from root, 22,414,235 from flower, and 23,416,205 from fruit, were filtered and trimmed using Trimmomatic (Bolger et al. 2014) resulting in 19,843,227, 23,007,618, 22,277,820, and 23,250,076 sequences respectively. The four datasets were independently assembled using Bridger (Chang et al. 2015) resulting in 66,802 leaf, 66,937 root, 88,196 flower, and 81,260 fruit transcripts. Those datasets were merged using CD-HIT (Li and Godzik 2006) with a threshold of 99% sequence identity and lead to a unified dataset of 205,757 transcripts. GMST (Tang et al. 2015) identified 96,700 of such transcripts to

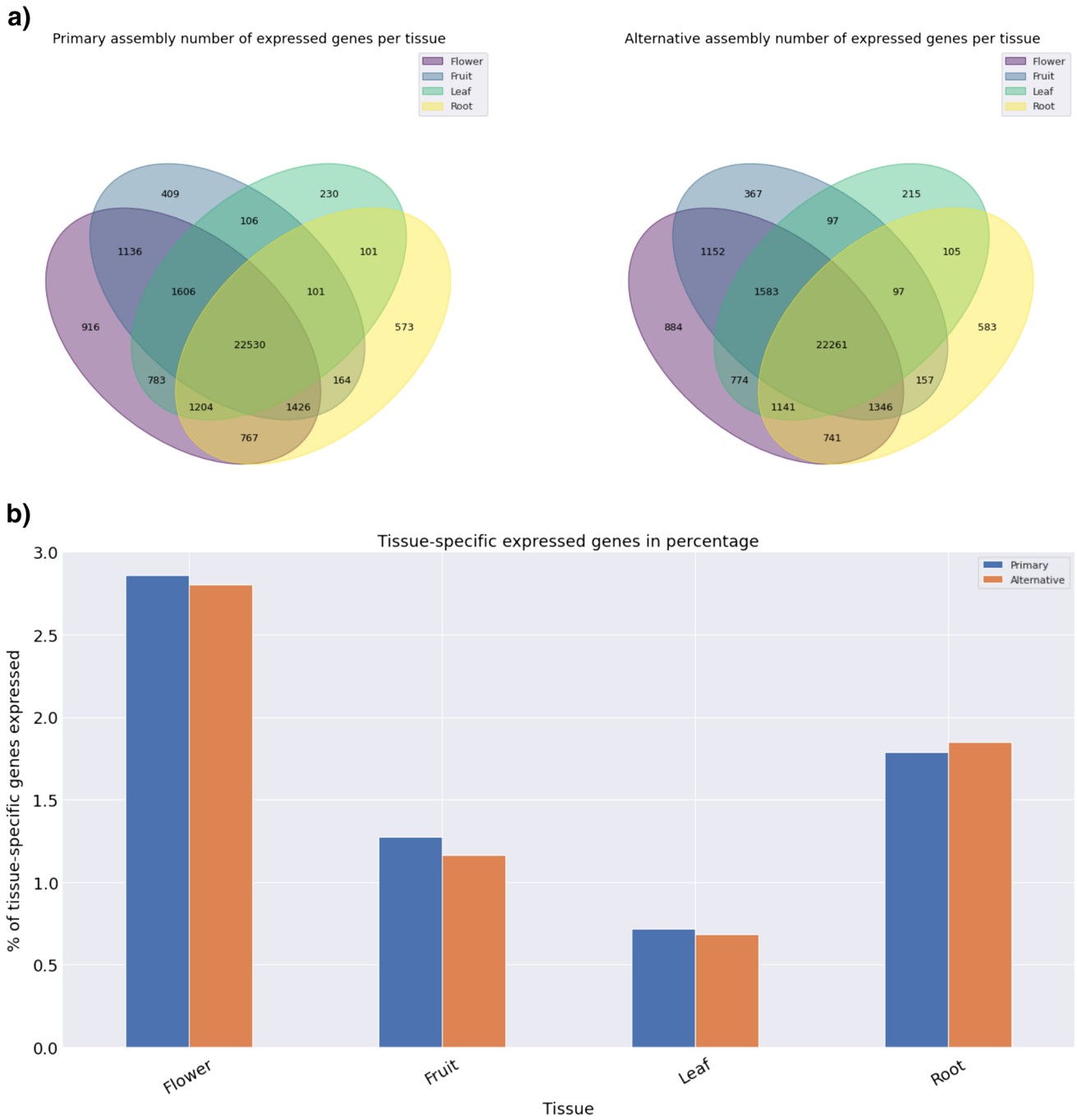

Fig. 3 **A** Venn diagram with number of expressed genes in flower, fruit, leaf, and root for the two haplotypes. **(B)** Bar plot representing the percentages of genes uniquely expressed in flower, fruit, leaf, and root in for the two haplotypes

be more likely to code for proteins. The alignment of those transcripts against the UniProt Uniref100 Viridiplantae dataset allowed to further filter 24,880 putative complete transcripts. GenomeThreader (Gremme et al. 2005) was able to completely align 11,964 transcripts on the primary assembly, and a genomic region comprising 1kbps upstream and downstream the alignment was retained for each transcript in order to create a GeneBank file to train AUGUSTUS (Stanke et al. 2008). After the training and optimization process, AUGUSTUS was used to predict genes on both primary and alternative assemblies using RNA-seq alignment results as biological hints to further enhance the prediction results. AUGUSTUS results were finally aligned using BLAST + on the Viridiplantae Uniprot Uniref100 database to filter out

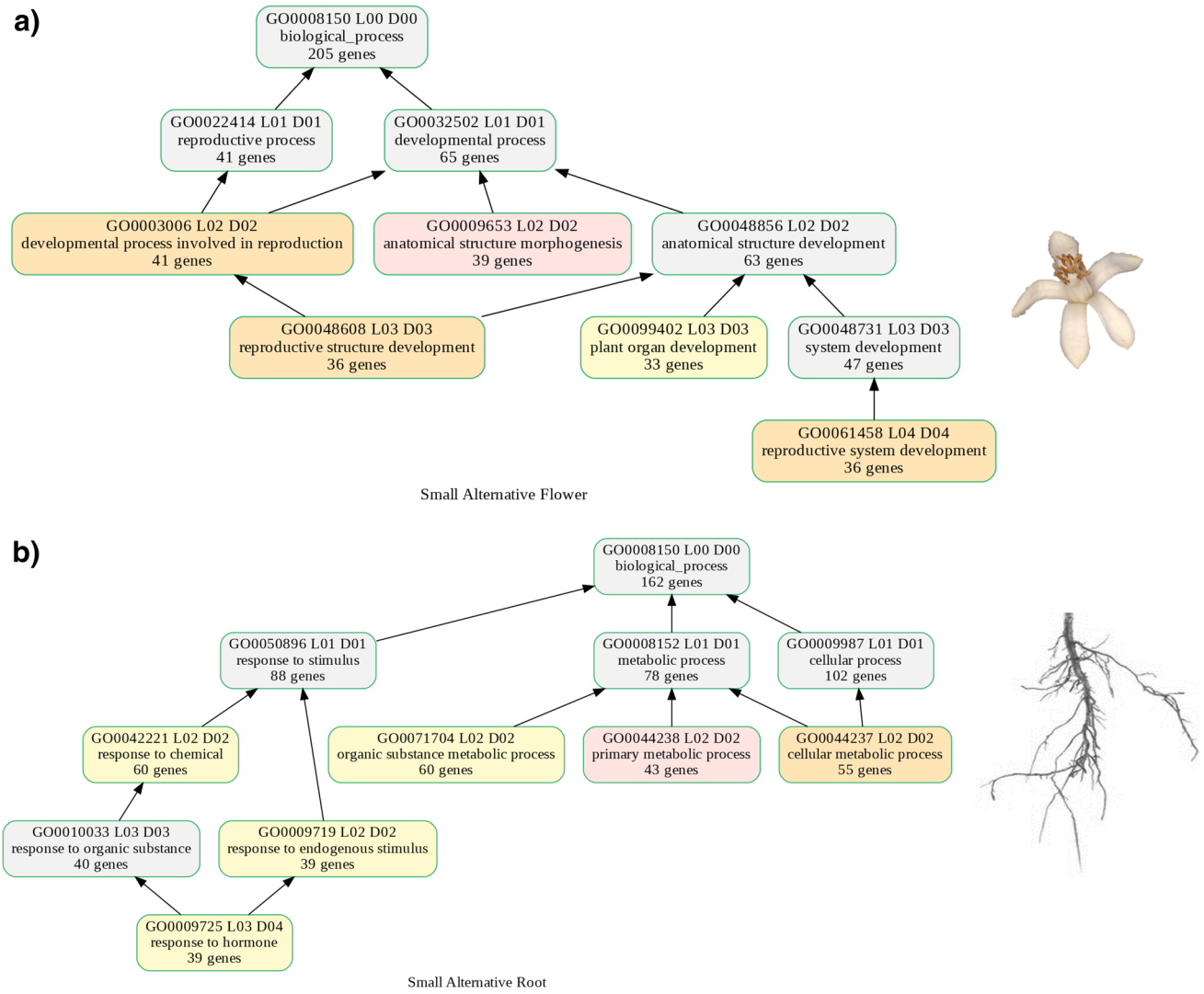

Fig. 4 Enriched GO terms for the genes involved in 'biological process' and mapped in the alternative assembly in flower **(A)** and root **(B)**. The different color indicates different *p* values, such as light red corresponds to a *p* value < 0.005, light orange corresponds to a *p* value < 0.01, and yellow corresponds to a *p* value < 0.05. Small enrichments are created using only the best significant GO terms

shorter transcripts with no BLAST hit. The final prediction consists of 35,020 and 34,577 predicted transcripts on primary and alternative assembly respectively (Table 2); the distribution of the transcripts along the nine chromosomes is represented in Fig. 2B.

The amino acid sequences of the predicted transcripts of the two haplotypes were merged and used, together with the proteome of ten citrus species, as input for OrthoFinder (Emms and Kelly 2019, version 2.5.4). The result shows a good consistency in comparison to the other species as reported in Supplementary Table 4. The same amino acid sequences were also employed to compute a rooted phylogenetic tree. The highest aminoacidic divergence among the eleven samples analyzed was detected for the Citrus-related species *Atlantia buxifolia*, *Poncirus trifoliata*, and *Fortunella hindsii*, while lemon showed the highest similarity with *Citrus medica* in agreement with their parental relationship (Supplementary Fig. 5). Lemon showed the highest relative frequency of both orthogroups (64.6%) and genes in species-specific orthogroups (17.9%) among the eleven species in analysis confirming the high quality of the gene assembly.

Among the predicted transcripts analyzed, 22,530 and 22,261 were detected in all four tissues in the primary and alternative assembly respectively (Fig. 3A). Overall, the number of predicted transcripts detected in the two assemblies was highly consistent in all the sets depicted in the Venn diagrams (average difference equal to 4.2% with standard deviation = 0.03). The set composed by the fruit-specific transcripts showed the highest relative difference (11.4%)

between primary (409 transcripts) and the alternative (367 transcripts). The highest number of tissue-specific transcripts was detected in Flower (916 and 884 in the primary and alternative assembly) while the lowest value was detected in Leaf (230 and 215, Fig. 3) with a relative frequency compared to all the other predicted transcripts equal to 2.86% and 2.81% for Flower and 0.72% and 0.68% for Leaf in the primary and alternative assembly respectively (Fig. 3B).

Gene Ontology functional annotation was performed using eggNOG and resulted in 27,607 and 27,249 annotated transcripts respectively for primary and alternative assembly. InterProScan assigned 28,077 and 27,684 transcripts at least one annotation term, while the BLAST + alignment on the Viridiplantae Uniprot Uniref100 dataset resulted in 29,469 and 29,058 hits with a putative known protein. The relative frequency of the genes according to the GO categories is depicted in Supplementary Fig. 6. As for the genes involved in 'Biological Process', the categories 'protein metabolic process' (2307 and 2245 genes in the primary and alternative respectively) and 'signal transduction' (45 and 57) showed the highest and lowest abundance respectively. Among the genes related to the category 'Molecular function', the highest number of genes was detected for 'transferase activity' (2291 and 2178), while the lowest was related to genes involved in 'oxygen binding' (6 and 3). Among the five sub-categories detected for the GO category 'Cellular component', the genes involved in the synthesis of the 'plasma membrane' were the most abundant (2308 and 2235), while those related to the lysosome showed the lowest relative frequency (70, 60; Supplementary Fig. 6). A GO analysis was also performed on the most represented 100 genes within the lemon-specific orthogroup (Supplementary Table 5). Among those, 68 were involved in biological processes, while 22 and 10 were related to cellular component and molecular function respectively.

In addition, an enrichment analysis was performed using GOATOOLS (Klopfenstein et al. 2018) on the four tissues and for each of the three GO aspects (Supplementary Fig. 7). To provide a more readable GO graph, Fig. 4 shows a smaller version of the GO enrichment for leaf and root, containing only the most significant GO categories. As shown in Fig. 4A, several GO terms related to 'biological process' were significantly more represented in flower compared to the other tissues; notably significant differences were detected for GO terms related to reproductive system development, reproductive structure development, plant organ development, developmental process involved in reproduction, anatomical structure morphogenesis for a total of 185 genes. As for the root tissue, several GO terms were significantly more represented compared to all the annotated genes. Some of these GO terms were related to response to exogenous or endogenous stimuli (response to chemical: 60 genes; response to hormone: 39 genes; response to endogenous stimulus: 39 genes), while others were related to metabolic process (organic substance metabolic process: 60 genes; primary metabolic process: 43 genes; cellular metabolic process: 55 genes), Fig. 4B. The same analysis was also performed for the top 100 represented genes within the species-specific orthogroup showing significant differences especially among the GO terms related to the cellular component (Supplementary Fig. 8).

Conclusions

Here, we present the first reference genome of lemon. Sequencing has been carried out combining short (Illumina) and long (Oxford Nanopore) sequencing approaches, while genomic annotation was based on the analysis of four tissues to provide a comprehensive overview of the genes expressed in lemon. The high-quality draft genome will represent a milestone toward the identification of the genetic determinism of traits of agronomical interest and the setup of marker-assisted breeding effort to select novel lemon selections characterized by superior characteristics (i.e., enhanced fruit quality and/or resistance to severe disease such as mal secco) (Poles et al. 2020; Catalano et al. 2021).

Supplementary Information The online version contains supplementary material available at https://doi.org/10.1007/s11295-021-01528-5.

Author contribution **MDG:** conceptualization, investigation, writing—original draft; **MM:** conceptualization, formal analysis, investigation, writing—original draft; **MM:** conceptualization, formal analysis, investigation, writing—original draft; **CC:** investigation; **TM:** conceptualization; **DZ:** investigation; **CA:** investigation; **CM:** investigation, writing—review and editing; **DG:** investigation, writing—review and editing, funding acquisition; **LMS:** conceptualization, writing—review and editing, funding acquisition; **BL:** conceptualization, formal analysis, investigation, writing—original draft; **GA:** conceptualization, writing—review and editing, funding acquisition.

Funding Projects 'Development of Resistance Inductor against Citrus Vascular Pathogens' (Sviluppo di Induttori di Resistenza a Patogeni Vascolari negli Agrumi, S.I.R.P.A., http://www.progettosirpa.it/home, 08CT7211000254) and 'Fruit Crops Resilience to Climate Change in the Mediterranean Basin' (FREECLIMB, https://primafreeclimb.com/) and 'Valutazione di genotipi di agrumi per l'individuazione di fonti di resistenza a stress biotici e abiotici' (Linea 2 del Piano della Ricerca di Ateneo 2020, University of Catania) are supporting the proposed work related to new biotechnological approaches carried out to unlock genetic basis of mal secco resistance and to obtain new tolerant genotypes. The APC was funded by Fondi di Ateneo 2020–2022, University of Catania, linea Open Access. Mario Di Guardo took part on this work in the frame of the PON 'AIM: Attrazione e Mobilità Internazionale', project number 1848200–2.

Declarations

Conflict of interest The authors declare no competing interests.

Data archiving statement The genome assembly sequences and gene predictions have been submitted to the citrus genome database (https://www.citrusgenomedb.org/Analysis/1462349) where they can be downloaded and accessed through the genome browser and BLAST services. Raw data have been submitted to NCBI's SRA under the bioproject id PRJNA732837.

References

Alonge M, Soyk S, Ramakrishnan S et al (2019) RaGOO: fast and accurate reference-guided scaffolding of draft genomes. Genome Biol 20:1–17

Barry GH, Caruso M, Gmitter FG (2020) Chapter 5—Commercial scion varieties. In: Talon M, Caruso M, Gmitter FG (eds) The genus citrus. Woodhead Publishing, pp 83–104

Bateman A, Martin MJ, Orchard S et al (2021) UniProt: the universal protein knowledgebase in 2021. Nucleic Acids Res 49:D480–D489. https://doi.org/10.1093/nar/gkaa1100

Bolger AM, Lohse M, Usadel B (2014) Trimmomatic: a flexible trimmer for Illumina sequence data. Bioinformatics 30:2114–2120. https://doi.org/10.1093/bioinformatics/btu170

Cabanettes F, Klopp C (2018) D-GENIES: dot plot large genomes in an interactive, efficient and simple way. PeerJ 6:e4958

Camacho C, Coulouris G, Avagyan V et al (2009) BLAST+: architecture and applications. BMC Bioinformatics 10:421. https://doi.org/10.1186/1471-2105-10-421

Carbonell-Caballero J, Alonso R, Ibañez V et al (2015) A phylogenetic analysis of 34 chloroplast genomes elucidates the relationships between wild and domestic species within the genus citrus. Mol Biol Evol 32:2015–2035. https://doi.org/10.1093/molbev/msv082

Catalano C, Di Guardo M, Distefano G et al (2021) Biotechnological approaches for genetic improvement of lemon (Citrus limon (l.) burm. f.) against mal secco disease. Plants 10:1–16. https://doi.org/10.3390/plants10051002

Chang Z, Li G, Liu J et al (2015) Bridger: a new framework for de novo transcriptome assembly using RNA-seq data. Genome Biol 16:1–10. https://doi.org/10.1186/s13059-015-0596-2

Chin C-S, Peluso P, Sedlazeck FJ et al (2016) Phased diploid genome assembly with single-molecule real-time sequencing. Nat Methods 13:1050–1054

Di Tommaso P, Chatzou M, Floden EW et al (2017) Nextflow enables reproducible computational workflows. Nat Biotechnol 35:316–319

Emms DM, Kelly S (2019) OrthoFinder: phylogenetic orthology inference for comparative genomics. Genome Biol 20:1–14. https://doi.org/10.1186/s13059-019-1832-y

Food and Agriculture Organization of the United Nations. FAOSTAT (1997) Statistical Database. [Rome] :FAO

Gremme G, Brendel V, Sparks ME, Kurtz S (2005) Engineering a software tool for gene structure prediction in higher organisms. Inf Softw Technol 47:965–978. https://doi.org/10.1016/j.infsof.2005.09.005

Gulsen O, Roose ML (2001) Chloroplast and nuclear genome analysis of the parentage of lemons. J Am Soc Hortic Sci 126:210–215. https://doi.org/10.21273/jashs.126.2.210

Huerta-Cepas J, Szklarczyk D, Heller D et al (2018) eggNOG 5.0: a hierarchical, functionally and phylogenetically annotated orthology resource based on 5090 organisms and 2502 viruses. Nucleic Acids Res 47:D309–D314. https://doi.org/10.1093/nar/gky1085

Huson DH, Scornavacca C (2012) Dendroscope 3: an interactive tool for rooted phylogenetic trees and networks. Syst Biol 61:1061–1067. https://doi.org/10.1093/sysbio/sys062

Jones P, Binns D, Chang HY et al (2014) InterProScan 5: genome-scale protein function classification. Bioinformatics 30:1236–1240. https://doi.org/10.1093/bioinformatics/btu031

Klopfenstein DV, Zhang L, Pedersen BS et al (2018) GOATOOLS: a Python library for Gene Ontology analyses. Sci Rep 8:1–17. https://doi.org/10.1038/s41598-018-28948-z

Langmead B, Trapnell C, Pop M, Salzberg SL (2009) Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. Genome Biol 10. https://doi.org/10.1186/gb-2009-10-3-r25

Li W, Godzik A (2006) Cd-hit: a fast program for clustering and comparing large sets of protein or nucleotide sequences. Bioinformatics 22:1658–1659. https://doi.org/10.1093/bioinformatics/btl158

Mehl F, Marti G, Boccard J et al (2014) Differentiation of lemon essential oil based on volatile and non-volatile fractions with various analytical techniques: a metabolomic approach. Food Chem 143:325–335. https://doi.org/10.1016/j.foodchem.2013.07.125

Muccilli V, Vitale A, Sheng L et al (2020) Substantial equivalence of a transgenic lemon fruit showing postharvest fungal pathogens resistance. J Agric Food Chem 68:3806–3816. https://doi.org/10.1021/acs.jafc.9b07925

Nguyen L-T, Schmidt HA, von Haeseler A, Minh BQ (2015) IQ-TREE: a fast and effective stochastic algorithm for estimating maximum-likelihood phylogenies. Mol Biol Evol 32:268–274. https://doi.org/10.1093/molbev/msu300

Nicolosi E, Deng ZN, Gentile A et al (2000) Citrus phylogeny and genetic origin of important species as investigated by molecular markers. Theor Appl Genet 100:1155–1166. https://doi.org/10.1007/s001220051419

Ou S, Su W, Liao Y et al (2019) Benchmarking transposable element annotation methods for creation of a streamlined, comprehensive pipeline. Genome Biol 20:275. https://doi.org/10.1186/s13059-019-1905-y

Poles L, Licciardello C, Distefano G et al (2020) Recent advances of in vitro culture for the application of new breeding techniques in citrus. Plants 9:1–25. https://doi.org/10.3390/plants9080938

Rhie A, Walenz BP, Koren S, Phillippy AM (2020) Merqury: reference-free quality, completeness, and phasing assessment for genome assemblies. Genome Biol 21:1–27. https://doi.org/10.1186/s13059-020-02134-9

Roach MJ, Schmidt SA, Borneman AR (2018) Purge Haplotigs: allelic contig reassignment for third-gen diploid genome assemblies. BMC Bioinformatics 19:1–10

Simão FA, Waterhouse RM, Ioannidis P et al (2015) BUSCO: assessing genome assembly and annotation completeness with single-copy orthologs. Bioinformatics 31:3210–3212

Singh B (1981) Establishment of first gene sanctuary in India for Citrus in Garo Hills. Concept Publishing Company

Stanke M, Diekhans M, Baertsch R, Haussler D (2008) Using native and syntenically mapped cDNA alignments to improve de novo gene finding. Bioinformatics 24:637–644. https://doi.org/10.1093/bioinformatics/btn013

Sun Y, Singh Z, Tokala VY, Heather B (2019) Harvest maturity stage and cold storage period influence lemon fruit quality. Sci Hortic (amsterdam) 249:322–328. https://doi.org/10.1016/j.scienta.2019.01.056

Tang S, Lomsadze A, Borodovsky M (2015) Identification of protein coding regions in RNA transcripts. Nucleic Acids Res 43:1–10. https://doi.org/10.1093/nar/gkv227

The UniProt Consortium, UniProt: the universal protein knowledgebase in 2021 (2021) Nucleic Acids Research 49(D1):D480–D489, https://doi.org/10.1093/nar/gkaa1100

Vurture GW, Sedlazeck FJ, Nattestad M et al (2017) GenomeScope: fast reference-free genome profiling from short reads. Bioinformatics 33:2202–2204. https://doi.org/10.1093/bioinformatics/btx153

Walker BJ, Abeel T, Shea T, et al (2014) Pilon: an integrated tool for comprehensive microbial variant detection and genome assembly improvement. PLoS One 9:e112963

Wang L, He F, Huang Y et al (2018) Genome of wild mandarin and domestication history of mandarin. Mol Plant 11:1024–1037. https://doi.org/10.1016/j.molp.2018.06.001

Wang X, Xu Y, Zhang S et al (2017) Genomic analyses of primitive, wild and cultivated citrus provide insights into asexual reproduction. Nat Genet 49:765–772. https://doi.org/10.1038/ng.3839

Wu GA, Prochnik S, Jenkins J et al (2014) Sequencing of diverse mandarin, pummelo and orange genomes reveals complex history of admixture during citrus domestication. Nat Biotechnol 32:656–662. https://doi.org/10.1038/nbt.2906

Wu GA, Terol J, Ibanez V, et al (2018) Genomics of the origin and evolution of Citrus. Nature. https://doi.org/10.1038/nature25447

Wu J, Fu L, Yi H (2016) Genome-wide identification of the transcription factors involved in citrus fruit ripening from the transcriptomes of a late-ripening sweet orange mutant and its wild type. PLoS ONE 11:1–22. https://doi.org/10.1371/journal.pone.0154330

Zimin AV, Marçais G, Puiu D et al (2013) The MaSuRCA genome assembler. Bioinformatics 29:2669–2677

Zimin AV, Puiu D, Luo M-C et al (2017) Hybrid assembly of the large and highly repetitive genome of Aegilops tauschii, a progenitor of bread wheat, with the MaSuRCA mega-reads algorithm. Genome Res 27:787–792

Zimin AV, Salzberg SL (2020) The genome polishing tool POLCA makes fast and accurate corrections in genome assemblies. PLoS Comput Biol 16:e1007981